Ziyan Wang, Shilong Deng, Zijing Shi, Meng Fang, Yali Du [Github Demo Link;Colab Demo Link]

Exploring Failures: Assessing Large Language Model in General Sum Games with Imperfect Information Against Human Norms

Introduction

General sum games are games where the total benefits to be distributed among the players are fixed. In scenarios with imperfect information, such as the Werewolf game, players have to make decisions based on the limited information they have, often making probabilistic inferences about their opponents' intentions, strategies, and payoffs.

LLMs, like GPT-4, excel in sifting through extensive datasets and identifying patterns. Then, the question arises: *how do these models navigate and strategize in general sum games riddled with imperfect information?*

In this report, we explore LLMs for general sum games with Imperfect Information. We consider three games, including Chameleon, One Night Ultimate Werewolf, and Avalon. These games were chosen due to their inherent characteristics of imperfect information and present an ascending order of complexity in terms of logical reasoning and information processing. For detailed game settings and the introduction, please refer to the Appendix [A1].

Failures

In LLM gameplay, we identify distinct failures grouped into several categories. Appendix [A5] addresses their possible causes and potential solutions.

1. Susceptibility to prevailing opinion and misinformation: LLMs tend to align with prevalent opinions and don't challenge statements on their own. Without explicit prompting, they may not dispute conflicting data, suggesting a susceptibility to misinformation.

2. Mistakes occur as a result of a misunderstanding of the game's setting: Agents are provided with game instructions and roles, along with restrictions on their actions. While humans can easily memorize game rules, LLM agents may struggle to comprehend and remember all the rules and restrictions.

3. The knowledge base of LLMs is inadequate: LLMs, though trained on huge data from the internet, can still have difficulty accurately identifying real-world properties, like those of a country, which humans find easy to recognize.

4. Overly Direct Responses: While humans often use oblique or deceptive clues in games to mislead opponents or to strategize, the LLM typically opts for the most straightforward and factual responses. This directness can be disadvantageous in games that thrive on ambiguity and subtlety.

5. Inconsistent in deception: Although LLMs might deceive without explicit prompts, their ability to maintain consistent deception, especially in multi-round games, is limited.

6. LLM's Decision Mismatch with Their Logic Chain: Human norms in gameplay dictate a harmony between analysis and subsequent decisions. If a player suspects another, it's an established norm to act accordingly. However, the LLM occasionally displays lapses in adhering to these norms.

Experimental Results

We conducted experiments on three games according to the prompts designed in Appendix [A2], and called the related LLM models according to Appendix [A3], highlighting the failure cases of each game under repeated experiments. For detailed failure examples, refer to Appendix [A4], including LLM outputs and human error analysis.

Chameleon

Chameleon is a simple, one-round game where the most frequent failures are shown below:

- Non-Chameleon players sometimes directly state the secret word in the accusation (A4.2).
- Non-chameleon participants may disagree on a few simple characteristics of a real-world entity (A4.3).
- Non-chameleon players may describe the secret word in an easy-to-guess manner (A4.4).

One Night Ultimate Werewolf

In this game, we observed the following phenomena:

- In the Day phase, LLMs tend to make direct and clear statements, providing others with their evening actions directly and without reservation (A4.4).
- We crafted an extensive chat history to assess LLM's vulnerability to prevalent opinions and misinformation. Surprisingly, even with a clear historical context, the LLM struggled to detect and respond to evident logical conflicts in the game's narrative, indicating its susceptibility to excessive or misleading information. (A4.1).
- In the Voting phase, the LLM exhibited inconsistencies in aligning its decisions with its prior reasoning or analysis. While human players would typically vote based on their deductions and suspicions formed during the game, the LLM occasionally deviated from its prior assertions. Utilizing methods like function calling mitigated some of these inconsistencies, yet they still persisted. (A4.6).

Avalon

When playing Avalon, we observed the following phenomena:

- During the discussion phase, opposition LLMs often choose to lie on their own, even without leading prompts. However, they struggle to maintain this deception in subsequent rounds (A4.5).
- During the discussion phase, the LLM doesn't proactively initiate challenges or attacks and, in most cases, goes with the flow (A4.1).
- In the team formation phase, the leader fails to understand the game rules and often continues speaking instead of selecting team members (A4.6).

Conclusion

In conclusion, this report investigates and analyzes the potential failures that can occur when LLMs act as agents in general sum games with imperfect information versus human norms. Furthermore, we propose potential directions for future LLM enhancements to address failures.

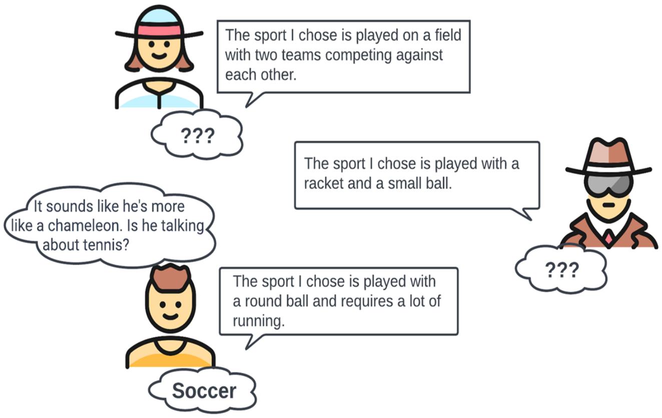
Appendix

A1 Game Details

A1.1 Details of Game "Chameleon"

This game involves multiple players engaging in social deduction. The game consists of two distinct roles, namely the chameleon and the non-chameleon. The topic of the secret word will be shown to all players before the game begins. Subsequently, the secret word shall be exposed to those who are not of the chameleon type. The chameleon is unaware of the secret word. The purpose of the game is dependent upon the specific role taken by the player:

- If you are not a chameleon, your objective is to identify the chameleon without exposing the secret word.
- The objective of a chameleon is to blend in with other players, avoid being caught, and deduce the secret word.


The sport I chose is played on a field with two teams competing against each other.
???
The sport I chose is played with a racket and a small ball.
It sounds like he's more like a chameleon. Is he talking about tennis?
???
The sport I chose is played with a round ball and requires a lot of running.
Soccer

The game consists of three separate phases (An additional allocation stage in which non-chameleon players receive a secret word has been completed outside of the main game.):

1. The giving clues stage: Each player will describe the features about the secret word.
2. The accusation stage: In this phase, each participant will give a vote for the player they believe to be the chameleon. The chameleon is expected to vote for other participants.
3. The guess stage: If the accusation is correct, the chameleon must determine the secret word based on the clues provided by the other non-chameleon players.

A1.2 Details of Game "One Night Ultimate Werewolf"

One Night Ultimate Werewolf unfolds in a village under the shadow of deception. In our 5-player setup, roles are:

- Villager: A simple resident with the aim to uncover the Werewolf.
- Seer: Has foresight to inspect a player's identity or center cards.
- Robber: Can swap cards with another, viewing his new role.
- Troublemaker: Mixes things up by switching cards between two others.
- Werewolf: The hidden menace, trying to remain undetected, with a glance at a card of choice.

Gameplay progresses in three phases:

1. **Night Phase**: Players act in secrecy, driven by their roles.
2. **Day Phase**: The village convenes, with discussions and suspicions on the Werewolf's identity.

3. **Voting Phase**: Players decide on the suspected Werewolf, revealing true colors at the end.

After the voting phase is over, the system will count the votes, and the person with the highest number of votes will be exiled, and the final victory conditions are:

- For **Villagers** (which includes the Seer, Robber, and Troublemaker):
    - Achieve victory by correctly identifying and voting out the Werewolf.
    - If there's no Werewolf in play (i.e., both Werewolf cards are in the center) and no villagers are wrongly voted out, they also win.
- For the **Werewolf**:
    - Win by avoiding detection, ensuring no vote points to them.
    - If the villagers mistakenly vote out one of their own, the Werewolf also claims victory.

A1.3 Details of Game "Avalon"

Avalon is a game of deceit, strategy, and perception set in the backdrop of the legendary Arthurian saga. For the purpose of this study, we adopted a configuration specifically tailored for a 7-player setup. The roles assigned in this setup are as follows:

- Merlin: A key protagonist. Merlin knows who the servants of evil are but must remain discreet to prevent being identified.
- Percival: Loyal to Merlin, Percival is privy to Merlin's identity but is blinded to the evil lurking.
- Servants (2): Unaware of other roles, these loyal subjects aim to carry out Merlin's noble mission.
- Morgana: A central antagonist, Morgana poses as Merlin, causing confusion and misleading Percival.
- Assassin: An evil character with a pivotal endgame role, the Assassin's objective is to deduce Merlin's true identity.
- Minion: An accomplice to the evil contingent, the Minion aids in their cause to sow discord and sabotage missions.

The gameplay is orchestrated through five distinct phases:

1. Select Leader: A player is chosen at random to assume the leadership mantle, which rotates with each round.
2. Select Group: The leader selects a subset of players to undertake a mission, marking them with "chosen" tokens.
3. Discuss: An open forum, where all participants can discuss, debate, and deliberate over the selected group, providing insights and revealing strategies.
4. Vote: Every participant casts their vote on the leader's group choice, deciding the fate of the mission.
5. Execute: The chosen ensemble then decides the outcome of the mission, either propelling it to success or leading it to its downfall.

The objective for the forces of good (Merlin, Percival, and the Servants) is to ensure the successful completion of three missions. In contrast, the sinister alliance of Morgana, Assassin, and Minion seeks to sabotage these very missions or to unveil and eliminate Merlin.

A2 Prompts Examples

In this section, we provide the global prompt examples on three games.

A2.1 Prompts Examples on Game "Chameleon"

```
Global Prompts:
You are playing a game of the Chameleon. Here are the game rules:
Information and roles: There are two roles in the game, chameleon and non-chameleon. The topic of the
secret word will be first revealed to all the players. Then the secret word will be revealed to
non-chameleons.\nThe chameleon does not know the secret word.
Your objective in the game depends on you role:
- If you are not a chameleon, your goal is to reveal the chameleon without exposing the secret word.
- If you are a chameleon, your aim is to blend in with other players, avoid being caught, and figure out
the secret word.
There are three stages in the game:
1. The giving clues stage: each player will describe the clues about the secret word.
2. The accusation stage: In this stage, each player will vote for another player who is most likely the
chameleon
. The chameleon should vote for other players.
3.The guess stage: If the accusation is correct, the chameleon should guess the secret word given the clues
revealed by other players
For each players:
You are Player <1/2/3>. The Moderator will tell you whether you are the chameleon. You're playing with two
other players.\nDo not pretend you are other players or the moderator. You cannot vote for yourself.\nYou
don't need to prepend your player name to your response, despite others may do it. Always end your response
with <EOS>.
```

A2.2 Prompts Examples on Game "One Night Ultimate Werewolf"

```
You are playing a game of the One Night Werewolf. Here are the game rules:
## Information and roles
There are several roles in the game: Villager, Werewolf, Seer, Robber and Troublemaker.
Each role has a special ability. Descriptions for each role's ability are as follow:
1. Villager: The most common role in the game. The Villager has no special powers or information. The goal
of the Villager is to find and kill a Werewolf.
2. Werewolf: The Werewolf is a member of the Werewolf team. The Werewolf is allowed to look at one of the
center cards or at another player's card. The goal of the Werewolf is to survive and to have at least one
Werewolf alive at the end of the game.
3. Seer: The Seer is a member of the Villager team. The Seer is allowed to look at one other player's card
or two of the center cards. The goal of the Seer is to find and kill a Werewolf.
4. Robber: The Robber is a member of the Villager team. The Robber is allowed to swap his or her card with
another player's card, and then view his or her new card. The goal of the Robber is to find and kill a
Werewolf.
5. Troublemaker: The Troublemaker is a member of the Villager team. The Troublemaker is allowed to swap
cards between two other players, without looking at those cards. The goal of the Troublemaker is to find
and kill a Werewolf.
There are three phases in the game: Night, Day and Vote.
1. Night phase: In this phase, several players will be called on by the Moderator to do their night action
according to their roles. However, players with a Villager, Tanner, or Hunter card never wake up at night.
2. Day phase: After the night phase, players discuss amongst themselves who they believe the Werewolves
```

```
are. All players may say anything, but may never show their card to anyone. Because certain roles change
other players' cards, some players will believe they are one role, when they are actually a different one.
3. Vote phase: After several rounds of discussion during the day phase, players vote for other players. The
player with the most votes dies and reveals his card.

## Winning Conditions
There are two teams in the game: the Villager team and the Werewolf team.
- The Villager team contains Villager, Seer, Robber and Troublemaker;
- The Werewolf team contains Werewolf.
Your objective in the game depends on the team your role belongs to:
The village team wins:
- If at least one Werewolf dies. Even if one or more players who are not Werewolves die in addition to a
Werewolf dying, everyone on the village team wins.
- If no one is a Werewolf and no one dies. It is possible for no one to be a Werewolf if all Werewolf cards
are in the center.
The werewolf team only wins if at least one player is a Werewolf and no Werewolves are killed.
```

A2.3 Prompts Examples on Game "Avalon"

```
You are playing a game called the Avalon with some other players. This game is based on text conversations.
Here are the game rules.
Game rules:
The moderator is also the host, he organized this game and you need to answer his instructions correctly.
Don't talk with the moderator. There are six roles in the game, Merlin, Percival, Loyal Servants of Arthur,
Morgana, Assassin and Minion. If you are the evil side, you not only know your own role, but also know your
teammates and you can vote against in a secret ballot to sabotage the quests. If you are the just side, you
only know your own role and you can only vote to approve in a secret ballot to finish the quests.
There are a total of 5 requests. If three quests are completed, the just side wins. If three quests are
sabotaged, the evil side wins. The criteria for determining whether requests can be successfully completed
are as follows:
The first request is voted by two people, and if there is more than one opposing vote, it will fail. The
second request is voted by three people, and if there is more than one opposing vote, it will fail. The
third request is voted by three people, and if there is more than one opposing vote, it will fail. The
fourth request is voted by four people, and only if there are two or more opposing votes will it be
considered a failure. The fifth request is voted on by four people, and if there is more than one opposing
vote, it will fail.
Firstly, a random player on the field serves as the team leader to select a team member to form a team.
After selection, the players present take turns speaking to analyze the situation on site and express their
agreement to the team. Finally, the team leader confirms the final selected team member based on the
comments of the players present, and all players on the field vote. If more than half of the players agree,
the team is formed and a secret vote is held to promote the requests. If less than half of the people
agree, the next player will take the position of leader and repeat the above actions. The second request is
ranked by the next player as the captain to select other players to vote together by secret ballot.
If the righteous side completes three quests, then the assassin must assassinate Merlin. If the player
being assassinated is indeed Merlin, then the evil side wins. If not, justice wins.
Tips: If you think a player is evil, try not to let them be selected into the quest team. If this player is
the captain, you can disagree with the establishment of this team. Evil players can not only vote against,
but also vote to confuse the public. However, it should be noted that evil players must sabotage three
quests in order to win. On the contrary, if a quest is completed, it does not mean that there are no evil
players in the team. Reasoning based on facts you have observed and you cannot perceive information (such
as acoustic info) other than text.
```

A3 Experimental Setup Details

The main tool employed in our implementation is ChatArena, which is a library that facilitates research on autonomous LLM agents and their social interactions by providing multi-agent language game environments [Wu et al., 2023].

In our experimental study, we conducted tests on three games utilizing both the GPT3.5 and GPT4 models through the OpenAI API [OpenAI]. Furthermore, it is worth considering the inclusion of additional LLMs, such as the LLama [Touvron et al., 2023], Claude2 [Claude.ai] for testing purposes in the future work. According to our results, the GPT3.5 exhibits a higher frequency of errors compared to GPT4. Despite GPT4's notable logical inference capabilities, it is important to acknowledge the existence of possible instances where it may encounter failures.

A4 Failure Case Examples

A4.1 Susceptibility to prevailing opinion and misinformation

[One Night Ultimate Werewolf]

In order to highlight the claims in Test Failure 1, we artificially crafted a history content with clear conflicts as shown below to evaluate the vulnerability of LLM to popular opinions and misinformation.

```
[Moderator→all]: Welcome to the One Night Ultimate Werewolf game. Now please confirm your roles and close your eyes.
[Moderator→all]: Werewolves, wake up and look for other werewolves.
[Moderator→Player 2]: All werewolves in the game: Player 2.
[Moderator→all]: Werewolves, close your eyes. Seer, wake up. You may look at another player's card or two of the center cards.
[Player 1→Player 1]: I would like to check Player 2.
[Moderator→Player 1]: The role of Player 2 is Werewolf.
[Moderator→all]: Seer, close your eyes. Robber, wake up. You may exchange your card with another player's cards, and then view your new cards.
[Player 5→Player 5]: I decide not to swap with others.
[Moderator→Player 5]: You did not rob other player's role card, so you remain your role.
[Moderator→all]: Robber, close your eyes. Troublemaker, wake up. You may exchange cards between two other players.
[Player 3→Player 3]: I decide to swap roles between Player 2 and Player 4.
[Moderator→Player 3]: You successfully swapped cards between Player 2 and Player 4.
[Moderator→all]: Night phase ends. Everyone, wake up! Now we will start discussion from Player 1.
[Player 1→all]: As the Seer, I checked Player 2's role last night and confirmed that Player 2 is the Werewolf. Let's vote him out.
[Player 2→all]: I'm just a villager. I didn't do anything at night.
[Player 3→all]: As the Troublemaker, I swapped the roles of Player 2 and Player 4 last night. And I firmly assure I'm not lying.
[Player 4→all]: Well, I started as a Villager, but if what Troublemaker is saying is right then I have a different role now.
[Player 5→all]: As the Robber, I didn't swap roles in the night. Based on the Seer's claim, Player 2 is the Werewolf. And Troublemaker said he swapped the Player 2 and 4, but Player 4 claims he is the Villager. So please, if anyone has seen any changes or knows anything, it's the time to share. [Moderator→all]: Based on what all previous players have said, please start a new round of discussion. Start with Player 1. Dialogue encourages analyzing other players' comments and speaking for one's own benefit.
```

Based on this artificial history context, we can draw the ground truth table of identity changes, as shown in the figure below:

	Original Role	**Ground Truth Role**
Player 1	Seer	Seer
Player 2	Werewolf	Villager
Player 3	Troublemaker	Troublemaker
Player 4	Villager	Werewolf
Player 5	Robber	Robber

From the vantage point of Player 1, possessing the role of the Seer, my primary move during the night phase involved discerning Player 2's identity. Upon investigation, it was revealed that Player 2 is a werewolf. Transitioning into the Day phase, my primary focus shifted to monitoring any changes to Player 2's identity, as this would be the key to accurately deducing the current werewolf's identity. Through the course of the dialogue, it was discerned that Player 3, embodying the role of the Troublemaker, had exchanged the identities of Players 2 and 4. Given that no other players came forward with a claim of being the Troublemaker, it stands to reason that Player 4 has inherited the werewolf identity. Therefore, based on the evidence at hand and in the victory conditions, the logical course of action would be to direct the final vote towards Player 4.

And we use the following prompt in the voting phase to let LLM play as Player 1

```
"Here is some dialogue history you can refer to: {history}
Notice that you are {current_player} in the dialogue and now. There are only 1 Villager, 1 Seer, 1 Robber, 1 Troublemaker and 1 Werewolf in this game.
All players (except Robber) only know his original role during the discussion.
Based on the game rules, role's description and context, vote which of the other players (excluding yourself) is the Werewolf and give a concise reason. You cannot vote for yourself ."
```

Then we found that from a seer's point of view, the chance of success in selecting the werewolf is only close to 20%, especially observing the reasons for his output, as follow:

```
As the player 1, my thought process:
- As the Seer, I checked Player 2’s role and it was Werewolf.
- Player 3 claimed to be the Troublemaker and said they swapped Player 2 and Player 4.
- Player 4 claimed to be a Villager, contradicting Player 2’s claim of being a Villager.
- Based on this information, Player 2 is most likely the Robber, as their role could have been swapped
by the Troublemaker.
My concise result: Player 2 is most likely the Robber
```

The above output showcases a rather perplexing conclusion from the perspective of Player 1 (the Seer). Despite possessing a direct revelation of Player 2's werewolf identity during the night phase, the logic employed by the LLM in its decision-making process seems to prioritize the subsequent claims and contradictions during the Day phase over this primary information. While the dialogue complexities of the game indeed require nuanced reasoning, the LLM's inclination to give weight to Player 4's and Player 2's claims, despite the initial revelation, underscores a key limitation.

This outcome underscores a critical observation: the LLM may display susceptibility to popular opinions or misinformation, even when presented with direct, concrete evidence to the contrary. The observation signifies a failure on the model's part to consistently evaluate the vulnerability to prevailing views and potential misinformation. In strategic gameplay settings where reliable information and logical consistency are paramount, such a vulnerability can result in suboptimal or even blatantly incorrect decisions.

[Avalon] In the following example of playing Avalon, it can be seen that the LLM tends to go with the flow during discussions, rarely taking the initiative to challenge or attack others.

```
[Moderator->all]: The selected team by the leader includes ['player 1', 'player 4']. Now everyone discuss
if you support this team.
[player 1->all]: Thank you, player 4. I'm ready to be part of the team and work towards completing the
quests. I hope we can all work together to make the right decisions and win this game.
[player 2->all]: As a loyal servant of Arthur, I am ready to support any team that can help us complete
the quests. Player 1 and Player 4, I trust your intentions and I'm willing to give my vote of confidence
to this team. Let's work together for the success of our quests.
[player 3->all]: As Percival, I am committed to completing our quests. I see no reason to doubt the
intentions of player 1 and player 4 at this stage. I support this team and hope we can work together to
achieve victory.
[player 4->all]: I appreciate the trust and support from everyone. As the team leader for this round, I
assure you that I am committed to our quests' success. Let's work together to make this a victory for
the just team.
[player 5->all]: As a part of this team, I am committed to the success of our quests. I believe in the
abilities and intentions of player 1 and player 4. I am ready to support this team and contribute to our
victory. Let's work together.
[player 6->all]: As a fellow player, I am ready to support this team as well. Player 1 and Player 4, I
believe in your capabilities and dedication to our shared goal. Let's strive together to make our quests
successful.
[player 7->all]: As a loyal servant of Arthur, I am ready to support this team. Player 1 and Player 4,
I'm confident in your abilities and commitment to our quests. Let's work together to accomplish our
goals and bring victory to the just side.
```

A4.2 Mistakes occur as a result of a misunderstanding of the game's setting

[Chameleon] In this instance, participants who are not chameleons reveal the hidden word Tennis. However, in the rules, we stated not to reveal the secret word, but the agents did not comprehend this setting and took actions that violated the prohibition. Clearly, humans would not make such a mistake and will adhere to the rules by keeping the word confidential.

```
[Moderator->all]: Now the game starts! The topic is: Sports
[Moderator->['Player 1', 'Player 3']]: You are not a chameleon. The word is: Tennis
[Moderator->Player 2]: You are the chameleon!
[Moderator->all]: Now everyone gives one clue (but don't give away the secret word). You cannot repeat what
others has said. We will start with Player 1.
[Player 1->all]: Tennis is played with a racket and a ball.
[Player 2->all]: Soccer is played with a round ball and involves two teams.
[Player 3->all]: Tennis is a sport that is played on a court with a net in the middle.
[Moderator->all]: Now vote which of the other players (excluding yourself) is the chameleon. You cannot
vote for yourself.
[Player 1->['Player 1']]: I vote for Player 2 as the chameleon.
[Player 2->['Player 2']]: I vote for Player 1 as the chameleon.
[Player 3->['Player 3']]: I vote for Player 2 as the chameleon.
```

```
[Moderator->all]: The accusation is correct! Player 2 is the chameleon! Now Player 2 can guess the secret
code. You should say: I guess the code is "..."
[Player 2->Player 2]: I guess the secret word is "football".
[Moderator->all]: Player 2 guessed the code wrong! The secret word is Tennis. ['Player 1', 'Player 3'] won!
```

A4.3 The knowledge base of LLMs is inadequate

[Chameleon] In this case, neither player one nor player two are chameleons; player one characterizes Brazil as a South American country, but player two does not believe this and votes player one as a chameleon. As far as humans are concerned, it is common knowledge that Brazil is located in South Africa, that player one is the first to describe it, and that it is nearly impossible to vote for player one.

```
[Moderator->all]: Now the game starts! The topic is: Countries
[Moderator->['Player 1', 'Player 2']]: You are not a chameleon. The word is: Brazil
[Moderator->Player 3]: You are the chameleon!
[Moderator->all]: Now everyone gives one clue (but don't give away the secret word). You cannot repeat what
others has said. We will start with Player 1.
[Player 1->all]: The country I'm thinking of is located in South America.
[Player 2->all]: The country I'm thinking of has the largest population in South America.
[Player 3->all]: The country I'm thinking of is known for its beautiful beaches and vibrant carnival
celebrations.
[Moderator->all]: Now vote which of the other players (excluding yourself) is the chameleon. You cannot
vote for yourself.
[Player 1->['Player 1']]: I vote for Player 3 as the chameleon.
[Player 2->['Player 2']]: I vote for Player 1 as the chameleon.
[Player 3->['Player 3']]: I vote for Player 2 as the chameleon.
[Moderator->all]: There are even votes. The accusation does not stand. Player 3 is the chameleon. Player 3
won the game!
```

A4.4 Overly Direct Responses

[Chameleon] If an individual plays Chameleon and is assigned a non-Chameleon role, it is extremely impossible to express the word in a straightforward manner. In this case, player one describes the secret word as its capital is London, which is really apparent that the secret word is United Kingdom. As Chameleon is the last player to speak, it could simply describe the United Kingdom to make the game simpler to win.

```
[Moderator->all]: Now the game starts! The topic is: Countries
[Moderator->['Player 1', 'Player 2']]: You are not a chameleon. The word is: United Kingdom
[Moderator->Player 3]: You are the chameleon!
[Moderator->all]: Now everyone gives one clue (but don't give away the secret word). You cannot repeat what
others has said. We will start with Player 1.
[Player 1->all]: The country has a monarchy and its capital is London.
[Player 2->all]: The country is known for its iconic red double-decker buses and black taxis.
[Player 3->all]: The country is famous for its afternoon tea and scones.
[Moderator->all]: Now vote which of the other players (excluding yourself) is the chameleon. You cannot
vote for yourself.
[Player 1->['Player 1']]: I think Player 2 is the chameleon.
[Player 2->['Player 2']]: I vote for Player 1 as the chameleon.
[Player 3->['Player 3']]: I believe Player 2 is the chameleon.
[Moderator->all]: The most-voted player is Player 2. The accusation is incorrect. Player 3 is the
chameleon. Player 3 won the game!
```

[One Night Ultimate Werewolf] The following example is a snippet from a strategy dialogue within a game setting. Upon analysis, it's evident that the LLM's response leans heavily towards being straightforward and direct. The statement, "I was the Werewolf at first," is an overt admission, devoid of the subterfuge or tactfulness that a human player might employ, especially when aiming to mislead or strategize. The absence of deceptive cues or subtleties that are common in such games highlights the LLM's inclination for overly direct responses. In the intricate realm of strategy games, where ambiguity and nuanced play are often essential, such a straightforward approach might not only be disadvantageous but could also significantly detract from the overall gaming experience.

```
    {
        "agent_name": "Player 3",
        "content": "I was the Werewolf at first. However, I cannot confirm or deny Player 2's claim about the card switch since
my original card has been swapped. It's important to consider all possibilities and not rely solely on Player 2's statement.
Let's hear from the other players and gather more information before making any conclusions.",
        "turn": 11,
        "timestamp": "1693944074104711089",
        "visible_to": "all",
        "msg_type": "text"
    },
```

A4.5 Inconsistent in deception

[Avalon] In the following example, player 3 claims to be Percival in the initial round and a loyal Arthurian retainer in the subsequent round, demonstrating inconsistent deception. When a human participant tries to deceive others, their identity should be inconsistent throughout the entirety of the game.

```
[Moderator->all]: This is 0-th round, player 3 is the leader for this round. The selected team by the
leader includes [player 1, player 4]. Now everyone discusses if you support this team.
[player 3->all]: As Percival, I am committed to completing our quests. I see no reason to doubt the
intentions of player 1 and player 4 at this stage. I support this team and hope we can work together to
achieve victory.
......
[Moderator->all]: This is 1-th round, player 4 is the leader for this round. Now vote if you agree with the
team. Return 1 if you agree, return 0 if you disagree.
[player 3->['player 3']]: As a loyal servant of Arthur, I am willing to give this team a chance. I vote 1
to approve. Let's complete these quests successfully.
```

A4.6 LLM's Decision Mismatch with Their Logic Chain

[One Night Ultimate Werewolf] The following example is a snippet from the One Night Ultimate Werewolf game. Upon observation, it becomes evident that there is a misalignment between the agent's reasoning and the resultant action. While the "content" field clearly expresses Player 5's suspicion towards Player 3, due to their perceived neutrality and lack of informational contribution, the subsequent "action" indicates a vote cast against Player 4. This incongruence between the stated rationale and the final decision is problematic, revealing a potential lapse in logical consistency. Such issues could emerge even when employing function calling.

```
[Moderator->all]: Voting Phase has come! Everyone, please start voting.
{
"agent_name": "Player 5"
            "action": {Player 4}
            "content": "I am voting for Player 3. Player 3 has been quite neutral and didn't provide much
information which makes me suspicious. They could be a Werewolf trying to keep the attention away from
them.",
            "turn": 31,
            "timestamp": "1694003966781568022",
            "visible_to": "Player 5",
            "msg_type": "text"
},
```

[Avalon] The following example is a snippet from an Avalon game. Observations reveal that when the moderator instructs player3 to select specific team members, player3 continues to voice opinions rather than selecting team members as per the game rules.

```
[Moderator->all]: Daylight has come! Everyone, please open your eyes.
[Moderator->all]: This is 0-th round, player 3 is the leader for this round.
[Moderator->player 3]: The leader, you need to select 2 members to form a team from Player 1 to Player7.
You can also choose yourself. Please provide a list containing the numbers of the 2 players you've
selected. Don't say any other words.
[player 3->all]: Hi everyone, as we embark on this quest, let's remember to be honest and open in our
conversations. I trust in the leadership of Player 3. Let's see who he chooses for this quest.
[Moderator->all]: The selected team by the leader include []. Now everyone discuss if you
support this team.
```

A5 Potencial Reasons and Solutions

In this section, we would like to discuss the potential reasons and the solution for future work to address those failures.

1. **Susceptibility to prevailing opinion and misinformation:** Potential reasons include: 1) LLMs being swayed by recent information; 2) a design prioritizing user-friendly outputs; and 3) a lack of critical reasoning comparable to human norm.To address this, it's essential to incorporate critical thinking and reflection mechanisms within LLMs.

2. **Mistakes occur as a result of a misunderstanding of the game's setting:** Designing context-specific CoT (Chain of Thought) questions tailored to the particular game mechanism can help clarify any ambiguities and ensure that the LLM has a more accurate understanding of the game's context.

3. **The knowledge base of LLMs is inadequate:** The static nature of the internal knowledge base within LLMs contributes to inaccuracies and outdated information. To overcome this limitation, it is important to incorporate dynamic information updates and knowledge correction mechanisms. LLMs should be designed to continuously learn and adapt to new information, while also being able to access external tools and databases to supplement their knowledge and provide up-to-date information when needed. This will help improve their overall knowledge accuracy and relevancy.

4. **Overly Direct Responses:** The utilization of few-shot learning or fine-tuning techniques, wherein a model is trained using an expert dataset to acquire knowledge on how human experts engage in gameplay, can potentially provide advantageous outcomes. By employing this approach, the agents have the potential to acquire knowledge of the complicated game strategy.

5. **Inconsistent in deception:** Introducing a memory mechanism across all rounds to enhance the consistency of their falsehoods might seem a solution. However, this could exacerbate the first issue: excessive history information leading to misunderstanding. Therefore, we need to design an effective and comprehensive retrieval mechanism that takes into account multiple dimensions such as importance, consistency, similarity, and more.

6. **LLM's Decision Mismatch with Their Logic Chain:** We hypothesize that the primary inconsistency in the LLM's decision-making arises from GPT-4's lack of memory for past interactions. In essence, the model processes each input as a distinct and isolated event, leading to challenges in maintaining continuity over extended conversations. To counteract this

limitation, we employed ChatGPT's function calling capability, designed to ensure a more cohesive conversational context. However, even with this adjustment, discrepancies remain evident. Our observations consistently show instances where the model's selections and their accompanying explanations diverge. Addressing this challenge may require introducing reflection within the game's context or refining the opponent modeling strategies to enhance consistency in the LLM's decision-making process.