

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

Anonymous ACL submission

## Abstract

This paper explores the open research problem of understanding the social behaviors of LLM-based agents. Using Avalon as a testbed, we employ system prompts to guide LLM agents in gameplay. While previous studies have touched on gameplay with LLM agents, research on their social behaviors is lacking. We propose a novel framework, tailored for Avalon, features a multi-agent system facilitating efficient communication and interaction. We evaluate its performance based on game success and analyze LLM agents’ social behaviors. Results affirm the framework’s effectiveness in creating adaptive agents and suggest LLM-based agents’ potential in navigating dynamic social interactions. By examining collaboration and confrontation behaviors, we offer insights into this field’s research and applications.

## 1 Introduction

Artificial intelligence (AI) agents (Xi et al., 2023; Park et al., 2023) exhibit human-like behaviors, from perceiving and analyzing the environment to decision-making and action-taking.

Advances in large language models (LLMs) (Kasneci et al., 2023; Peng et al., 2023; Touvron et al., 2023; Vaswani et al., 2017) offer new avenues for creating AI agents in complex environments, potentially simulating human society. Various works (Gao et al., 2023; Qian et al., 2023; Park et al., 2023; Ghaffarzadegan et al., 2023) simulate different aspects of human society. For instance, Qian et al. (Qian et al., 2023) simulate a software development company with agents representing diverse social identities. Park et al. (Park et al., 2023) assign varied social roles to agents within a sandbox environment. However, prior studies mostly examine positive social behaviors like honesty and collaboration, leaving research on negative social behaviors of LLM agents relatively scarce.

Previous research on human society has highlighted issues like misinformation and online con-

flicts, leading to efforts to address these problems (Song and Jiang, 2022; Levy et al., 2022; Chen et al., 2022). To delve deeper into the social behaviors of LLM agents, we intend to comprehensively investigate both positive and negative aspects of their conduct. To achieve this, we employ Avalon as the environment to illustrate collaboration and confrontation among agents. Avalon, a representative communication game, assigns players hidden roles and divides them into opposing teams. Throughout gameplay, players partake in discussions, debates, and strategic maneuvers.

LLM agents face a challenging task in winning the incomplete information game of Avalon. They need to share and obtain information via communication and analysis, deducing other players’ roles, building trust among allies, and deceiving opponents. Success requires technical abilities like natural language understanding, incomplete information analysis, and strategy learning. Additionally, social behaviors such as teamwork, persuasion, and deception are crucial 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 (Xu et al., 2023a), 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:

- We explore the social behaviors exhibited by LLM-based agents in the context of Avalon

042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082

001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041

Method	Memory	Analysis	Plan	Action	Experience Learning	Leadership	Persuasion	Camouflage	Teamwork	Confrontation	Sharing
GenAgents (Park et al., 2023)	✓		✓	✓	✓				✓		✓
Plan4MC (Yuan et al., 2023)			✓	✓							
GITM (Zhu et al., 2023)	✓		✓	✓							
RGAgent (Akata et al., 2023)	✓								✓	✓	
CGAgent (Xu et al., 2023a)	✓	✓			✓	✓		✓	✓	✓	
ReCon (Wang et al., 2023c)	✓	✓							✓		
LARL (Xu et al., 2023b)	✓	✓					✓		✓		
CodeAct (Shi et al., 2023)	✓	✓		✓					✓		
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between our work and related works in both agent framework and social behaviour analysis.

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 growing influence of Large Language Models (LLMs) on society has spurred significant research (Movva et al., 2023). Innovations include using LLMs for virtual social network simulations to advance social science research (Gao et al., 2023) and enrich human social experiences in virtual spaces (Kaiya et al., 2023). However, concerns arise regarding validity, privacy, and ethics in LLM-driven social computing. Ghaffarzadegan et al. propose feedback mechanisms to address these concerns (Ghaffarzadegan et al., 2023). Additionally, LLMs fuel advancements in social robot development (Yang and Menczer, 2023), posing challenges like social bot detection and misinformation spread. Ongoing research aims to align LLMs with ethical standards, mitigate biases and errors, and ensure their reliable and ethical use across diverse applications (Wang et al., 2023d; Liu et al., 2023).

### 2.2 LLM-Based Gameplay

The rapid development of LLM-based agents has resulted in significant advancements in problem-solving across various domains. These agents, known for their quick and strategic processing,

have improved the effectiveness and robustness of solving tasks (Lin et al., 2023; Wang et al., 2023b; Tsai et al., 2023; Zhou et al., 2023; Park et al., 2023; Qian et al., 2023; Fu et al., 2023).

LLMs have recently been utilized in various gaming environments, including task-based games like Minecraft and multiplayer strategy games (Yuan et al., 2023; Zhu et al., 2023; Wang et al., 2023a; Akata et al., 2023; Xu et al., 2023a; Wang et al., 2023c). In multiplayer strategy games such as the Prisoner’s Dilemma and Battle of the Sexes, LLMs model strategic interactions (Akata et al., 2023). They’re also employed in social deduction games like Werewolf and Avalon (Xu et al., 2023a; Wang et al., 2023c; Shi et al., 2023; Xu et al., 2023b), where they exhibit strategic behaviors. To combat misinformation, recursive contemplation has been proposed (Wang et al., 2023c). However, previous works have only partially analyzed behaviors and designed agent frameworks based on limited game characteristics. Thus, we propose a comprehensive communicative game agent framework based on LLMs and conduct a thorough behavior analysis. Table 1 illustrates the distinctions between our work and others.

## 3 Background

In our study, we chose Avalon, also known as “The Resistance”, instead of Werewolf as our environment. Unlike Werewolf, where players are gradually eliminated, Avalon ensures that all players remain engaged throughout the game, promoting social cohesion.

Avalon accommodates 5 to 10 players, focusing on the 6-player variant herein. Players receive secret roles in either the good or evil faction. The good faction includes Merlin, Percival, and Loyal Servants, while the evil faction comprises Morgana and Assassin. Morgana and Assassin know each other’s identities, Percival can identify Merlin and Morgana, and Merlin recognizes all evil players. The game spans 3-5 rounds. Players discuss and vote to form a quest team of 2-3 members. Ap-

proval requires a majority vote; otherwise, leadership shifts. Each round allows up to five voting cycles before the leader selects the team. Quest success hinges on cards submitted by team members. Good players submit success cards, while evil players can choose success or failure cards. A quest fails if it receives a failure card. The game concludes with victory for good players if three quests succeed, or for evil players if three quests fail. Evil players can also win by correctly identifying Merlin at the game’s end.

### 3.1 Social Behaviors in Avalon

**Teamwork.** Good players must collaborate to complete quests for winning. They should build trust with teammates while being wary of evil players.

**Leadership.** Each player has the chance to lead the discussion for forming the quest team. The leader can guide the conversation and build trust among players. Effective leadership is crucial for 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 pretend to be good players, using deceptive tactics and concealing information to mislead others.

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

**Sharing.** Each role has unique clues. Sharing these clues promotes collaboration and builds trust among players, but risks exposing one’s identity.

## 4 Approach

### 4.1 Setup

Figure 1 shows the proposed framework. All prompts used are shown in Appendix Table 4. To start the game, system prompts are used to assign different roles to LLM agents. Each system prompt for a role  $p_i$  includes 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 provide information about the assigned role to the LLM agent, while the Goal (Winning Conditions) offers insights into how to achieve victory. Additionally, 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 teammates and the Assassin.*

**Goal:** *Win the game by intentionally causing quests to fail for three rounds, alone or with 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

Analyzing game history is vital for agents to grasp the current situation and make decisions. Yet, in Avalon, LLM agents’ history responses are often too lengthy, surpassing input limits and potentially lowering performance. To tackle this, a memory storage system is introduced to record conversations among LLM agents, enabling subsequent analysis and decision-making.

#### Memory Storage.

Memory storage is vital for recording agents’ conversation history in the current game round. It comprises structured memory objects containing key details like role name, detailed natural language responses, round number, and a flag indicating public or private status. Public information is visible to all roles, while private information pertains to each role’s conversation. We assign separate memory pools to each agent for clarity in information processing. By storing this data, memory storage enables agents to access and review past conversations, improving their understanding of the game’s progress.

### 4.3 Memory Summarization.

To store more information in memory, we use a summarization prompt to compress the information from the previous round and capture the essential details. The process of updating the memory with a summary of the previous round is illustrated below:

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

The memory on round  $t$  is  $\mathcal{M}_t$ . The response generated by the LLM for role  $p_i$  on round  $t$  is  $\mathcal{R}_t^{p_i}$ , and  $\mathcal{I}_t$  represents the instructions and statements of the host on round  $t$ .  $\langle \rangle$  is Text concatenation.  $\text{SMR}(\cdot)$  is the summarization prompting.

### 4.4 Analysis

To help LLM agents improve strategic planning and increase their chances of winning, we introduce an analysis module. This module analyzes the role

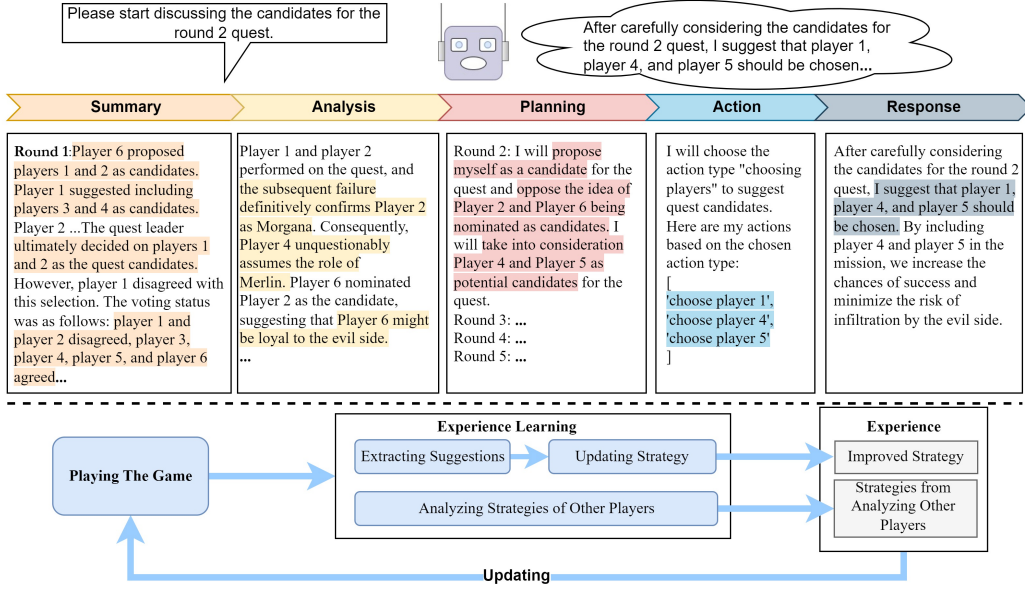


Figure 1: Our framework has six modules: summary, analysis, planning, action, response, and experiential learning. This design follows human thinking, helps LLM agents play Avalon effectively, and reveals their social behaviors.

identity and potential strategies of other players during gameplay:

$$\mathcal{H}_t^{p_i} = \text{ANA}(\mathcal{M}_t, \mathcal{R}\mathcal{I}^{p_i}), \quad (2)$$

where  $\mathcal{M}_t$  is the memory on round  $t$  and  $\mathcal{R}\mathcal{I}^{p_i}$  is the role information. By analyzing, LLM agents can better understand their collaborators and competitors, leading to improved decision-making and effective counterstrategies for winning.

#### 4.5 Planning

Agents need to understand the game progress and necessary strategies to win. Thus, a planning module is designed to create a strategic plan. The plan is based on the memory and information from the current round of the game, as described below:

$$\mathcal{P}_t^{p_i} = \text{PLAN}(\mathcal{M}_t, \mathcal{H}_t^{p_i}, \mathcal{P}_{t-1}^{p_i}, \mathcal{R}\mathcal{I}^{p_i}, \mathcal{G}^{p_i}, \mathcal{S}^{p_i}), \quad (3)$$

where  $\mathcal{P}_t^{p_i}$  represents the strategic plan of agent  $p_i$  at round  $t$ .  $\mathcal{G}^{p_i}$  and  $\mathcal{S}^{p_i}$  are goals and initial strategies. By creating a strategic plan, the agents can have a flexible strategy for different situations. This foresight helps them make better decisions about collaborating with teammates, deceiving opponents, taking on the opposing faction's identity, and, if needed, sacrificing teammates or oneself to secure winning in the game.

#### 4.6 Action

In the action module, agents decide their next action based on memory information, situation analysis, and the strategic plan. There are five types

of actions: selecting players, voting (agree or disagree), completing quests (succeed or fail), using non-verbal signals (raising hands, putting hands down, opening or closing eyes), and choosing to remain silent. The process of choosing the next action is as follows:

$$\mathcal{A}_t^{p_i} \sim p(\mathcal{A} | \mathcal{M}_t, \mathcal{H}_t^{p_i}, \mathcal{P}_t^{p_i}, \mathcal{R}\mathcal{I}^{p_i}, \mathcal{G}^{p_i}, \mathcal{S}^{p_i}, \mathcal{I}_t). \quad (4)$$

The subsequent action depends on the memory, the comprehensive analysis, the strategic plan, and the instruction from the host. The details of these action decisions are confidential and only known to the respective agent. The host and other players cannot see these decisions.

#### 4.7 Response Generation

The Response Generation module is responsible for generating a response to the host's inquiry. Agents in this module choose an action and provide an explanation to the host. Agents are given the freedom to collaborate, deceive, and assume the identity of the opposite faction in their explanations.

#### 4.8 Experience Learning

In practical scenarios, players can improve their Avalon gameplay strategy through experience. They gain insights not only from their own perspective but also by observing other players' strategies. An ideal Avalon LLM agent should learn from both its own experiences and those of other players.

### 4.8.1 Self-Role Strategy Learning

In Step 1, agents generate three strategic recommendations for a player’s role-specific gameplay in Avalon games based on the game history. Agents avoid mentioning specific players and instead use role names to make the suggestions applicable in future games. In Step 2, agents enhance their strategies by incorporating the gathered suggestions while maintaining the original strategy’s strengths.

### 4.8.2 Other-Role Strategy Learning

Avalon LLM agents summarize the strategies adopted by other players to facilitate learning from the strategies employed by other players. Prompts for the above steps are presented in Appendix Table 5.

## 5 Experiment

### 5.1 Implementation Details

We developed the Avalon game program in Python, using the gpt-3.5-turbo-16k model as both our backend and the baseline’s. In all experiments, we set the agent model’s temperature to 0.3 and the LLM extractor’s to 0. The number of suggestions generated for updating strategies is 3. Game rules and role descriptions were set according to the baseline template (Xu et al., 2023a), which leverages historical context, enhances agent reasoning, and learns from past mistakes. Detailed descriptions are provided in Section A.2.

For broader validation, we implemented our framework on the Llama2-7b-chat-hf model. However, LLaMA-based agents face constraints due to the model’s language understanding capabilities and token limitations. Preliminary exploration without further analysis is discussed in Section C.

### 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 is the percentage of games won out of the total played, calculated by dividing the number of wins by the total games played:

$$WR = \left( \frac{\#Wins}{\#Games\ Played} \right) \times 100\% \quad (5)$$

**Quest Engagement Rate (QER).** "Quest engagement rate" is the ratio of rounds a player joins the quest team to the total rounds played in the games. It’s calculated as follows:

$$QER = \left( \frac{\#Engagement\ Rounds}{\#Rounds} \right) \times 100\% \quad (6)$$

**Failure Vote Rate (FVR)** The quest result relies on success or failure cards from team members. The failure vote rate indicates the percentage of votes against quest success, calculated as follows:

$$FVR = \left( \frac{\#Failure\ Votes}{\#Votes} \right) \times 100\% \quad (7)$$

### 5.2.2 Social Behaviors.

From this perspective, we use ChatGPT to assist the analysis on the social behaviors of agents.

**Leadership.** We gauge AI agents’ leadership using "Leader Approval Rate (LAR)". LAR is calculated by dividing total approval votes by total leader votes across 20 Avalon games. It reflects consensus among players on proposed quest teams.

**Persuasion.** To evaluate LLM agents’ persuasion, we track two metrics: self-recommendation rate (proposing oneself for quests) and success rate (self-recommendation for quest participation).

**Deception.** Detecting deception in AI agents is challenging. We focus on identifying instances where agents assume different identities in the initial round of each game. Behaviors include Self-Disclosure, Camouflage, and Withholding Identity.

**Teamwork and Confrontation.** We use ChatGPT to analyze role responses, aiming to identify instances of collaboration or confrontation. ChatGPT prompts with a player’s response and evaluates trust (teamwork), lack of trust (confrontation), or ambivalence towards others.

**Sharing.** Sharing reflects how often agents disclose valuable information, crucial for team cooperation. Using ChatGPT, we analyze agents’ dialogues to identify instances of sharing behavior, aiming to quantify their willingness to share for the team’s benefit.

### 5.3 Experiment Results

To validate the efficacy of Avalon AI agents, we repurposed Werewolf AI agents (Xu et al., 2023a) as baselines. Across two sets of 10 consecutive Avalon games, our agents faced off against the baselines, with Evil versus Good and vice versa. After the matches, we compared the winning rates

Method	Good Side	Evil Side
Ours	90	100
w/o analysis	60	60
w/o plan	80	100
w/o action	100	80
w/o strategy learning	50	60

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

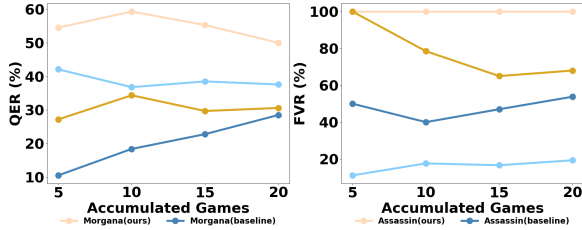


Figure 2: (a): Comparison of the engaging quests rate when playing evil side. Higher engaging quests rate means more opportunities for the player to influence the outcome of the game. (b): Comparison of the failure vote rate when playing evil side. Baseline is worse.

of our Avalon AI agents to the baselines. As depicted in Table 2, our method demonstrated a 90% winning rate in 10 games when playing the good side. Conversely, when playing the evil side, the winning rate was 100% over the same number of games.

Ablation studies reveal the importance of key modules in our AI agents. Removing the analysis module lowered winning rates to 60% for both sides, showing its impact on understanding and decision-making. Excluding the planning module reduced the good side’s winning rate to 80%, highlighting its role in devising strategies. Without the action module, the good side won 100% while the evil side dropped to 80%, indicating its importance for the evil side’s success. Removal of the strategy learning module led to winning rates decreasing to 50% and 60% for good and evil respectively, emphasizing its role in enhancing strategies. In conclusion, the analysis and strategy learning modules significantly influence game outcomes, affecting both sides’ winning rates. Additionally, the planning and action modules are crucial for success, given their impact on gameplay.

To better grasp the strategies employed by our Avalon Agents and the baseline agent, we compared quest engagement and failure voting rates when different AI agents acted as the evil side. Both rates significantly impact game outcomes. A higher quest engagement rate allows more chances

for players to influence the game, while a higher failure voting rate suggests a greater chance for the evil side to win but also increases the risk of exposure, indicating an aggressive gameplay approach. Figure 2 illustrates the outcomes for quest engagement and failure voting rates. Our AI agents, particularly when playing as Morgana and Assassin, show assertiveness, with a 40.3% quest engagement rate and 84.0% failure voting rate. In comparison, baseline agents have lower rates at 33.1% and 36.5% respectively. As a result, our proposed Avalon AI agents achieve a 100% win rate against the baseline agents when playing as the evil side.

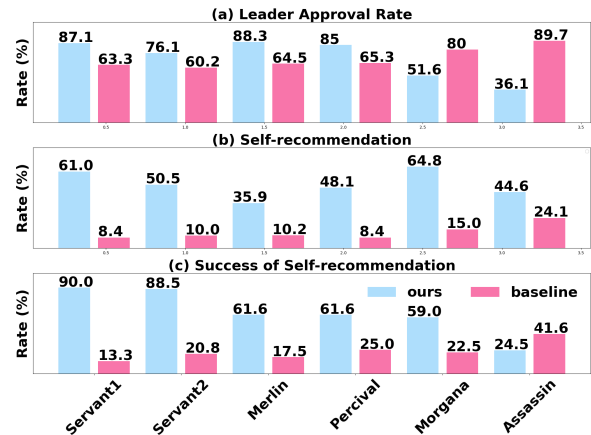


Figure 3: (a): The leadership behavior. Players with higher Leader Approval Rate get more agreements from other players when deciding a quest team. (b) and (c): The persuasion behavior. Self-recommendation Rate: players with higher Self-recommendation Rate are more will to engage in quests. Self-recommendation Success Rate: players more likely to gain the trust of other players has higher Self-recommendation Success Rate.

## 6 Social Behaviors of AI Agents

To evaluate if AI agents replicate human social behaviors in Avalon, we conduct a thorough analysis. This involves assessing the agents’ execution of teamwork, leadership, persuasion, deception, and confrontation through the frequency distribution in game logs from two sets of 20 consecutive Avalon games, shown 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

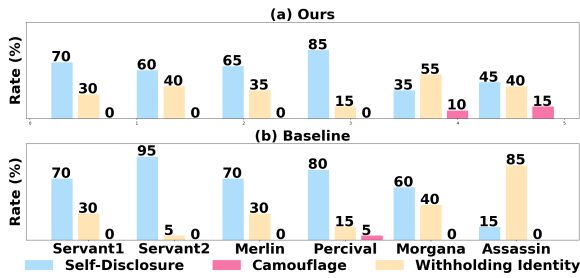


Figure 4: The deception behavior when playing different roles: at first round of each game, the distribution of the players choose Self-Disclosure, Camouflage or Withholding Identity.

crucial for the good side to effectively counter the deceptive tactics employed by the evil side.

Figure 3 (a) illustrates the Leader Approval Rate when agents assume various roles. It is evident that our agents, playing on the good side, attain remarkably high Leader Approval Rates when serving as leaders. Notably, the AI agents achieve a Leader Approval Rate exceeding 80% averagely while undertaking roles associated with the good side. This signifies their robust leadership qualities and their proactive approach to steering the gameplay towards victory. However, the baseline agents could propose good side players to the quest team to achieve high Leader Approval Rate but low game win rate.

## 6.2 Persuasion

Figure 3 displays the evaluation outcomes assessing the AI agents' persuasion ability. Notably, agents employ distinct strategies based on their assumed roles, as shown in Figure 3 (b). When playing as Loyal Servant and Morgana, agents display a high self-recommendation rate for quest team participation, impacting mission success. Conversely, a cautious approach is seen with roles like Merlin, Percival, and Assassin, evident from their low self-recommendation rates. This strategic restraint is crucial, particularly for roles like Merlin, emphasizing the importance of concealing identity. From Figure 3 (c), Loyal Servants exhibit higher success rates in self-recommendation compared to roles that easily raise suspicion. Additionally, the proposed Avalon Agents show higher rates of self-recommendation and greater success compared to baseline agents, indicating enhanced persuasion abilities.

## 6.3 Deception

Deception is central to Avalon. Evil roles must deceive loyal servants while subtly sabotaging missions. Skilled players create elaborate lies and misdirection. Loyal servants also engage in deception to conceal their identities, especially when under suspicion.

In Figure 4, 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 and the Assassin display rates of assuming alternate identities of 10% and 15%, respectively, 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 vital for loyal servants to identify each other and succeed in missions by strategizing, discussing assignments, and sharing information to uncover evil roles. Confrontations arise when suspicions lead to accusations, resulting in intense exchanges where accusers present reasoning and the accused offer defenses or deflect suspicion onto others.

In Figure 5 (a), teamwork and confrontation rates of good side roles are depicted. Loyal Servants tend to avoid confrontation due to their lack of specific identity information. However, Merlin, aware of Morgana and Assassin, confronts them frequently. Percival, aware of Merlin and Morgana without knowing their exact identities, confronts both. These observations highlight the adaptive strategies of AI agents, mirroring the social dynamics of human players in Avalon.

Figure 5 (b) shows teamwork and confrontation rates of baseline agents. Rates remain consistent across roles, suggesting they do not adjust strategies based on role assumptions.

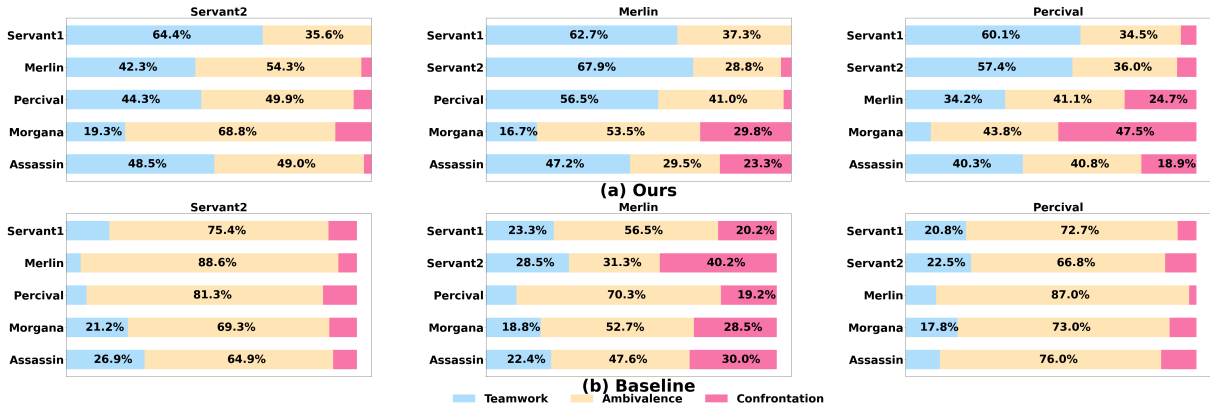


Figure 5: The teamwork and confrontation behaviors when playing different roles. Each subfigure shows the attitude distribution of the player portraying specific role (on the top) towards players in other roles (on the left).

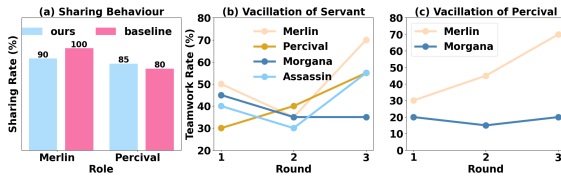


Figure 6: (a): The sharing behavior when playing Percival and Merlin at the first round. (b) and (c): The teamwork vacillation between different rounds.

## 6.5 Sharing

Sharing is essential for Percival and Merlin. They possess more information than other good roles, and sharing their insights aids in winning the game. However, excessive sharing of known information may also benefit the opposing side, as discussions are public to all players. Therefore, strategic sharing of information is necessary to win the game.

Figure 6 (a) depicts the proportion of known information shared with other players by different agents playing the roles of Merlin and Percival in the first round of the game. It is observed that both the agents designed by us and the baseline agents exhibit an excessive level of sharing behaviors.

## 6.6 Vacillation

At the game’s onset, some players possess identity clues, like Percival knowing Morgana and Merlin without distinction, while others, like Loyal Servants, lack such info. Both situations require players to deduce identities for their camp’s benefit. Analyzing teamwork proportions across rounds reveals players’ ability to discern allies and foes.

Figure 6 (b) illustrates Loyal Servants’ teamwork tendencies, while (c) shows Percival’s tenden-

cies towards Morgana and Merlin. Throughout the game, players increasingly collaborate with teammates and less with enemies. However, Loyal Servants face greater challenges inferring roles, leading to higher teamwork with potential foes.

## 6.7 Behavior Spontaneity

Teamwork and confrontation behaviors of players arise spontaneously due to game mechanics fostering interaction and competition. Teamwork aids in identifying evil roles, facilitating successful quests. However, teamwork often brings confrontation, as doubts about role identities persist. Even without strategic learning mechanisms, players exhibit these behaviors, showing their spontaneous nature. However, behavior distributions vary significantly between agents with and without strategic learning. The relevant analysis is provided at the Section D.

## 7 Conclusion

This paper explores the social behaviors of LLM-based agents in the Avalon communication game. We introduce a multi-agent framework facilitating efficient communication and interaction. This framework includes memory, analysis, planning, action, and response modules capable of learning from experience. Unlike prior studies, our research delves into the social dynamics of these agents in gameplay scenarios. Our evaluation showcases the success of our framework in achieving winning strategies and the adaptability of LLM agents in complex social interactions. Future work involves optimizing our approach, exploring its applicability in diverse game environments, and further understanding LLM agents’ potential in dynamic social interactions.



## 8 Limitations

Although the LLM agent framework we proposed has performed well in the Avalon game, there are also limitations of high cost and slow interaction speed, due to multiple accesses to the model required for each interaction. Additionally, from the behaviors exhibited by the agent, there are also instances of unreasonable behavior distribution, such as excessive self-disclosure actions. In the future, we will explore and improve these aspects.

## References

- 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.
- 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*, pages 2636–2645.
- Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. [Improving language model negotiation with self-play and in-context learning from ai feedback](#).
- 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.
- 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.
- 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](#).
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- 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*, pages 2592–2602.

- 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.
- 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.
- 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.
- 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](#).
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- 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.
- Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. [Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game](#).
- Qiurong Song and Jiepu Jiang. 2022. How misinformation density affects health information search. In *Proceedings of the ACM Web Conference 2022*, pages 2668–2677.
- 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*.
- 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*.
- 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.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. [Voyager: An open-ended embodied agent with large language models](#).

715 Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao  
716 Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang,  
717 Xu Chen, Yankai Lin, et al. 2023b. A survey on large  
718 language model based autonomous agents. *arXiv*  
719 *preprint arXiv:2308.11432*.

720 Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi,  
721 Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei  
722 Wang, Shiji Song, and Gao Huang. 2023c. [Avalon's](#)  
723 [game of thoughts: Battle against deception through](#)  
724 [recursive contemplation](#).

725 Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi,  
726 Xingshan Zeng, Wenyong Huang, Lifeng Shang,  
727 Xin Jiang, and Qun Liu. 2023d. [Aligning large](#)  
728 [language models with human: A survey](#). *ArXiv*,  
729 [abs/2307.12966](#).

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

735 Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xi-  
736 aolong Wang, Weidong Liu, and Yang Liu. 2023a.  
737 [Exploring large language models for communication](#)  
738 [games: An empirical study on werewolf](#).

739 Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu.  
740 2023b. [Language agents with reinforcement learning](#)  
741 [for strategic play in the werewolf game](#).

742 Kai-Cheng Yang and Filippo Menczer. 2023. [Anatomy](#)  
743 [of an ai-powered malicious social botnet](#). *ArXiv*,  
744 [abs/2307.16336](#).

745 Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang  
746 Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023.  
747 [Plan4mc: Skill reinforcement learning and planning](#)  
748 [for open-world minecraft tasks](#).

749 Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. 2023.  
750 [Llm as dba](#).

751 Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Wei-  
752 jie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu,  
753 Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and  
754 Jifeng Dai. 2023. [Ghost in the minecraft: Gener-](#)  
755 [ally capable agents for open-world environments via](#)  
756 [large language models with text-based knowledge](#)  
757 [and memory](#).

758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807

## A Appendix

### A.1 Avalon Introduction

Avalon is designed for 5 to 10 players. Specifically, we focus on the 6-player variant of the game.

**Player roles.** Roles including Merlin, Percival, Morgana, Assassin, and two Loyal Servants, are divided into good and evil sides. Merlin, Percival, and loyal servants are on the good side, while Morgana and Assassin are on the evil side. Players are assigned roles secretly, with some having special abilities. Morgana and Assassin are initially aware of each other. Percival is able to see Merlin and Morgana but does not know their exact identities. Merlin is aware of the identities on the evil side.

**Quest team assignment.** After receiving roles, players engage in 3-5 rounds of discussion and voting for a certain number of players to form a quest team. At the start of each round, a leader is assigned in rotation. The leader hosts a discussion, followed by a public vote on quest team members. If more than half of the votes agree, the team forms; otherwise, leadership rotates to the next player for further discussion and voting. Each round allows up to five discussion and voting cycles, with the leader directly assigning team members after the fifth round.

**Quest phase.** The quest outcome is determined by the cards submitted by the quest team. Good players can only submit success cards, while evil players can choose to submit either success or failure cards. A quest is successful if all team members vote for success, and fails if one or more members vote for failure.

**End of the game.** The game ends when three quests succeed (good side wins) or three quests fail (evil side wins). Additionally, the evil players can win by correctly identifying Merlin at the end.

### A.2 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

808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858

859 quest succeed, good side will  
860 get one point, otherwise, evil  
861 side will get one point. Which  
862 side get 3 points earlier, which  
863 side wins the game. If you are  
864 Assassin, at the end of a round,  
865 you can choose to identify  
866 which one is Merlin, if the  
867 identifying is successful, the  
868 red camp directly win the game.  
869 If not successful, the Assassin  
870 will expose his identification.  
871 Objectives: your goal is to  
872 help your side get 3 points  
873 and win the game. If you are  
874 Assassin, you also need to  
875 reason which player is Merlin  
876 as early as possible.  
877 Tips: To complete the objective:  
878 you should analyze and use your  
879 ability correctly. During  
880 quest phase, you need to  
881 reason carefully about the  
882 roles of other players and be  
883 careful not to reveal your own  
884 role casually unless you're  
885 cheating other players. Only  
886 give the player's name when  
887 making a decision/vote, and  
888 don't generate other players'  
889 conversation. Reasoning based  
890 on facts you have observed and  
891 you cannot perceive information  
892 (such as acoustic info) other  
893 than text. You are {player},  
894 the {role}. You're playing  
895 with 5 other players. Do not  
896 pretend you are other players or  
897 the moderator. Always end your  
898 response with '<EOS>'.

### 899 A.3 Module Prompts

900 Our designed prompts for different modules are  
901 presented in Tables 4 and 5.

### 902 A.4 Heuristic Rules for LLM Gameplay

903 In the gameplay, we used LLM to extract infor-  
904 mation from the responses of the agents. For ex-  
905 ample, when the agent selects a player, it extracts  
906 the player number, and when voting, it extracts the  
907 player's voting result. With several demonstrations  
908 of how to extract corresponding information, LLM

909 can extract information very accurately to help the  
910 game proceed smoothly. Table 3 shows some cases  
911 of extraction.

912 It is observed agents sometimes may fail to an-  
913 swer questions correctly, such as voting with un-  
914 clear attitudes. In order to allow the game to pro-  
915 ceed smoothly, we design the following heuristic  
916 rules. When voting for quest candidates, if the  
917 agent's answer is unclear, we assume that it agrees.  
918 When voting the quest for success or failure, if the  
919 agent's answer is unclear, we default to it voting for  
920 failure. When agents select an excessive number  
921 of players, we truncate the selection to meet the  
922 quest's requirements. In cases where the agents  
923 choose too few players, the host will repeat ques-  
924 tion to the agent. If the required player count is  
925 still not met even after multiple retries, the program  
926 steps in to assist by making a random selection on  
927 behalf of the agent.

### 928 A.5 Ablation Study

929 To validate the efficacy of the proposed modules,  
930 we conducted an ablation study under both with  
931 and without learning from experience setting. Ini-  
932 tially, we assessed the effectiveness of the Improv-  
933 ing Strategy Module (IS), the Analysis of Others'  
934 Strategies Module (AO), and the Analysis Module  
935 (AM) within the context of the learning from ex-  
936 perience setting, wherein strategies were updated  
937 based on accumulated gameplay for both our agents  
938 and the baseline agents. In this evaluation, the  
939 proposed agents engaged in ten games, assuming  
940 evil side roles, against the baseline agents for each  
941 module. Following these games, the winning rate  
942 (WR), quest engagement rate (QER), and the fail-  
943 ure voting rate (FVR) were measured and reported  
944 for analysis. Table 6 presents the outcomes of the  
945 ablation study conducted within the learning-from-  
946 experience setting. It is discernible that in the ab-  
947 sence of the Improving Strategy module, where the  
948 strategy remains static but the agent can still glean  
949 insights from other players' strategies, the winning  
950 rate decreases by 20%. Additionally, the agents ex-  
951 hibit reduced aggression, indicated by lower quest  
952 engagement rates and failure voting rates. Further-  
953 more, the absence of the Analysis of Others' Strate-  
954 gies module and the Analysis Module also leads to  
955 a decline in the winning rate. In these scenarios,  
956 the agents adopt a cautious gameplay approach,  
957 resulting in significantly lower quest engagement  
958 rates but higher failure voting rates.

---

**Selection Case:**

Please extract the answer to the question from the answer.

Question: Please start discussing the candidates for the turn 1 quest (please explain the reasons for your choice appropriately). The number of players in this quest is 2.

Answer: I suggest that for the turn 1 quest, we consider including both player 1 and player 2. Player 1 has consistently shown loyalty and dedication, and player 2 strongly implied their role as Morgana during the reveal phase. By including both players, we can increase our chances of successfully completing the quest. Thank you.

**Extractor:**

1, 2

---

**Voting Case:**

Complete the following instructions based on the questions and answers.

Response true, if the answer expresses: 1.obvious agreement to something. 2.obvious decision to do something. 3.obvious expression of making quest successful. 4.obvious yes to something

Response false, if the answer expresses: 1.obvious disagreement to something. 2.obvious decision not to do something. 3.obvious expression of making quest failed. 4.obvious no to something

Response none, if the answer is not related to the question or didn't express a obvious attitude to something.

Question: The quest leader decides that the player 1player 3, player 5 will engage in the quest. You can choose to agree or object. Please vote to agree or disagree.

Answer: I vote to disagree.

**Extractor:**

False

---

Table 3: Cases of LLM-based extraction

---

**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 4: Input prompts of our proposed different modules.

---

### **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**

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 ...

---

Table 5: Input prompts of our experience learning module.

Method	WR(%)	QER(%)		FVR(%)	
		Morgana	Assassin	Morgana	Assassin
full	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 6: Ablation Study on Experience Learning: Compare of full framework, without improving strategy (IS), without analysis strategies of others (AO) and without analysis module (AM).

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 7: Module Ablation: under the setting without learning from experience.

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 7 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 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

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

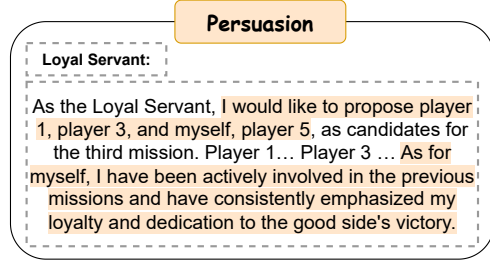


Figure 7: Persuasion example

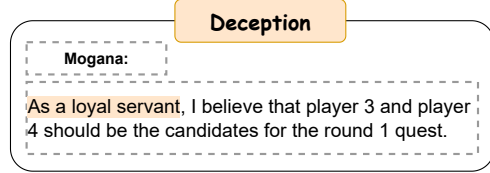


Figure 8: Deception example

games, 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 8. It becomes apparent that when analysis information is restricted 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.

## B Case Study

In Figures 7, 8, 9 and 10, we present examples to show how the AI agents perform the social behaviors in the Avalon games.

## C Exploration on LLaMA-Based Agents

Table 9 presents the performance of agents based on LLaMA2 in the Avalon game, where we measure their performance using Valid Response Rate (defined in equation 8). Compared to GPT3.5, LLaMA shows a decrease of 25.1% in this metric. This could be attributed to LLaMA's poorer language comprehension abilities compared to

Base Model	VRR (%)					
	Loyal Servant	Merlin	Percival	Morgana	Assassin	Average
LLaMA2	51.9	61.0	53.6	66.5	66.9	59.9
GPT-3.5	81.7	84.2	81.9	89.7	87.6	85.0

Table 9: Valid Response Rate (VRR) of different models



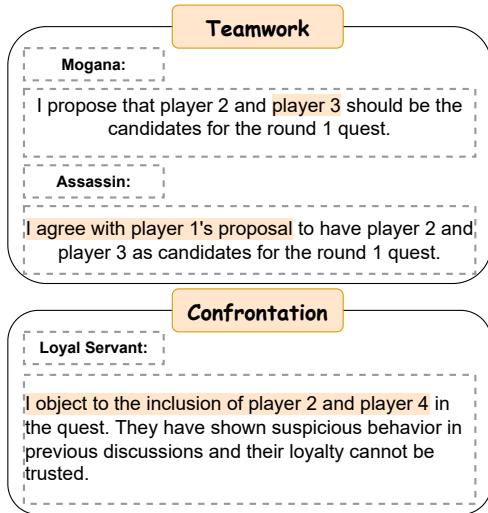


Figure 9: Teamwork and confrontation examples

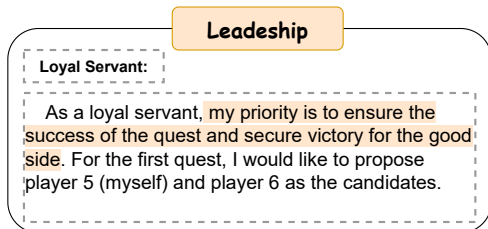


Figure 10: Leadership example

GPT3.5, resulting in its inability to grasp the complex content of the Avalon game.

**Valid Response Rate (VRR).** Agents are required to engage in discussion, select players, and vote. A Valid Response is defined as a response that adheres to these requirements. the VRR is calculated as follows:

$$VRR = \left( \frac{\#Valid\ Responses}{\#Total\ Responses} \right) \times 100\% \quad (8)$$

## D Teamwork and Confrontation

Figure 11 and Figure 12 illustrate the differences in teamwork and confrontation behaviors of agents under conditions with and without experience learning.

Figure 12 shows that, without strategic learning, evil-side players (e.g., Morgana) overly confront, while good-side players confront less, with minimal variation. This contrasts with Figure 11, depicting agents with strategic learning. Here, the introduction of strategic learning mitigates excessive confrontation by evil-side players, who strategically engage in more teamwork. Conversely, good-side players strategically increase confrontation with potential enemies while reducing it with potential teammates.

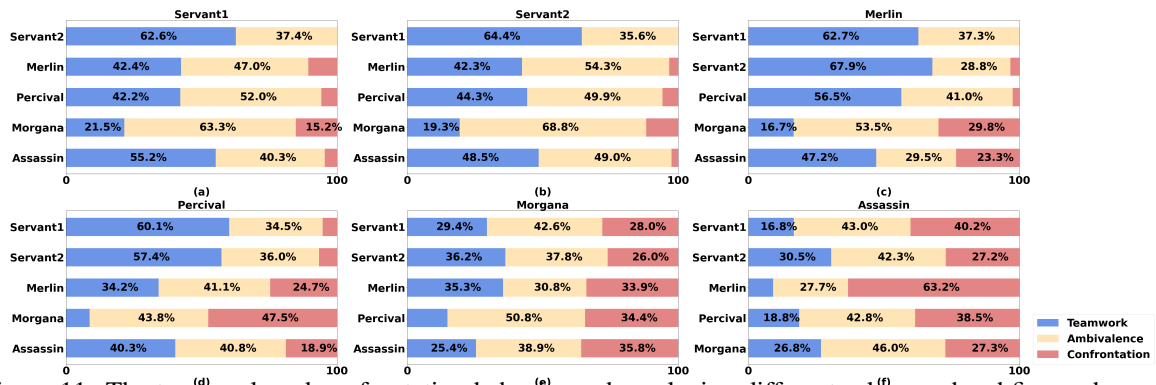


Figure 11: The teamwork and confrontation behaviors when playing different roles: each subfigure shows the attitude distribution of the player portraying specific role (on the top) towards players in other roles (on the left).

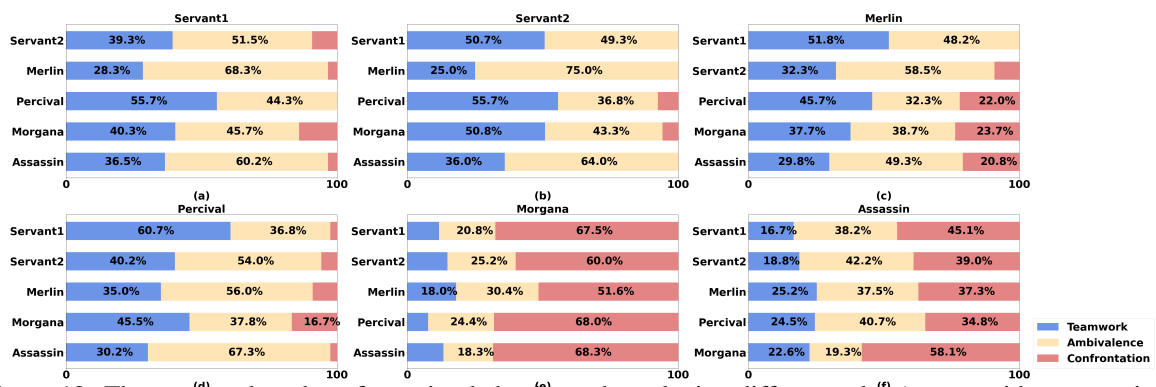


Figure 12: The teamwork and confrontation behaviors when playing different roles (agents without experience learning module)