LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay

Yihuai Lan[1*], Zhiqiang Hu[2*], Lei Wang[3], Yang Wang[4], Deheng Ye[5], Peilin Zhao[5],
Ee-Peng Lim[3], Hui Xiong[1], Hao Wang[1†]
[1]The Hong Kong University of Science and Technology (Guangzhou), [2]Singapore University of Technology and Design
[3]Singapore Management University, [4]Verily Life Sciences, [5]Tencent
{yihuailan, haowang}@hkust-gz.edu.cn

ABSTRACT

This paper aims to investigate the open research problem of uncovering the social behaviors of LLM-based agents. To achieve this goal, we adopt Avalon, a representative communication game, as the environment and use system prompts to guide LLM agents to play the game. While previous studies have conducted preliminary investigations into gameplay with LLM agents, there lacks research on their social behaviors. In this paper, we present a novel framework designed to seamlessly adapt to Avalon gameplay. The core of our proposed framework is a multi-agent system that enables efficient communication and interaction among agents. We evaluate the performance of our framework based on metrics from two perspectives: winning the game and analyzing the social behaviors of LLM agents. Our results demonstrate the effectiveness of our framework in generating adaptive and intelligent agents and highlight the potential of LLM-based agents in addressing the challenges associated with dynamic social environment interaction. By analyzing the social behaviors of LLM agents from the aspects of both collaboration and confrontation, we provide insights into the research and applications of this domain.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent planning**; **Natural language processing**; • **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

Large Language Models, LLM Agents, Social Game Playing

1 INTRODUCTION

Artificial intelligence (AI) agents [21, 37] refer to the entities that are able to perform the human-like behaviors, from perceiving and analyzing the environment, to making decisions and taking actions. The rapid advancements in large language models (LLMs) [16, 22, 29, 31] have ushered in new possibilities for creating AI agents in complex environment, which sheds light on the potential of LLM-based agents simulating the human society. Various works [11, 12, 21, 23] have been proposed to simulate different aspects of human society. For example, Qian et al. [23] aim to simulate the software development company, with agents from different social identities. Park et al. [21] use the sandbox environment and assign diverse social roles to agents, simulating the agent's lives within the game world. However, these previous works mostly focus on inspecting the positive social behaviors only, such as honesty, collaboration, and etc. While relevant studies on the negative social behaviors of LLM agents are still rare so far.

Much prior research in human society has identified issues such as misinformation and conflicts on the web, prompting investigations and solutions to address these concerns [5, 17, 26]. To gain a deeper understanding of the social behaviors of LLM agents, we aim to conduct comprehensive investigations into both positive and negative aspects of their behavior. To this end, we utilize Avalon as the environment to demonstrate the collaboration and confrontation among agents. Avalon is a representative communication game, in which the players are assigned with hidden roles and divided into two opposing teams. During gameplay, players engage in discussions, debates, and strategic actions.

It is non-trivial for the LLM agents to win this incomplete information game, as they are required to share and obtain information through communication and analysis, including deducing other players' roles, building trust among allies, and deceiving opponents. On the one hand, the agents must possess multiple technical abilities, such as natural language understanding, incomplete information analysis, and strategy learning and planning, to succeed. On the other hand, the social behaviors of LLM agents, including teamwork, persuasion, and deception, are also essential for success in Avalon gameplay.

To investigate the LLM-based agent society, we propose a novel framework for the agents to play Avalon. Specifically, we adopt ChatGPT as the players and assign various roles to agents. We adopt system prompts to guide LLM agents to play Avalon automatically. Following human's thinking methodology, we incorporate multiple modules, including memory storage and summarization, analysis and planning, game action and response generation, and experience learning. We utilize a competitive baseline approach [38], to elaborate the efficacy of our proposed framework. We also carefully analyze the social behaviors of LLM agents, and observe clear collaboration and confrontation between agents during the gameplay.

Our contributions can be summarized as:

*Both authors contributed equally to this research.
†The corresponding author.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

- We explore the social behaviors exhibited by LLM-based agents in the context of Avalon gameplay. We reveal the various aspects of these behaviors, including teamwork, leadership, persuasion, deception, and confrontation.
- We design an effective framework to play Avalon, which presents superior performance compared with the baseline method. We also carefully analyse the relationship between the module design and agents' social behaviors, providing comprehensive experiment discussions.
- Our findings have the potential to contribute to a better understanding of the role of LLM-based agents in social and strategic contexts, and shed light on the implications of these behaviors in such environments.

2 RELATED WORK

2.1 LLMs' Impact on Society

The impact of LLMs on various aspects of society has increasingly drawn the attention of researchers [20]. One innovative application of LLMs is in virtual social network simulations. Specifically, Gao et al. [11] propose to build a social-network simulation system to foster advancements in social science research. Furthermore, Kaiya et al. [15] demonstrate LLM-based agents also have the potential to enhance human social experiences within virtual environments. However, despite the growth of social computing research driven by LLMs, concerns regarding validity, privacy, and ethical implications have emerged [25]. To address these concerns, a discussion is proposed on developing feedback-enriched models for social systems, through the utilization of generative agent based models [12].

LLMs also promote the development of social robots. The detection of social bots presents both challenges and opportunities in this domain [9]. Yang et al. [39] observe AI-enabled social robots may introduce risks such as promoting dubious websites and disseminating harmful content. To mitigate these issues, a comprehensive survey [35] on alignment technologies has been conducted. This survey aims to tackle problems such as misunderstanding human instructions, generating potentially biased content, or experiencing hallucinations in LLMs. By evaluating the alignment of LLMs, Liu et al. [19] propose to promote the deployment of reliable and ethical LLMs across various applications.

2.2 LLM-Based Agents

The rapid development of LLM-based agents has led to advancements in various aspects of problem-solving. These agents, which are equipped with fast and deliberate thought, have demonstrated increased efficiency and robustness in addressing complex tasks [18, 30, 33]. For instance, LLMs have been successfully applied to tasks that associated with database administrators [41]. Moreover, numerous studies [4, 21, 23, 27, 36] have focused on using multi-agent systems for solving complex tasks. In [21], Park et al. develop a simulated society comprised of multiple unique LLM agents, which has generated human-like behaviors. Further, several multi-agent frameworks have been proposed, including AgentVerse [4], AutoGen [36], and MetaGPT [13], all designed to facilitate various collaborative tasks. Additionally, some social complex tasks, such as court simulations and software development, become possible through the collaboration between LLM-based agents [8, 23, 27]. It is observed multi-agent debate systems have shown promise in addressing tasks like open-ended question-answering and dialogue response generation [3]. For example, LLM-based agents in negotiation scenario can learn strategies from a third agent [10].

However, most existing research in this domain focuses on either purely collaborative or debative relationships between agents. While in our study, it is essential to consider the combination of cooperation and confrontation in agent relationships, as this better reflects the complexity of real-world interactions and warrants further exploration.

2.3 Gameplay with LLMs

LLMs have been utilized in a wide range of gaming contexts [1, 6, 7, 14, 24]. Plan4MC [40] employs the planning capabilities of LLMs to break down the complex task execution process in Minecraft into a series of basic skills. The GITM framework [42] integrates planning, feedback, and text-based memory mechanisms to create Generally Capable Agents (GCAs) [28] in Minecraft. Additionally, the lifelong learning agent VOYAGER [32] is able to continuously explore the Minecraft environment.

LLMs have also been applied to multi-player strategy games, such as the Prisoner's Dilemma and Battle of the Sexes [2]. To explore the application scenarios with rich communication, several works [34, 38] have designed LLM-based agents for multi-player social deduction games such as Werewolf and Avalon, which demand effective player communication. These studies have shown that LLM-based agents exhibit strategic behaviors such as trust, confrontation, camouflage, and leadership [38]. In [34], recursive contemplation has been proposed to enhance agents' ability to discern and counteract misinformation.

3 BACKGROUND

In our study, we use Avalon as the environment, in stead of Werewolf. It is known both of them are social deduction games. However, Werewolf eliminates players as the game progresses, which can lead to potentially less participation for players who are eliminated early. In contrast, Avalon's gameplay structure promotes social cohesion by ensuring that all players remain involved throughout the entire game. Therefore, Avalon is our selected game environment.

3.1 Avalon Introduction

Avalon, also known as "The Resistance", is a social deduction board game. This game is adapted for 5-10 players. In our work, we adopt the 6-player version of the game.

Player roles. The roles include Merlin, Percival, Morgana, Assassin and two Loyal Servants. These roles are divided into the good and evil sides. Among them, Merlin, Percival and loyal servants belong to the good side, while Morgana and Assassin belong to the evil side. Each player will be assigned a role secretly, where some roles may have special abilities. Specifically, Morgana and Assassin know each other at the beginning. Percival sees both Merlin and Morgana but does not know their exact identities. Merlin knows the identity of the evil characters, but they do not know who Merlin is.

Quest team assignment. After players receiving their respective roles, they will engage in 3-5 rounds of discussion and vote process to assemble a quest team, which consists of 2 or 3 players. At the

beginning of a round, a leader will be assigned on a rotational basis. The leader hosts the discussion, then all players need to vote on the quest team members in public. If more than half of the votes are in agreement, the quest team will be successfully assembled. Otherwise, the leader moves to the next player and the players discuss and vote again. In each round, discussion and vote process can be conducted up to five times. The leader can directly assign quest members after the fifth discussion.

Quest phase. The quest outcome is determined by the success or failure cards submitted by the quest team members. Good players can only submit success cards, while evil players have the option to submit either success or failure cards. A quest is deemed successful if all team members vote for success, whereas it fails if one or more members cast a failure vote.

End of the game. The game ends when either three quests succeed, resulting in victory of the good side, or three quests fail, leading to victory of the evil side. However, the evil players can still secure a win at the end by accurately identifying which player is Merlin.

3.2 Social Behaviors in Avalon

Teamwork. Good players must work together to successfully complete quests and ultimately win the game. To this end, they need to establish trust with their teammates while remaining cautious of the evil players.

Leadership. Each player has the chance to lead the discussion for assembling the quest team. This allows the leader to steer the conversation and facilitate the development of trust among players. Effective leadership can be a crucial factor in achieving victory.

Persuasion. Players must use their communication skills to persuade others to believe their claims, trust their judgments, and support their decisions.

Deception. Evil players must hide their true intentions and identities by pretending to be good players. This involves misleading others through deceptive tactics and strategically concealing information.

Confrontation. Disagreements and conflicts will arise during the game. Players must tackle these confrontations and work towards resolving them.

4 APPROACH

Figure 1 displays the proposed Avalon LLM agents framework, and the following section will introduce each module.

4.1 Setup

To start the game, system prompts are used to assign these roles to different LLM agents, respectively. A system prompt for a role p_i include several important components: Role Information $\mathcal{RI}^{p_i}$ (Role Name and Role Introduction), Goal $\mathcal{G}^{p_i}$ (Winning Conditions), and Abstracted Strategy $\mathcal{S}^{p_i}$ for gameplay. The Role Name and Role Introduction inform the LLM agent about its assigned role, while the Goal (Winning Conditions) provides insights into how the winning can be achieved. Furthermore, the Initial Playing Strategy outlines the high-level planning for the LLM agent to take specific actions during gameplay. Below is a specific example of a system prompt for the role of Margana:

Role: `Morgana`.

Role Introduction: `In identification phase, you can identify who the Assassin is, as well as your teammate.`
Goal: `Win the game by deliberately making the quests fail for three rounds, either through yourself or your teammates.`
Initial Strategy: `You always pretend to be a loyal servant and recommend yourself as a candidate for quests, and let the quests fail.`

4.2 Memory Storage and Summarization

The analysis of game history is of paramount importance for agents as it enables them to assess the current situation comprehensively and make informed decisions regarding subsequent actions. However, the cumulative word count of the responses generated by LLM agents in Avalon often exceed the input length limit of LLM and can be too large to fit within a single prompt. Additionally, LLM's performance may decrease when processing long inputs. To address this, a memory storage is created to store a detailed record of the conversations among LLM agents, which can then be used for subsequent analysis and decision making.

4.2.1 Memory Storage. Memory storage plays a crucial role in maintaining a comprehensive record of agents' conversation history in the current round of the game. It is a structured list of memory objects, each containing important information such as the role name, a detailed response in natural language description, the corresponding round number, and a flag indicating whether the object is public or private. Public information is the content that every role can see, while private information is specific to the conversation that occurred for a particular role. For example, the information revealed after the statement "Merlin, open your eyes and see the agents of evil". We allocate distinct memory pools to each agent for the purpose of subsequent analysis and decision-making. This approach ensures clarity and organization in processing information, aiding agents in their tasks. By storing these valuable pieces of information, the memory storage enables agents to access and review past conversations. This facilitates a better understanding of the game progress.

4.2.2 Memory Summarization. As the Avalon game progresses, the stored information in memory significantly increases, surpassing the input length limitation of LLM. In order to harness a greater amount of information from memory for subsequent analysis and decision-making, we employ a process wherein we summarize previous information before incorporating the memory objects from the current round into the memory. Additionally, we utilize a summarization prompt to condense the memory objects from the previous round, rather than merely storing them verbatim. This summarization is more concise and also retains essential information. The process of updating the memory with a summarization of the previous round is illustrated below:

$$\mathcal{M}_t = \left\langle \text{Summarize}(\mathcal{M}_{t-1}), \left(\mathcal{R}_t^{p_1} \cdots, \mathcal{R}_t^{p_6}, \mathcal{I}_t \right) \right\rangle. \quad (1)$$

The memory on round t is represented as $\mathcal{M}_t$. The memory object containing the response generated by the LLM for role p_i on round t is denoted as $\mathcal{R}_t^{p_i}$, and $\mathcal{I}_t$ represents the instructions and statements of the host on round t. $\langle\rangle$ denotes text concatenation.

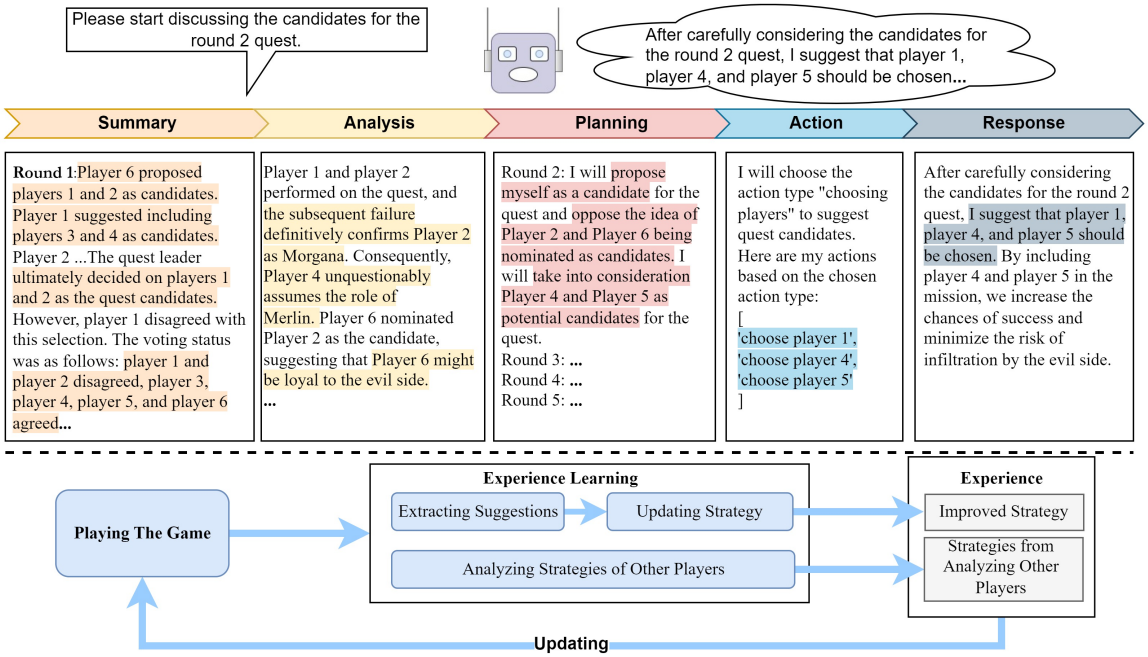


Figure 1: Our proposed framework consists of six modules, i.e. summary, analysis, planning, action, response and experience learning. This pipeline design follows the thinking methodology of humans, which helps the LLM agents play Avalon effectively and reveals the social behaviors of LLM agents.

Summarize$(\cdot)$ is the summarization function based on the summarization prompting approach. The specific summarization prompt is shown in Table 1.

4.3 Analysis and Planning

4.3.1 Analysis. To assist LLM agents in enhancing their strategic planning and increasing their chances of winning, we introduce an extra analysis module. This module is developed to thoroughly analyze the role identity of other players and strategies other players may potentially employ during gameplay, which can be formulated as below:

$$\mathcal{H}_t^{p_i} = \text{Analyze}\left(\mathcal{M}_t, \mathcal{RI}^{p_i}\right), \tag{2}$$

where $\mathcal{M}_t$ represents the memory on round t, and role information $\mathcal{RI}^{p_i}$ is provided by the system prompt. By providing this comprehensive analysis, LLM agents will be equipped with a deeper understanding of their collaborator and competitors, enabling them to make better decisions and develop effective counterstrategies that help lead to winning. The example prompt is illustrated in Table 1.

4.3.2 Planning. Agents must possess a profound understanding of the game overall progress and the necessary strategies to secure a victory in the present situation. Consequently, a planning module is deployed to aid agents in formulating a strategic plan before taking any action. This plan is based on the memory and information derived from the current round of the game, as outlined below:

$$\mathcal{P}_t^{p_i} = \text{Plan}\left(\mathcal{M}_t, \mathcal{H}_t^{p_i}, \mathcal{P}_{t-1}^{p_i}, \mathcal{RI}^{p_i}, \mathcal{G}^{p_i}, \mathcal{S}^{p_i}\right), \tag{3}$$

where $\mathcal{P}_t^{p_i}$ denotes the strategic position of agent p_i at round t, while $\mathcal{M}_t$ and $\mathcal{H}_t^{p_i}$ represent the memory and comprehensive analysis at round t respectively. Additionally, $\mathcal{RI}^{p_i}$, $\mathcal{G}^{p_i}$, and $\mathcal{S}^{p_i}$ signify the essential information such as role details, goals, and initial strategies of agent p_i. By formulating a strategic plan, the agents will possess an adaptable strategy tailored to the prevailing circumstances. This strategic foresight enables them to make informed decisions concerning collaboration with teammates, strategic deception of opponents, assuming the identity of the opposing faction, and, when necessary, making the difficult choice of sacrificing either teammates or oneself to secure victory in the game. The example prompt to illustrate this concept can be found in Table 1.

4.4 Action and Response Generation

4.4.1 Action. Within the action module, agents will determine their next course of action by evaluating available memory information, conducting a thorough analysis, and adhering to the established strategic plan. There are five types of actions the agents can take including choosing players, voting (agree or disagree), performing quests (make quest succeed or fail), using non-verbal signals (raising hands up, put hands down, open eyes or close eyes), and choosing to remain silent. The process of choosing the next action with available information is illustrated below:

$$\mathcal{A}_t^{p_i} \sim p\left(\mathcal{A}|\mathcal{M}_t, \mathcal{H}_t^{p_i}, \mathcal{P}_t^{p_i}, \mathcal{RI}^{p_i}, \mathcal{G}^{p_i}, \mathcal{S}^{p_i}, \mathcal{I}_t'\right). \tag{4}$$

The determination of the subsequent action is contingent upon the utilization of the memory $\mathcal{M}_t$, the comprehensive analysis $\mathcal{H}_t^{p_i}$, the strategic plan $\mathcal{P}_t^{p_i}$, and the specific instruction provided by the host in round t. Within the action module, agents select a particular action in alignment with the established strategic plan. It is imperative to note that the details of these action decisions remain confidential and are accessible exclusively to the respective agent. Neither the host nor other players have visibility into these

decisions. The prompt that illustrates the action decision making process is presented in Table 1.

4.4.2 Response Generation. Following meticulous analysis and decision-making procedures, the response to the host's inquiry is generated within the Response Generation module. Within this response, agents delineate the selected action and provide a corresponding explanation to the host. It is important to note that in crafting these explanations, agents are not bound by truthfulness. Consequently, agents are afforded the flexibility to collaborate with teammates, deceive opponents, and assume the identity of the opposite faction through natural language descriptions. The prompt elucidating the process of response generation is shown in Table 1.

4.5 Experience Learning

In practical scenarios, humans can enhance their Avalon gameplay strategy through accumulated game experience. Furthermore, players not only accumulate insights from their individual perspectives but also from observing strategies employed by other players. Consequently, an ideal Avalon LLM agent should possess the ability to draw from both its own experiences and the collective experiences of other players.

4.5.1 Self-Role Strategy Learning. In Step 1, agents are tasked with generating three strategic recommendations for a player's role-specific gameplay in Avalon games. These suggestions are derived from an analysis of the entire game history. To ensure the applicability of these suggestions in future games, agents are instructed to refrain from directly referencing specific players, using the respective role names instead.

In Step 2, following the initial suggestion extraction process, agents proceed to enhance their strategies by incorporating the three suggestions gathered, while preserving the merits of the original strategy.

4.5.2 Other-Role Strategy Learning. In Step 1, to facilitate the acquisition of strategies from other players, the Avalon LLM agents are instructed to first summarize the strategies employed by their counterparts.

In Step 2, to facilitate the learning process from strategies employed by other players, we incorporate the extracted suggestions and summaries of these strategies into the Avalon game instructions provided to the agents. The prompts of the above steps are presented in Table 2 of Appendix.

5 EXPERIMENT

5.1 Implementation Details

We implemented the Avalon game program in Python. The gpt-3.5-turbo-16k model is served as our and baseline's backend LLMs. For all experiments, we set the temperatures of the agent model and the LLM extractor to be 0.3 and 0 respectively. The number of suggestion generating for updating strategies is 3. We set the game rules and role descriptions referring to the template of baseline [38], which retrieves and reflects historical context, enhances agent reasoning, and also learns from past mistakes. The descriptions set in system prompt for both our method and baseline can be seen in Section 8.1.

5.2 Evaluation Metrics

We evaluate the performance of our framework based on metrics from two perspectives.

5.2.1 Gameplay Outcome and Strategy. From this perspective, we use metrics associated with the gameplay outcome and strategies to quantitatively evaluate the performance of the proposed agents and the baseline agents.

Winning Rate (WR). The winning rate typically refers to the percentage of games a particular player or team wins out of the total number of games played. It's simply the number of wins divided by the total number of games played, expressed as following:

$$WR = (\frac{Number\ of\ Wins}{Total\ Number\ of\ Games\ Played})x100\% \tag{5}$$

Quest Engagement Rate (QER). The term "quest engagement rate" denotes the proportion of rounds in which a specific player participates in the quest team out of the total number of rounds played during the games. The calculation of the quest engagement rate (QER) is depicted by the following equation:

$$QER = (\frac{Number\ of\ Egagement\ Rounds}{Total\ Number\ of\ Rounds})x100\% \tag{6}$$

Failure Vote Rate (FVR) The quest outcome is determined by the success or failure cards submitted by the quest team members. The failure vote rate refer the percentage of votes against the success of a quest that leads to its failure. The failure vote rate is computed with the following equation:

$$FVR = (\frac{Numner\ of\ Failure\ Votes}{Total\ Number\ of\ Votes})x100\% \tag{7}$$

5.2.2 Social Behaviors. From this perspective, we employ metrics related to social behaviors, such as teamwork, leadership, persuasion, deception, and confrontation, to assess the social interactions exhibited by AI agents in Avalon games.

Leadership. To assess the leadership of the AI agents, we use a metric called "Leader Approval Rate (LAR)". This rate is calculated by dividing the total number of approval votes by the total number of votes when agents serve as the leader in 20 consecutive Avalon games. The Leader Approval Rate indicates the consensus among other players regarding the quest team members proposed by the leader. We use ChatGPT to count the votes.

Persuasion. To assess LLM agents' persuasion skills, we use two metrics: self-recommendation rate and success rate. The self-recommendation rate is calculated by dividing the number of rounds where agents propose themselves for the quest team by the total rounds played. The success rate represents the proportion of rounds where self-recommendation leads to active participation in the quest team. ChatGPT is used to detect instances of LLM agents' self-recommendation and determine its success.

Deception. Detecting deception behaviors in AI agents poses a considerable challenge. Our evaluation centers on identifying instances where the agents assume other identities during the initial round of each game. We categorize AI agent behaviors into three distinct types: Self-Disclosure, Camouflage, and Withholding Identity. To analyze these behaviors, we utilized ChatGPT to assess the responses of different roles during the first round of each game.

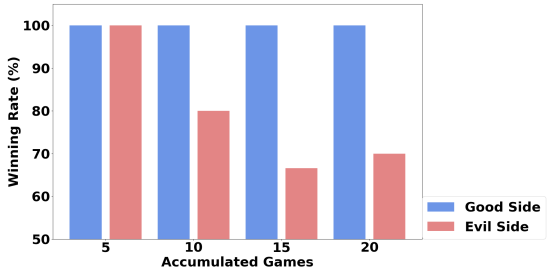


Figure 2: Results of the gameplay between ours and baseline. We present the winning rates of our method being good and evil sides.

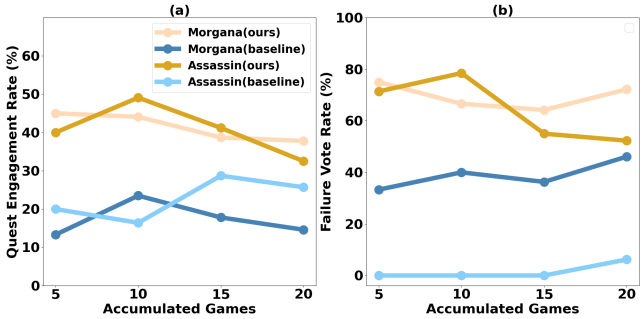


Figure 3: (a): Comparison of the engaging quests rate when playing Evil side. Higher engaging quests rate means more opportunities for the player to influence the outcome of the game. (b): Comparison of the failure vote rate when playing Evil side. Baseline has a lower rate.

Teamwork and Confrontation. We utilize ChatGPT to scrutinize the responses of various roles, with the specific objective of identifying instances where agents either collaborate with their fellow players (teamwork) or engage in confrontations with others (confrontation). This analysis is achieved by prompting ChatGPT with a specific player's response and evaluating the trust (teamwork), lack of trust (confrontation), or ambivalence towards other players. The specific prompts used for these metrics are detailed in the Section 8 of Appendix.

5.3 Experiment Results

To validate the efficacy of our proposed Avalon AI agents, we utilized Werewolf AI agents [38] and adapted them for Avalon games as baseline agents. We conducted two series of 20 consecutive Avalon games, where our Evil Side played against the baseline Good Side for 20 games, and conversely, our Good Side played against the baseline Evil Side for another 20 games. Following the gameplay sessions, we conducted a comparative analysis of the winning rates achieved by the proposed Avalon AI agents in contrast to the baseline.

As depicted in Figure 2, our method demonstrated a 100% winning rate in 20 games when playing the good side. Conversely, when playing the evil side, the winning rate was 70% over the same number of games. Furthermore, an observation was made regarding a decrease in the winning rate, transitioning from 100% to 66.6%, as the number of consecutive Avalon games increased from 5 to 15. This decline may be attributed to the learning mechanism employed by the baseline, where responses from winning games carry a high

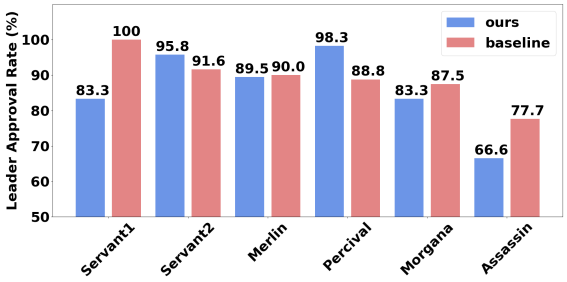


Figure 4: The leadership behavior when playing different roles: Players with higher Leader Approval Rate get more agreements from other players when deciding a quest team.

probability of being utilized to enhance future gameplay. Consequently, the proposed Avalon AI agents engaged in competitive matches against the baseline agents. Notably, our Avalon Agents demonstrated an increase in the winning rate from 66.6% to 70% between games 15 and 20.

To gain a deeper understanding of the strategies utilized by our Avalon Agents and the baseline agent, we conducted a comparative analysis of the quest engagement rate and failure voting rate when various AI agents assumed the roles of the evil side. Both the quest engagement rate and the failure voting rate hold considerable influence over the game's outcome. A higher quest engagement rate signifies increased opportunities for players to sway the game's result. Moreover, a higher voting rate indicates a greater probability for the evil side to secure victory but also heightens the risk of exposing their identity, reflecting an aggressive gameplay strategy. The outcomes regarding the quest engagement rate and the failure voting rate are presented in Figure 3. It is evident that our proposed AI agents adopt an assertive approach when assuming the roles of Morgana and Assassin, with an average quest engagement rate of 35.2% and a failure voting rate of 62.3% when joining the quest team. In contrast, the baseline agents exhibit a quest engagement rate of only 20.2% and a relatively low failure voting rate of 26.2%. Consequently, our proposed Avalon AI agents achieve a 70% winning rate against the baseline agents when playing as the evil side.

6 SOCIAL BEHAVIORS OF AI AGENTS

When humans play Avalon, the game involves teamwork, leadership, persuasion, deception, and confrontation. Teamwork is essential as players collaborate to complete missions, while leadership skills are needed to select the right team members. Persuasion is used to engage themselves into the quest team or convince others of one's innocence or guilt, and deception is a central element where players must hide their true intentions and even assueme the identity of the opposing faction. Confrontation occurs when players challenge each other's statements, leading to intense interactions. These elements create a dynamic and engaging social experience in the game.

In order to assess whether the AI agents naturally emulate social behaviors observed in human players during Avalon games, we conduct a comprehensive analysis. This investigation focuses on evaluating the agents' execution of social behaviors, with a specific emphasis on analyzing the frequency distribution of teamwork, leadership, persuasion, deception, and confrontation. The entirety

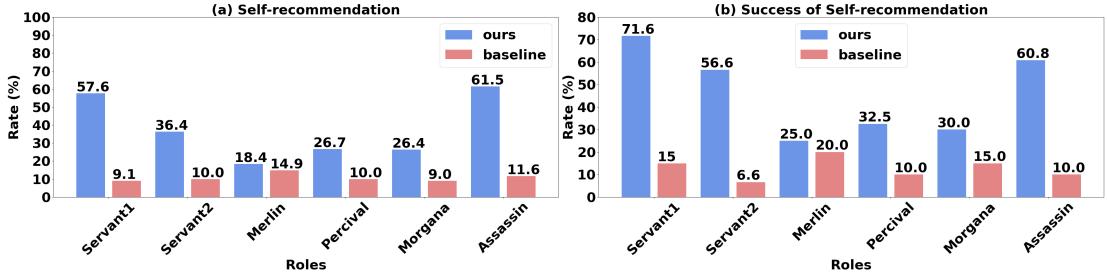


Figure 5: The persuasion behavior when playing different roles: Self-recommendation Rate: players with higher Self-recommendation Rate are more will to engage in quests. Self-recommendation Success Rate: players more likely to gain the trust of other players has higher Self-recommendation Success Rate.

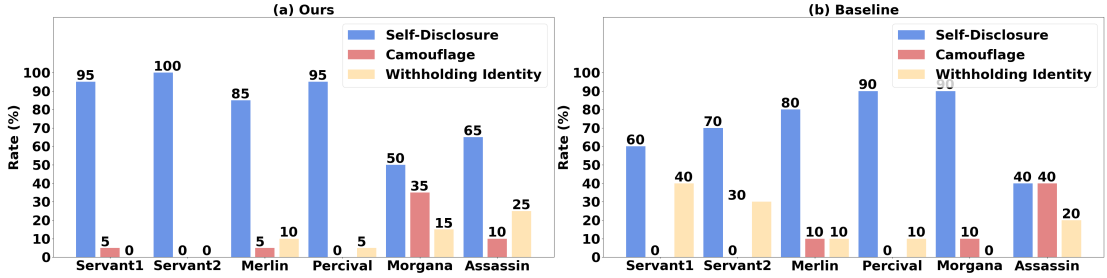


Figure 6: The deception behavior when playing different roles: at first round of each game, the distribution of the players choose Self-Disclosure, Camouflage or Withholding Identity.

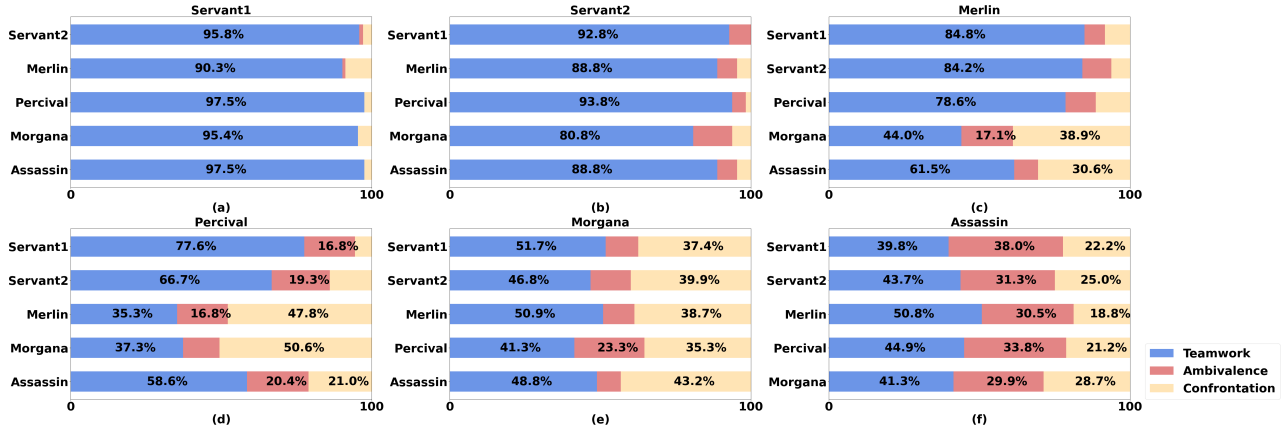


Figure 7: The teamwork and confrontation behaviors when playing different roles (Ours): each subfigure shows the attitude distribution of the player portraying specific role (on the top) towards players in other roles (on the left).

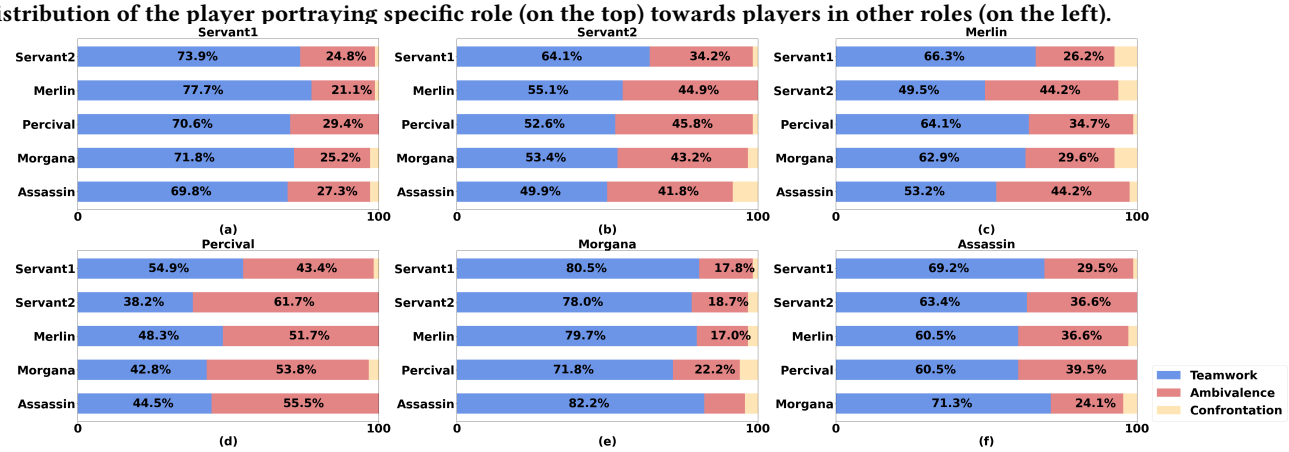


Figure 8: The teamwork and confrontation behaviors when playing different roles (Baseline): each subfigure shows the attitude distribution of the player portraying specific role (on the top) towards players in other roles (on the left).

of our analysis is founded upon the game logs extracted from the two sets of 20 consecutive Avalon games, as depicted in Figure 2.

6.1 Leadership

Leadership skills come into play when players take charge of discussions and decision-making processes. A good leader can steer the conversation, guide suspicions, and rally the loyal servants to make informed decisions. Leadership abilities are crucial for the good side to effectively counter the deceptive tactics employed by the evil side.

Figure 4 illustrates the Leader Approval Rate when agents assume various roles. It is evident that agents, whether playing on the good side or the evil side, attain remarkably high Leader Approval Rates when serving as leaders. Notably, the AI agents achieve a Leader Approval Rate exceeding 80% while undertaking roles associated with the good side. This signifies their robust leadership qualities and their proactive approach to steering the gameplay towards victory. It is worth mentioning that the agents' inclination to agree with others could contribute to the observed high Leader Approval Rates.

6.2 Persuasion

Persuasion is a key aspect of Avalon, especially during team discussions and mission planning phases. Players use persuasive techniques to convince others of their loyalty or to sow doubt about other players. Persuasion is vital for both good side roles, who aim to trust their allies, and evil side roles, who attempt to deceive and mislead the group.

Figure 5 presents the outcomes of the evaluation conducted to assess the persuasion ability of the AI agents. Notably, the agents exhibit distinct strategies contingent upon the roles they assume. When playing the roles of Loyal Servant and Assassin, the agents demonstrate a high self-recommendation rate to participate in the quest team, influencing the mission's success or failure. Consequently, the elevated success rate of self-recommendation observed in these roles underscores the potent persuasion abilities possessed by the AI agents. Conversely, a cautious approach is evident when the agents assume the roles of Merlin, Percival, and Morgana, as indicated by their low self-recommendation rates and corresponding success rates. This strategic restraint is crucial in Avalon gameplay, especially for players like Merlin, where concealing the true identity holds paramount importance. Furthermore, in comparison to the baseline agents, the proposed Avalon Agents exhibit higher rates of self-recommendation and a greater success rate in their self-recommendation efforts. This difference highlights the enhanced persuasion abilities demonstrated by the proposed agents.

6.3 Deception

Deception is at the heart of Avalon. Evil side roles must deceive the loyal servants by pretending to be one of them while subtly sabotaging missions. Skilled players can create elaborate lies and misdirections to confuse their opponents. On the other hand, loyal servants must also engage in deception to protect their true identities, especially when they are under suspicion.

In Figure 6, the rates of various behaviors exhibited by AI agents are displayed. Notably, the agents display a notably high tendency to reveal their identities at the commencement of the game, particularly among the roles associated with the good side. Intriguingly, in the roles of Morgana and Assassin, agents opt to either conceal or assume different identities without explicit instructions to do so in the initial strategy. Specifically, Morgana displays a 35% rate of assuming alternate identities, a strategy akin to that observed in human players, where Percival perceives both Merlin and Morgana but lacks precise knowledge of their identities. This spontaneous adoption of deceptive behaviors by AI agents stands out as a captivating observation, underscoring their adaptability and strategic acumen in the pursuit of game victory.

6.4 Teamwork and Confrontation

Teamwork is essential for the loyal servants to identify each other and work together to complete missions successfully. Players often collaborate to strategize, discuss mission assignments, and share information to identify the evil roles. While confrontation arises when suspicions lead to accusations. Accused players must defend themselves, leading to intense confrontations within the group. Accusers must present their reasoning, and the accused must either provide a convincing defense or attempt to deflect suspicion onto others.

In Figure 7, the teamwork and confrontation rates of good side roles are depicted. It is evident that Loyal Servants tend to collaborate with other players, given their lack of information about specific identities. Conversely, agents serving as Merlin, possessing knowledge of the identities of Morgana and Assassin, exhibit a high rate of confrontations with these adversaries. Additionally, Percival, who is aware of both Merlin and Morgana without knowing their exact identities, chooses to confront both of them. These observations highlight the AI agents' ability to employ appropriate strategies, utilizing teamwork and confrontation based on the available information. This behavior closely aligns with the social dynamics observed in human players during Avalon games.

Figure 8 illustrates the teamwork and confrontation rates of the baseline agents. It is apparent that the rates of teamwork and confrontation remain consistent across various roles. This indicates that the baseline agents do not adjust their strategies based on the specific roles they assume.

7 CONCLUSION

In this paper, we examine the social behaviors of LLM-based agents within the Avalon communication game. We propose a multi-agent framework that facilitates efficient communication and interaction among agents. This framework comprises memory, analysis, planning, action, and response modules, which can learn from experience. Unlike previous studies, our research focuses on the social dynamics of these agents in gameplay scenarios. Through our evaluation, we not only demonstrate the success of our framework in achieving game-winning strategies but also highlight the adaptability of LLM agents in complex social interactions, both collaborative and confrontational. Future efforts include optimizing our approach, investigating its applicability in different game environments, and gaining a deeper understanding of LLM's potential in dynamic social interaction.

REFERENCES

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. *arXiv preprint arXiv:2305.16867* (2023).

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. *ArXiv* abs/2305.16867 (2023). https://api.semanticscholar.org/CorpusID:258947115

[3] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *ArXiv* abs/2308.07201 (2023). https://api.semanticscholar.org/CorpusID:260887105

[4] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848 [cs.CL]

[5] Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Fei Li, Meihuizi Jia, and Jiaqi Li. 2022. EvidenceNet: Evidence Fusion Network for Fact Verification. In *Proceedings of the ACM Web Conference 2022*. 2636–2645.

[6] Prateek Chhikara, Jiarui Zhang, Filip Ilievski, Jonathan Francis, and Kaixin Ma. 2023. Knowledge-enhanced Agents for Interactive Text Games. *arXiv preprint arXiv:2305.05091* (2023).

[7] Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2023. An Appraisal-Based Chain-Of-Emotion Architecture for Affective Language Model Game Agents. *arXiv preprint arXiv:2309.05076* (2023).

[8] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration Code Generation via ChatGPT. arXiv:2304.07590 [cs.SE]

[9] Emilio Ferrara. 2023. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday* (2023). https://api.semanticscholar.org/CorpusID:259470739

[10] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback. arXiv:2305.10142 [cs.CL]

[11] Chen Gao, Xiaochong Lan, Zhi jie Lu, Jinzhu Mao, Jing Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *ArXiv* abs/2307.14984 (2023). https://api.semanticscholar.org/CorpusID:260202947

[12] Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. 2023. Generative Agent-Based Modeling: Unveiling Social System Dynamics through Coupling Mechanistic Models with Generative Artificial Intelligence. *ArXiv* abs/2309.11456 (2023). https://api.semanticscholar.org/CorpusID:262063524

[13] Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. 2023. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. *ArXiv* abs/2308.00352 (2023). https://api.semanticscholar.org/CorpusID:260351380

[14] Jun Inukai, Tadahiro Taniguchi, Akira Taniguchi, and Yoshinobu Hagiwara. 2023. Recursive Metropolis-Hastings Naming Game: Symbol Emergence in a Multi-agent System based on Probabilistic Generative Models. *arXiv preprint arXiv:2305.19761* (2023).

[15] Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe Agents: Generative agents for low-cost real-time social interactions. https://api.semanticscholar.org/CorpusID:263608891

[16] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[17] Sharon Levy, Robert E Kraut, Jane A Yu, Kristen M Altenburger, and Yi-Chia Wang. 2022. Understanding conflicts in online conversations. In *Proceedings of the ACM Web Conference 2022*. 2592–2602.

[18] Bill Yuchen Lin, Yicheng Fu, Karina Yang, Prithviraj Ammanabrolu, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2023. SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks. *ArXiv* abs/2305.17390 (2023). https://api.semanticscholar.org/CorpusID:258960143

[19] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hanguang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *ArXiv* abs/2308.05374 (2023). https://api.semanticscholar.org/CorpusID:260775522

[20] Rajiv Movva, S. Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2023. Large language models shape and are shaped by society: A survey of arXiv publication patterns. *ArXiv* abs/2307.10700 (2023). https://api.semanticscholar.org/CorpusID:259991588

[21] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]

[22] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[23] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. *ArXiv* abs/2307.07924 (2023). https://api.semanticscholar.org/CorpusID:259936967

[24] Jacob Sharf, Mustafa Omer Gul, and Yoav Artzi. 2023. CB2: Collaborative Natural Language Interaction Research Platform. *arXiv preprint arXiv:2303.08127* (2023).

[25] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the Emerging Norms of Using Large Language Models in Social Computing Research. *ArXiv* abs/2307.04280 (2023). https://api.semanticscholar.org/CorpusID:259501062

[26] Qiurong Song and Jiepu Jiang. 2022. How Misinformation Density Affects Health Information Search. In *Proceedings of the ACM Web Conference 2022*. 2668–2677.

[27] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. arXiv:2306.03314 [cs.AI]

[28] Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard, and Wojciech Marian Czarnecki. 2021. Open-Ended Learning Leads to Generally Capable Agents. arXiv:2107.12808 [cs.LG]

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[30] Chen Feng Tsai, Xiaochen Zhou, Sierra S Liu, Jing Li, Mo Yu, and Hongyuan Mei. 2023. Can Large Language Models Play Text Games Well? Current State-of-the-Art and Open Questions. *arXiv preprint arXiv:2304.02868* (2023).

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[32] 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:2305.16291 [cs.AI]

[33] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).

[34] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's Game of Thoughts: Battle Against Deception through Recursive Contemplation. arXiv:2310.01320 [cs.AI]

[35] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning Large Language Models with Human: A Survey. *ArXiv* abs/2307.12966 (2023). https://api.semanticscholar.org/CorpusID:260356605

[36] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *ArXiv* abs/2308.08155 (2023). https://api.semanticscholar.org/CorpusID:260925901

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

[38] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. arXiv:2309.04658 [cs.CL]

[39] Kai-Cheng Yang and Filippo Menczer. 2023. Anatomy of an AI-powered malicious social botnet. *ArXiv* abs/2307.16336 (2023). https://api.semanticscholar.org/CorpusID:260334464

[40] Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks. arXiv:2303.16563 [cs.LG]

[41] Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. 2023. LLM As DBA. arXiv:2308.05481 [cs.DB]

[42] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. arXiv:2305.17144 [cs.AI]

8 APPENDIX

8.1 Game Rules and Role Description

You are playing a game called the Avalon with some other players. This game is based on text conversations. Here are the game rules:
Roles: The moderator is also host, he organised this game and you need to answer his instructions correctly. Don't talk with the moderator. There are five roles in the game, Merlin, Percival, Loyal Servant, Morgana, Assassin. Merlin, Percival and Loyal Servant belong to good side and Morgana and Assassin belong to evil side There are two alternate phases in this game, reveal phase and quest phase. When it's reveal phase: You need follow the instructions of the moderator. You needn't worry about other players and moderator knowing what you say and do. No need to worry about suspicions from others during the phase. If you are Merlin, you can know which two players are Morgana and Assassin but you can't know which one is Morgana or Assassin specifically. If you are Percival, you can know which two players are Merlin and Morgana but you can't know which one is Merlin or Morgana specifically. If you are Morgana, you can know which player is Assassin.If you are Assassin, you can know which player is Morgana. If you are Loyal Servant, you can't get any information in this phase.
The quest phase includes 5 rounds. A round includes discussion, voting and execute quest. At every round, all players need to discuss about which players (candidates) will execute the quest at current round. And then all players need to vote if the candidates should execute the quest, if the agreement exceeds 1/2, the candidates will execute the quest, otherwise, discuss again and vote again. When executing quest, the candidates need to choose to make quest successful or failed. If all candidates choose to make quest successful, the quest will succeed. If anyone makes the quest failed, the quest will fail.
At the end of a round, if the quest succeed, good side will get one point, otherwise, evil side will get one point. Which side get 3 points earlier, which side wins the game. If you are Assassin, at the end of a round, you can choose to identify which one is Merlin, if the identifying is successful, the red camp directly win the game. If not successful, the Assassin will expose his identification.
Objectives: your goal is to help your side get 3 points and win the game. If you are Assassin, you also need to reason which player is Merlin as early as possible.
Tips: To complete the objective: you should analyze and use your ability correctly. During quest phase, you need to reason carefully about the roles of other players and be careful not to reveal your own role casually unless you're cheating other players. Only give the player's name when making a decision/vote, and don't generate other players' conversation. Reasoning based on facts you have observed and you cannot perceive information (such as acoustic info) other than text. You are {player}, the {role}. You're playing with 5 other players. Do not pretend you are other players or the moderator. Always end your response with '<EOS>'.

8.2 Module Prompts

Our designed prompts for different modules are presented in Tables 1 and 2.

8.3 Heuristic Rules for LLM Gameplay

In the gameplay, we used LLM to extract information from the responses of the agents. For example, when the agent selects a player, it extracts the player number, and when voting, it extracts the player's voting result. With several demonstrations of how to extract corresponding information, LLM can extract information very accurately to help the game proceed smoothly.

It is observed agents sometimes may fail to answer questions correctly, such as voting with unclear attitudes. In order to allow the game to proceed smoothly, we design the following heuristic rules. When voting for quest candidates, if the agent's answer is unclear, we assume that it agrees. When voting the quest for success or failure, if the agent's answer is unclear, we default to it voting for failure. When agents select an excessive number of players, we truncate the selection to meet the quest's requirements. In cases where the agents choose too few players, the host will repeat question to the agent. If the required player count is still not met even after multiple retries, the program steps in to assist by making a random selection on behalf of the agent.

8.4 Ablation Study

To validate the efficacy of the proposed modules, we conducted an ablation study under both with and without learning from experience setting. Initially, we assessed the effectiveness of the Improving Strategy Module (IS), the Analysis of Others' Strategies Module (AO), and the Analysis Module (AM) within the context of the learning from experience setting, wherein strategies were updated based on accumulated gameplay for both our agents and the baseline agents. In this evaluation, the proposed agents engaged in ten games, assuming evil side roles, against the baseline agents for each module. Following these games, the wining rate (WR), quest engagement rate (QER), and the failure voting rate (FVR) were measured and reported for analysis. Table 3 presents the outcomes of the ablation study conducted within the learning-from-experience setting. It is discernible that in the absence of the Improving Strategy module, where the strategy remains static but the agent can still glean insights from other players' strategies, the winning rate decreases by 20%. Additionally, the agents exhibit reduced aggression, indicated by lower quest engagement rates and failure voting rates. Furthermore, the absence of the Analysis of

Summarization:

Within the context of the Avalon game, please assist {Player i} in summarizing the conversations known to him from the current phase. These conversations are structured in JSON format, with "message" signifying the content of the conversation, "name" identifying the speaker, and "message_type" indicating the type of message relevant to {Player i}. Specifically,"public" implies that all players have access to the message, while "private" implies that only {Player i} has access to it.

Conversations: {conversations}.

Analysis:

Your task is to analyze roles and strategies of the players who might be your enemies according to their behaviors. The analysis should be no more than 100 words. The behaviors are summarized in paragraphs.

Your name is {Name} your role is {Role}.

The summary is {Summary}.

Planning:

Your task is to devise a playing plan that remains in harmony with your game goal and existing strategy, while also incorporating insights from your previous plan and current environment state.

{Role Information}

Goal: {Goal}

Strategy: {Strategy}

Your previous plan: {Plan}

Summary of previous rounds: {Summary}

Analysis about other players: {Analysis}.

Action:

Your objective is to make decisions based on your role, your game goal and the current game state. There are five types of actions you can take: choosing players, voting (agree or disagree), performing missions (make missions succeed or fail), using non-verbal signals (raise hands up, put hands down, open eyes, or close eyes), and choosing to remain silent. Only one action type can be selected at a time. If you decide to choose players, you can choose multiple players according to Host's question.

{Role Information}

Goal: {Goal}

Strategy: {Strategy}

Your current plan: {Plan}

Summary of previous rounds: {Summary}

Analysis about other players: {Analysis}.

Host's Instruction: {Instruction}.

Response:

Your task is to provide detailed response to the question of Host, in accordance with the provided actions. Your response should be no more than 100 words.

{Role Information}

Goal: {Goal}

Strategy: {Strategy}

Your current plan: {Plan}

Summary of previous rounds: {Summary}

Host's Instruction: {Instruction}.

current actions: {actions}

Table 1: Input prompts of our proposed different modules.

Others' Strategies module and the Analysis Module also leads to a decline in the winning rate. In these scenarios, the agents adopt a cautious gameplay approach, resulting in significantly lower quest engagement rates but higher failure voting rates.

Following the initial evaluation, we proceeded to assess the effectiveness of the Analysis Module, Planning Module, and Action Module under conditions where learning from experience was not incorporated. In this scenario, strategies were not updated for both our agents and the baseline agent. It is essential to note that the games were conducted independently, with no influence from previous games on future gameplay. Table 4 presents the results from the module ablation study conducted without incorporating learning from experience. It is discernible that the absence of the planning module results in a notable 20% decrease in the winning rate. Additionally, the Assassin exhibits a significantly lower quest engagement rate, indicating a tendency to overlook the mission objective

Self-Role Strategy Learning (Step 1)

Your task is to provide 3 suggestions for {player}'s playing strategy of the role {role} in Avalon games, according to the game log. The game log includes the summaries of different rounds of a game.

The roles of the players: {player-role mapping}

The summaries of a round game: {summary}

{player}'s game goal: {goal}

{player}'s playing strategy of role {role}:{current strategy}

Previous suggestions: {suggestions from last game}

Give your suggestions, No more than two sentences per suggestion and the suggestions should be general for future games (This implies that you should avoid referencing player x directly and instead use the respective role names when making your suggestion.) and effectively help him achieve his game goal in future games.

Self-Role Strategy Learning (Step 2)

Your task is to help {player} improve his playing strategy of the role {role} a Avalon game with suggestions.

{player}'s strategy: {current strategy}

Suggestions: {suggestions}

Please improve the strategy while retaining the advantages of the original strategy for him and the strategy should be no more than 2 sentences. Describe the strategy you provide using continuous sentences rather than bullet points or numbering.

Other-Role Strategy Learning (Step 1)

Your task is to help {player} analyze the strategies of other players in a Avalon game, according to the game log. The game log is summarized in paragraphs.

The roles of the players: {player-role mapping}

The summaries of rounds of the game: {summary}

Previous strategies of other roles: {previous strategies}

Your analysis should be no more than 100 words and the analysis should be general for future games (This implies that you should avoid referencing player x directly and instead use the respective role names when giving your analysis). And analyze together with previous strategies.

For example: The strategy of Merlin is that ... The strategy of Assassin is that... The strategy of ... is ...

Other-Role Strategy Learning (Step 2)

There are experience of previous games provided:

Suggestions from previous games: {suggestion}

Strategies of other roles from previous games: {other strategy}

Table 2: Input prompts of our experience learning module.

Method	WR(%)	QER(%)		FVR(%)	
		Morgana	Assassin	Morgana	Assassin
ours	80	44.1	49.1	66.6	78.5
w/o. IS	60	42.8	39.3	46.1	100
w/o. AO	70	18.3	8.3	100	100
w/o. AM	50	29.3	39	87.5	100

Table 3: Ablation Study on Experience Learning: Compare of full framework, without improving strategy (IS), without analysis strategies of others (AO) and without analysis module (AM).

without the guidance of a strategic plan. This underscores the critical importance of the planning module in ensuring that agents consistently progress toward winning the game.Furthermore, in the absence of both the analysis and action modules, the agents exhibit a slightly lower quest engagement rate. Despite this, they manage to maintain an impressive 80% winning rate.

In the final phase of our evaluation, we scrutinized the impact of analysis on all players, teammates and adversaries. In each configuration, our agents assumed the roles of the evil side in ten games,

Method	WR(%)	QER(%)		FVR(%)	
		Morgana	Assassin	Morgana	Assassin
all modules	90	55.5	58.3	93.7	100
w/o analysis	80	44.1	47.5	100	100
w/o. plan	60	55	16.6	90	100
w/o. action	80	45.6	45.6	100	100

Table 4: Module Ablation: under the setting without learning from experience.

Method	WR(%)	QER(%)		FVR(%)	
		Morgana	Assassin	Morgana	Assassin
all players	90	55.5	58.3	93.7	100
teammates only	80	26.8	48.1	62.5	100
adversaries only	90	38.3	45.3	92.3	100

Table 5: Analysis Module Ablation: under the setting without learning from experience. Analyzing different objects.

facing off against baseline agents aided by corresponding analysis information. The results, encompassing winning rate, quest engagement rate, and failure voting rate, are tabulated in Table 5. It becomes apparent that when analysis information is restricted

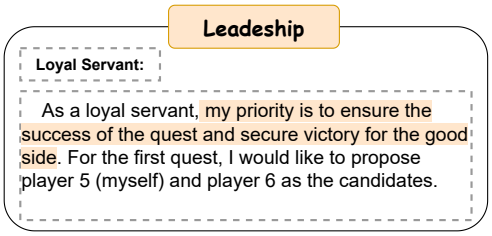


Figure 12: Leadership example

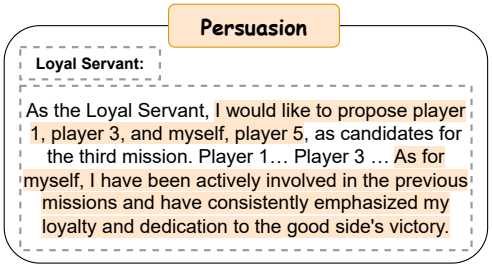


Figure 9: Persuasion example

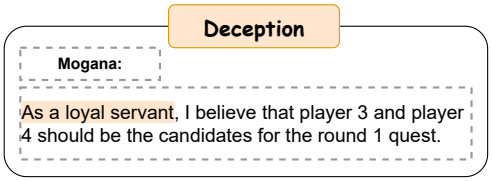
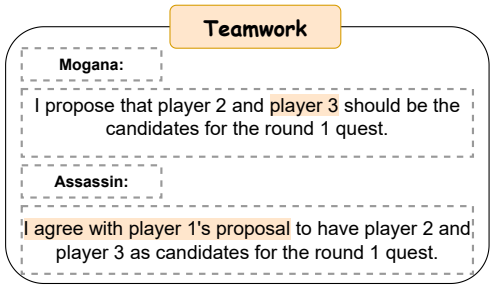


Figure 10: Deception example



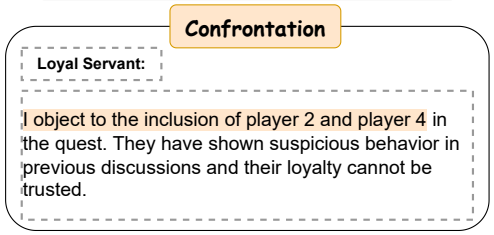


Figure 11: Teamwork and confrontation examples

solely to teammates, the winning rate declines by 10%. In response, our proposed AI agents adopt a less aggressive approach, evident in reduced quest engagement rates and failure voting ratings. However, when analysis information pertains exclusively to adversaries, there is a decrease in quest engagement rates while retaining the winning rate and failure voting rate. This phenomenon can be attributed to the strategic advantage gained by the Assassin, who can identify Merlin with the aid of analysis information on adversaries. Consequently, the analysis of adversaries proves to be paramount for the evil side's victory in Avalon games for AI agents.

9 CASE STUDY

In Figures 9, 10, 11 and 12, we present examples to show how the AI agents perform the social behaviors in the Avalon games.