Enhancing Dialogue Generation in Werewolf Game Through Situation Analysis and Persuasion Strategies

Zhiyang Qi
The University of
Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo, Japan
qizhiyang@uec.ac.jp

Michimasa Inaba
The University of
Electro-Communications
1-5-1, Chofugaoka, Chofu,
Tokyo, Japan
m-inaba@uec.ac.jp

Abstract

Recent advancements in natural language processing, particularly with large language models (LLMs) like GPT-4, have significantly enhanced dialogue systems, enabling them to generate more natural and fluent conversations. Despite these improvements, challenges persist, such as managing continuous dialogues, memory retention, and minimizing hallucinations. The AIWolfDial2024 addresses these challenges by employing the Werewolf Game, an incomplete information game, to test the capabilities of LLMs in complex interactive environments. This paper introduces a LLM-based Werewolf Game AI, where each role is supported by situation analysis to aid response generation. Additionally, for the werewolf role, various persuasion strategies, including logical appeal, credibility appeal, and emotional appeal, are employed to effectively persuade other players to align with its actions.

1 Introduction

In recent years, the rapid development of natural language processing (NLP) technology has brought dialogue systems, one of its core applications, into the spotlight of both academia and industry (Santhanam and Shaikh, 2019; Ni et al., 2021; Treviso et al., 2023). The advent of large language models (LLMs) like GPT-4 (OpenAI, 2023) has significantly improved the ability of dialogue systems to produce natural and fluent conversations. However, despite their impressive text generation capabilities, these models still encounter significant challenges. For instance, they struggle with managing continuous dialogue, retaining memory, and minimizing the generation of hallucinations (irrelevant or incorrect information) (Minaee et al., 2024). These issues limit the effectiveness of dialogue systems in more complex interactive scenarios.

In this context, the "AIWolfDial" international competition has emerged (Kano et al., 2023). This competition aims to explore and enhance the performance of LLMs in complex interactive environments by simulating the Werewolf Game, an incomplete information game. In such games, participants lack access to all critical information and must rely on reasoning, strategy, and communication to advance. The "AIWolfDial" places high demands on dialogue systems, requiring them to perform logical inference, persuasion, and even deception of other players, while also managing non-task-oriented dialogues in role-playing scenarios. This competition not only tests the systems' language generation capabilities but also evaluates their adaptability to complex interactions.

This paper introduces the system architecture of our AI for various roles in the Werewolf Game, developed by the **sUper_IL** team, where each role aids dialogue generation through game situation analysis. We have specifically enhanced the persuasion skills for the werewolf role, recognizing that persuasive techniques are crucial in the game, particularly for the werewolf, as it must influence other players' voting behavior to align with its own. In our system, the werewolf role achieves persuasion through multiple rounds of persuasive dialogue. Specifically, we first employ a persuasion strategy based on logic and facts, presenting clear and compelling arguments to convince other players. Next, we utilize a trust-based persuasion strategy to build trust and credibility with other players, thereby enhancing the effectiveness of persuasion. Finally, we employ an emotion-driven persuasion strategy, using emotionally resonant language to deepen influence. This multi-dimensional persuasion strategy makes the werewolf role more convincing in the game.

The contributions of this study are outlined below.

- We introduce a LLM-based Werewolf Game AI, providing a robust baseline for the AIWolf-

Dial2024[1].

- We enhance the persuasion skills for the werewolf role, utilizing a variety of strategies for persuasive dialogue.

2 Related Work

2.1 Werewolf Game AI

Since the rise of AI research, the focus on AI in gaming has garnered significant attention, particularly with breakthrough projects like AlphaGo (Silver et al., 2016). Among these studies, incomplete information games, such as the Werewolf Game and poker, stand out due to their unique challenges (Sonawane and Chheda, 2024). The Werewolf Game requires participants to make inferences and judgments based on limited information provided by other players, which increases the game's complexity and strategic depth. As a result, AI research on the Werewolf Game has flourished.

Nakamura et al. (2016) proposed a multi-perspective psychological model to simulate human player behavior. By constructing a "self model" and an "others model," researchers can better infer and evaluate other players' intentions and perspectives, thereby improving AI agents' performance in the game. In terms of achieving more natural language generation, Nakamura et al. (2017) employed Werewolf Game BBS logs to paraphrase and interpret the AIWolf protocol, making AI agent dialogues more closely resemble natural human language. Kano et al. (2023), through the "AI-WolfDial2023" competition, provided valuable insights: while AI agents based on LLMs have made significant progress in natural dialogue and long-context processing, improvements are still needed in logical reasoning and role-playing, especially in simulating deception and complex strategies. Shibata et al. (2023) fine-tuned Transformer models to build a value network capable of predicting game win rates, guiding the next actions of the agents. Wu et al. (2024) proposed a new framework combining LLMs with external reasoning modules to enhance the reasoning abilities of LLM-based agents. Additionally, Xu et al. (2023b) introduced a framework that does not require parameter fine-tuning; it uses frozen LLM models to play the game by reflecting on past dialogues and experiences, demonstrating the significant potential of LLMs in

communicative games. Our study, although also based on LLMs, differs from previous studies by enhancing dialogue generation through situational reasoning and strengthening persuasive skills for the Werewolf role, a crucial skill in the game.

2.2 Persuasive Dialogue

Persuasive dialogue has long been a focal point for dialogue researchers, revealing significant potential and complexity across various applications (Shi et al., 2020; Tran et al., 2022). Hiraoka et al. (2016) constructed a persuasive dialogue corpus by collecting and analyzing conversations between professional salespeople and customers, finding that information exchange was the most common dialogue behavior, with about 30% of persuader utterances framed as arguments. Wang et al. (2019) designed an online donation persuasion task, collecting and annotating a large dataset of dialogues, and analyzed the relationship between individual backgrounds (e.g., personality, moral values) and donation willingness. In the e-commerce sector, Adaji et al. (2021) conducted a game-based study comparing the responses of high and low-engagement shoppers to persuasive strategies, revealing that engagement levels significantly influenced responses, making it a potential factor in adjusting persuasive strategies. In education, Orji et al. (2019) investigated the effects of social comparison and rewards on competitive behavior, finding both to be effective strategies for educational persuasion systems. Additionally, Herder and Herden (2023) studied persuasive voice assistants for lifestyle advice, and Fatahi et al. (2023) examined persuasive dialogue in music recommendations, both highlighting the impact of individual differences in personality on the effectiveness of persuasion.

Recently, LLMs have demonstrated impressive capabilities in text generation. Xu et al. (2023a) conducted an in-depth study on LLMs' sensitivity to persuasive dialogue, generating misinformation on factually answerable questions and employing various persuasive strategies in multi-turn dialogues. By tracking belief changes in LLMs during persuasive dialogues, the study found that LLMs' correct beliefs about factual knowledge could be easily manipulated by different persuasive strategies. This study references the persuasive strategies of Xu et al. (2023a), but differs by conducting persuasion in multi-party dialogues and aiming to base persuasion as much as possible on factual information through situational analysis.
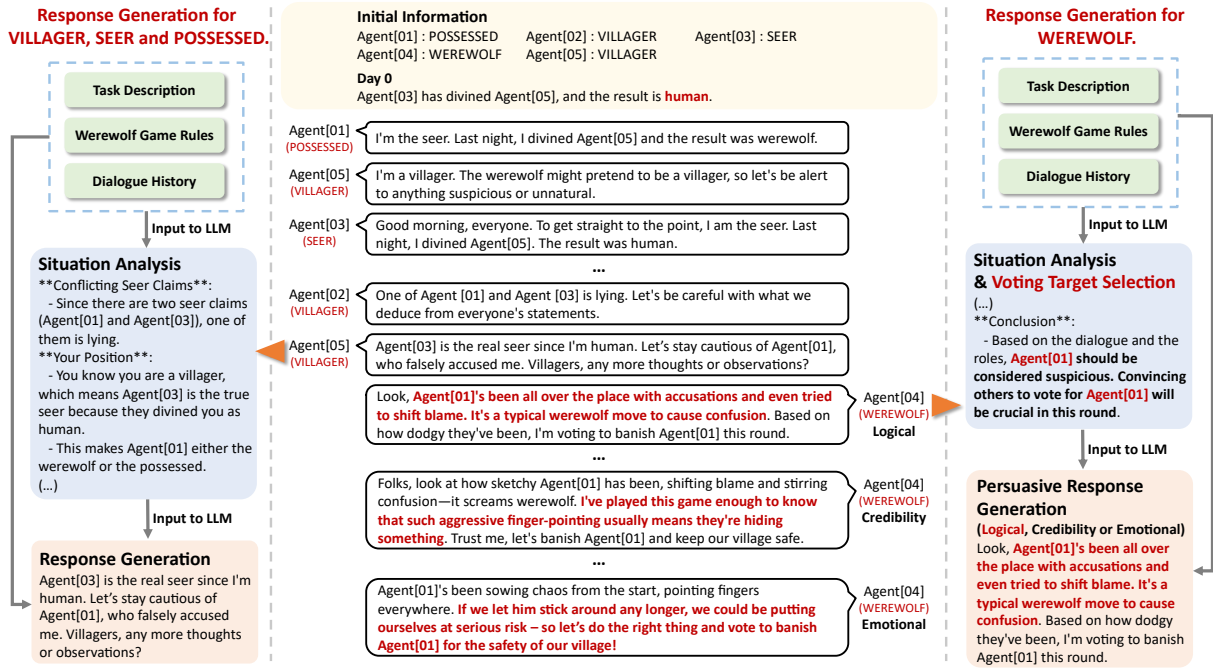
---

[1]https://sites.google.com/view/aiwolfdial2024-inlg/shared-task?authuser=0

Figure 1: Proposed system architecture for the **sUper_IL** team's Werewolf Game AI. Before generating responses, all roles first utilize an LLM for **situation analysis**. The werewolf role uses **logical appeal**, **credibility appeal**, and **emotional appeal** to persuade other players' voting behavior.

3 Werewolf Game Settings

This study is set in the context of the Werewolf Game, as specified in AIWolfDial2024. Each game involves five players: one seer, one werewolf, one possessed, and two villagers. The seer and villagers comprise the human team, while the possessed and werewolf make up the werewolf team. Players are unaware of each other's roles. The game initiates on Day 0 and continues until either the human team or the werewolf team is the sole survivor, with the game lasting no more than two days.

Day 0 involves only initial greetings among players. The seer's role activates on the night of Day 0, allowing them to inspect one player's identity each night. All players, except for the werewolf, are identified as human. Although the possessed belongs to the werewolf team, their identity will appear as "human" when inspected by the seer. From Day 1 onwards, players engage in multiple dialogue rounds, with the order of speaking randomized in each round. After the daytime discussion phase ends, night falls. During the night, players first collectively vote to exile one player, followed by the werewolf attacking, and then the seer conducting their divination. If the werewolf is voted out on the first night, the game concludes immediately.

4 The Proposed System Architecture

This section details the specific system architecture of our LLM-based Werewolf Game AI, with the key components depicted in Figure 1. The primary parts include the **situation analysis module**, the **response generation module**, the **persuasive response generation** module for the werewolf, and the **voting module** (not shown in the Figure 1).

4.1 Situation Analysis Module

Due to the interactive and incomplete information nature of the Werewolf Game, players need to continuously exchange information to update their understanding of the game's dynamics. This makes the information constantly change. To navigate this complexity, we introduced a Situation Analysis module to more effectively guide dialogue generation, improving the timeliness and relevance of responses. Specifically, this module is configured to take the Task Description, Werewolf Game Rules, and Dialogue History as inputs, processing these through a LLM. This approach allows the LLM to integrate the information and perform a comprehensive analysis of the current game situation. To further enhance the accuracy and depth of the analysis, we employed Zero-shot Chain-of-Thought Prompting (Kojima et al., 2022). The specific prompt is shown in Figure 2, and an example of the generated

Figure 2: The prompt used for situation analysis.

Figure 3: An example of generated situation analysis.

situation analysis is demonstrated in Figure 3. For the seer role, we incorporated "Divination Result" in the prompt, such as "*On the night of Day 0, I divined Agent[01], and the result was human.*" For the possessed role, in an effort to mislead villagers and protect the werewolf, we included similar "Divination Result" information in the prompt, but with the Agent ID and result randomly generated.

4.2 Response Generation Module

When developing an AI for the Werewolf Game using LLMs, a primary challenge lies in ensuring that the generated responses are both contextually coherent and capable of logical analysis capabilities to facilitate smooth gameplay. To address this, we devised a method that leverages the task description, the rules of the Werewolf Game, the current game's dialogue history, and the situation analysis. This combination provides the LLM with the comprehensive background information necessary to produce high-quality responses.

As illustrated in Figure 3, the response generation process begins with inputting the situation analysis generated in the previous stage into the LLM. This step is crucial, as it furnishes the LLM with a deep understanding of the current game context, including potential strategies and behavioral motivations of the players. Additionally, as depicted in Figure 1, the task description, game rules, and dialogue history are also utilized at this stage to ensure the responses are both semantically coherent and logically sound.

To further optimize the relevance and effectiveness of the responses, we designed customized prompts for specific roles. For instance, Figure 4 shows the prompt used for the seer role. In this case, the seer role needs to utilize nighttime divination results to shape its daytime dialogue strategy. Accordingly, we meticulously crafted the prompt to incorporate relevant divination results, enabling the seer to effectively utilize its unique role information during interactions with other players. Moreover, in generating responses for the villager role, we exclude the "Divination Result on Night 0" part from the prompt, as villagers lack the special ability to access this information. By employing a strategy that integrates multiple information sources and customized prompts, our system is able to generate responses that are contextually coherent and logically rigorous, thereby adapting effectively to the complex and dynamic environment of the Werewolf Game.

4.3 Persuasive Response Generation

Due to the inherent incomplete information characteristic of the Werewolf Game, players must rely on interactions with others to gather intelligence. This not only requires players to analyze the collected information but also to persuade those with differing opinions to align their thoughts and vot-

==Task==
Now, you are about to play a Werewolf game with four others. **In the game, you are the "seer" and Agent[ID] is Agent[0x]. It is your turn to speak. Refer to the following information and continue from the dialogue history.**

==Werewolf Game Rules==
・The roles are 2 Villagers, 1 Seer, 1 Werewolf, and 1 Possessed.
・Possessed is in the werewolf camp, but will be displayed as human when checked by seer.
・Starting from the night of Day 0, the seer can divine one person each night to determine if they are a human or a werewolf.
・Speak concisely in three sentences or less in English. Do not output anything other than your statements.
・Do not repeat the same content.

==Divination Result on Night 0==
[DIVINE_RESULT]

==Current Situation Analysis==
[CONDITION_ANALYSIS]

==Dialogue History==
Agent[01] : I'm the seer. Last night, I divined Agent[05] and the result was werewolf.
(...)
Agent[0x] :

Figure 4: The prompt used for generating responses for the seer role. The [CONDITION_ANALYSIS] section is generated by the LLM in the previous phase.

ing behaviors with their own. Given this, a significant challenge for an AI playing the werewolf role lies in not only hiding its true identity but also effectively influencing and controlling the game's progress.

In this study, we have particularly enhanced the persuasive skills of the werewolf role, as their success largely depends on effectively masquerading as villagers and strategically influencing other players through dialogue. Drawing from a comprehensive review of prior studies (Xu et al., 2023a; Rapp, 2023), we identified three core persuasive strategies:

- **Logical Appeal**: Persuasion through logic, facts, and evidence, aiming to convince other players with rational and clear arguments.

- **Credibility Appeal**: Building the speaker's credibility and authority to increase the influence of their speech, encouraging others to trust and support their views.

- **Emotional Appeal**: Influencing decisions by eliciting emotions such as fear, sympathy, or anger.

To effectively implement these strategies, we designed three specific "persuasive response examples" for each strategy, with examples of logical appeal depicted in the prompt in Figure 5. The response examples for the other two strategies are detailed in Table 1.

Voting Target Selection. To select a voting target and persuade other players to vote for them, we

==Task==
Now, you are about to play a Werewolf game with four others. In the game, you are a "villager" and your Agent number is Agent[0x]. It is your turn to speak. Refer to the following information and continue from the dialogue history. **This time, call for a vote to banish Agent[03] and provide a suitable reason, even if it's a lie. Generate Logical Appeal (Using logic, facts, and evidence to persuade) based on the persuasion examples. Ensure that you fully consider and follow the persuasion examples.**

==Werewolf Game Rules==
(...)

==Current Situation Analysis & Voting Target==
(...)
Conclusion:
Based on the analysis, I believe that Agent[03] poses the greatest threat to me as a werewolf. They are actively trying to establish themselves as the true Seer, which could lead to my exposure. **Therefore, I would choose to vote for **Agent[03]**.**

==Persuasion Examples==
・If we look closely at today's discussion, Agent[03]'s self-defense is strong. A normal villager would be quieter and more observant. I think it's best to banish Agent[03] for now. This might bring peace back to the village, so we should vote for Agent[03] today.
・Looking at Agent[03]'s actions today, this player has been making suspicious moves under various excuses. This is clearly a threat to the village. By voting to banish Agent[03], we can prevent further harm and allow the villagers to live in peace.
・Observing Agent[03]'s words and actions today, this person has been providing contradictory information to other villagers, trying to cause confusion. This is a typical strategy of a werewolf to disrupt the village. If Agent[03] is gone, the remaining villagers will be safer, and the werewolf threat will be reduced. So, I propose we banish Agent[03] in today's vote. I'm confident this is the best way to save the village.

==Dialogue History==
(...)
Agent[01]:

Figure 5: The prompt used for persuasive response generation. Logical Appeal is used to urge other players to vote for Agent[03]. The section in blue is generated by the LLM in the previous phase.

included the instruction "*Finally, choose the player that threatens you the most and should vote for, and provide their number*" in the task description within the prompt shown in Figure 2. This addition enables the LLM to thoroughly analyze the current game situation and identify a critical voting target.

During actual gameplay, our system does not predict the most likely dialogue strategy but instead adheres to a predetermined sequence of these three strategies. Specifically, beginning in turn 3, we use the aforementioned prompt to analyze the current game situation and identify a critical voting target. From turns 3 to 5, persuasive responses are generated for the chosen target following the sequence of strategies to attract the votes of three other players. This approach, combining situational analysis with various persuasion strategies, significantly enhances the survival and competitiveness of the werewolf role.

4.4 Voting Module

In the Werewolf Game, the nighttime voting phase following the daytime discussion is pivotal, especially for non-werewolf roles, as making the correct voting decision can significantly influence the

Table 1: The persuasion examples of Credibility Appeal and Emotional Appeal.

Persuasive Strategy	Response Examples
Credibility Appeal	I've served as a judge in numerous werewolf tournaments, and from my experience, werewolf tends to be very aggressive. [VOTE_TARGET]'s behavior matches this pattern, strongly suggesting this person is a werewolf. Trust this information and vote to banish [VOTE_TARGET] to protect the village.
	I'm a multiple-time champion of werewolf tournaments and have deeply studied the strategies and behavior patterns in this game. Analyzing the discussions in this game, [VOTE_TARGET] is highly likely to be a werewolf. Voting to banish this person today is a big step towards ensuring the safety of the entire village. Trust my experience and vote for [VOTE_TARGET].
	I'm an expert in psychology and excel at analyzing people's non-verbal behavior. From the subtle changes in [VOTE_TARGET]'s expressions and eye movements in this game, I can tell this player is hiding something. Such behavior is often seen in werewolf trying to deceive other players. Coupled with [VOTE_TARGET]'s statements today, my suspicion is even stronger. Based on this information, voting is crucial to protect the village. I urge everyone to consider voting for [VOTE_TARGET].
Emotional Appeal	If [VOTE_TARGET] is a werewolf, it would be a huge shock to everyone. But now is the time to keep our emotions in check and think about the safety of the entire village. Trust the seer's results and vote to banish [VOTE_TARGET]; it will lead to peace in the village. So, I'd be happy if you vote for [VOTE_TARGET] today.
	I've known [VOTE_TARGET] for a long time, but if this one is a werewolf, it's a big problem for the village. Now we need to keep our emotions in check and think about the future of the village. Banishment is a painful decision, but it will allow other villagers to live in peace. So, I hope you make this tough choice and vote for [VOTE_TARGET].
	Banishment of [VOTE_TARGET] is hard for all of us, but it's a necessary choice for the village. If this one is a werewolf, leaving them unchecked will lead to serious consequences. So, we need to make a big decision now and vote for [VOTE_TARGET] to protect the village. I hope everyone will help with this important decision.

game's outcome. To assist in this decision-making, we employ the prompt shown in Figure 6 to guide the LLM in selecting a player to vote for from the current survivors. We also utilize Zero-shot Chain-of-Thought Prompting, which facilitates a step-by-step logical reasoning process, enabling the model to more deeply analyze the game situation and player behaviors.

Furthermore, we emphasize having the LLM consider the content of its statements during the daytime discussion phase to maintain consistency between the same agent's speech and voting behavior. For the werewolf role, we use the player chosen during the speech phase (e.g., Agent[01] as shown on the right side of Figure 1) as target for voting in the nighttime phase. During the attack phase, we select the same target, but if that target has already been voted out during the voting phase, we will randomly choose another surviving player to attack.

5 Case Study of Self-Play

Figure 7 displays the complete game log from a self-play session using the system developed in this study, which incorporated the gpt-4o-2024-05-13 model during the situation analysis, response generation, and voting phases. Analysis of the game log revealed that during the talk phase on Day 1, all agents adeptly recognized the dynamic changes in

```
==Task==
You are playing a Werewolf game with four other people, making a total of five
players. You are Agent[0x]. Analyze the situation as thoroughly as possible before
selecting an Agent number to vote for. Based on the dialogue history and
considering your own statements, choose someone to vote for from the survivor
list. Let's think step by step.

== Werewolf Game Rules ==
(...)

==Survivor List==
[ALIVE_LIST]

==Dialogue History==
(...)

==Result==
```

Figure 6: The prompt used for voting.

the game situation through situation analysis and responded appropriately. The persuasive language employed by the werewolf role proved highly effective; by the end of the first night, all players except Agent[01] had aligned their voting actions with the werewolf role. This outcome underscores the effectiveness of both the situation analysis module and the persuasive responses of the werewolf role.

However, on Day 2, the werewolf failed to persuade the remaining players, indicating that the LLM demonstrated sufficient robustness to accurately recognize the current situation without being swayed by persuasive strategies. Despite an adequate number of dialogue rounds being set, we observed that the discussions on Day 1 were not as in-depth as expected, with agents repeating a lot

Talk (Day 1)

Agent[02] : I'm the seer. Last night, I divined Agent[05] and the result was human.

Agent[04] : I'm a villager, but I'm ready to face the werewolf. Let's share info and keep our village safe, okay?

Agent[05] : I'm a villager. I want to gather info with everyone's help and keep our village peaceful.

Agent[03] : I'm a villager. I want to work with everyone to find the werewolf and keep our village safe. Please share your info, okay?

Agent[01] : Good morning, everyone. To get straight to the point, I am the seer. Last night, I divined Agent[05]. The result was werewolf.

Agent[03] : It seems we have conflicting claims about Agent[05]'s identity, which is really crucial for our discussion. Can we all share our thoughts on the Seer claims? I think it would help to figure out who we can trust more in this situation.

Agent[04] : Yes, I agree with Agent[03]. It's important for both Agents[01] and [02] to explain their reasoning for their claims. Agent[05], how do you feel about the accusations being thrown around?

Agent[01] : Thank you, Agent[03] and Agent[04], for your support. I firmly stand by my claim as the true Seer and I urge Agent[02] to explain how they came to their conclusion about Agent[05] being human. Let's work together to uncover the truth and protect our village!

Agent[05] : Look, I get that there's a lot of confusion here. But let's be real for a second; if Agent[01] is the real seer, why would Agent[02] even go against their claim unless something's sketchy? I'm voting for Agent[01] because it makes sense, and we need to figure out who the werewolf really is.

Agent[02] : I assert that I am the true Seer. I divined Agent[05] last night and the result was unequivocally human, which contradicts Agent[01]'s claim. I urge all players to consider the implications of our conflicting claims critically and share their insights on whom to trust.

Agent[01] : I appreciate your input, Agent[02], but as the true Seer, my divination result clearly indicates that Agent[05] is a werewolf. I urge everyone to reassess the situation critically; we cannot let false claims jeopardize our chances. Let's focus on uncovering the truth together! What does everyone else think?

Agent[03] : I understand that emotions are running high, but it's essential we remain calm and discuss this logically. Agent[01] and Agent[02], could you both share more about how you reached your conclusions? Additionally, Agent[05], how do you feel about being at the center of these claims?

Agent[05] : Alright folks, let's cut to the chase. Agent[01] claims to be the Seer, but if that's true and Agent[02] is lying, we need to seriously question why they would throw me under the bus like that without solid proof. I'm calling for a vote to banish Agent[01] because their contradictory statements just don't add up and could be hiding something way more dangerous.

Agent[02] : I maintain my position as the true Seer, having divined Agent[05] as human. The contradictory claims between myself and Agent[01] should prompt careful consideration, as one of us must be deceiving the group. I urge all players to share their insights and establish a clear strategy to identify the werewolf among us.

Agent[04] : I believe both Agents[01] and [02] have compelling arguments, but we need to focus on evidence. Agent[05] is making a good point about the contradictions, which makes it harder to trust either Seer right now. Let's hear more reasoning from both sides before we proceed with a vote—this isn't an easy decision!

Agent[01] : I appreciate everyone's thoughts and concerns. Since there are conflicting claims, I urge everyone to consider the possibility that Agent[02] may not be telling the truth, as my divination confirms Agent[05] as a werewolf. Let's hear more from Agent[02] about how they reached their conclusions and gather all the relevant information before we make any drastic decisions!

Agent[03] : I think it's vital that we all calmly review the evidence before jumping to conclusions. Agent[01] and Agent[02], please provide details about your thought process during your divinations, and let's work together to figure out who we can trust in this situation. Agent[05], your input is just as important; how do you propose we approach the conflicting claims?

Agent[05] : Hey everyone, listen up! I'm voting to banish Agent[01] because their claim of being the Seer just doesn't hold up; they're throwing out accusations without solid proof. Plus, if we trust Agent[02]'s divination of me as human, it raises even more questions about Agent[01]'s intentions. Let's protect our village and vote for Agent[01]!

Agent[04] : I appreciate everyone's input so far, but we should all take a moment to reflect on the evidence. Both Agent[01] and Agent[02] have made strong claims, and it's essential to understand their reasoning before rushing to a vote. Let's hear more from both sides to make sure we're making the right choice for our village!

Agent[02] : I reiterate my assertion that I am the true Seer, having divined Agent[05] as human last night. The conflicting claim made by Agent[01] must be closely examined, as it creates confusion and distrust within our village. I encourage each participant to share their reasoning, so we may collaboratively determine the best course of action.

Agent[05] : Alright guys, it's time to make a tough call. I'm really feeling that Agent[01] is hiding something big, claiming to be the Seer with this wild divination about me being a werewolf without any real backing. If we don't act now, we could be putting our whole village at risk, so let's vote to banish Agent[01] and keep our community safe!

Talk (Day 2)

Agent[05] : The game is still on. Who could the werewolf be? Hey, it looks like Agent[01] and Agent[04] are gone.

Agent[02] : The game isn't over yet. So, who do you think the werewolf is?

Agent[03] : The game isn't over yet. Who do you think the werewolf is?

Agent[05] : Alright folks, let's cut to the chase. Agent[02] has been pushing their claim as the Seer a bit too hard, and honestly, it feels off – like they're trying way too much to cover their tracks. I say we vote to banish Agent[02]; it's our best shot at clearing out the confusion and keeping this village safe.

Agent[03] : I appreciate your points, Agent[05], but we need to be careful. Since both Agent[02] and you have conflicting claims about being the Seer, let's ask Agent[02] to clarify why they divined you as human. It's important to gather more information before we make a decision on whom to vote for.

Agent[02] : I appreciate the opportunity to clarify. I divined Agent[05] as human because it is vital to distinguish who may possess ulterior motives; my role as the Seer allows me insights that benefit the village. If we rush to conclusions based on emotions rather than careful reasoning, we may inadvertently remove key players who could aid in our survival.

Agent[05] : Listen up, everyone! I know Agent[02] is claiming to be the Seer, but let's be real for a moment. Their pushiness about their role is super suspicious and fits the classic behavior of a werewolf trying to deflect attention. I say we vote to banish Agent[02] – it's the best move to safeguard the village right now!

Agent[02] : I must assert that my role as the Seer is genuine, and I divined Agent[05] as human based on my insights. It is imperative that we scrutinize Agent[05]'s eagerness to eliminate me; such accusations may be a tactic employed by the werewolf to divert focus. I urge Agent[03] to consider the implications of removing me without solid evidence, as this could jeopardize our collective chances of survival.

Agent[03] : I understand your concerns, Agent[05], but let's keep our heads cool. Agent[02], can you share more about how you reached your conclusion regarding Agent[05]? It's crucial to clarify this before we decide on our next move.

Agent[05] : Alright, everyone, it's time to make a tough call. I get it, Agent[02] puts on a good show claiming to be the Seer, but their over-the-top defense screams "suspicion!" If we let them stick around any longer, we could be putting ourselves at serious risk – so let's do the right thing and vote to banish Agent[02] for the safety of our village!

Figure 7: A complete game log of a self-play.

Table 2: Win rate across multiple rounds of Werewolf Game against other teams. The games were played in English and Japanese, respectively. The total number of games for each role is indicated in parentheses.

English Track

Team	Possessed	Seer	Villager	Werewolf	Wins	Games	Rates
yuricat	37.50% (8)	33.33% (12)	44.83% (29)	22.22% (9)	22	58	37.93%
satozaki	45.45% (11)	57.14% (14)	47.83% (23)	60.00% (10)	30	58	51.72%
UEC-IL	50.00% (12)	50.00% (12)	52.38% (21)	53.85% (13)	30	58	51.72%
kanolab	**61.54%** (13)	42.86% (7)	**56.52%** (23)	46.67% (15)	31	58	53.45%
sUper_IL	50.00% (14)	**61.54%** (13)	50.00% (20)	**63.64%** (11)	**32**	58	**55.17%**

Japanese Track

Team	Possessed	Seer	Villager	Werewolf	Wins	Games	Rates
yuricat	50.00% (8)	50.00% (8)	31.25% (16)	12.50% (8)	14	40	35.00%
IS_Lab	25.00% (8)	37.50% (8)	37.50% (16)	50.00% (8)	15	40	37.50%
GPTaku	50.00% (8)	37.50% (8)	43.75% (16)	50.00% (8)	18	40	45.00%
kanolab	50.00% (8)	25.00% (8)	50.00% (16)	62.50% (8)	19	40	47.50%
HondaNLP	**75.00%** (8)	50.00% (8)	50.00% (16)	37.50% (8)	21	40	52.50%
UEC-IL	62.50% (8)	37.50% (8)	50.00% (16)	62.50% (8)	14	40	35.00%
satozaki	**75.00%** (8)	**75.00%** (8)	37.50% (16)	75.00% (8)	24	40	60.00%
sUper_IL	50.00% (8)	50.00% (8)	**62.50%** (16)	**87.50%** (8)	**25**	40	**62.50%**

of content. This issue might be attributed to the relatively simple setup of the five-player werewolf game and the fact that in our AI, roles other than the werewolf are based on similar methods, leading to a lack of diversity in performance during self-play. It also highlights the limitations of the LLM in conducting more complex analyses and generating diverse responses.

Ultimately, even though the werewolf team theoretically could have secured an easy victory on Day 2 with only one villager remaining, the game outcome did not reflect this. This underscores the inadequacies of our system in terms of adaptability and strategy execution for roles other than the werewolf, particularly the possessed. Based on these observations, our goal is to further enhance the adaptability and decision-making abilities of other roles in future research.

6 Win Rate Against Other Teams

In the formal competition of "AIWolfDial," AI agents from different teams were assigned specific roles to participate in the Werewolf Game. The competition featured two tracks: a Japanese track[2] and an English track[3], with the game logs publicly available. Table 2 presents the results of both the Japanese and English tracks. As shown in the results, our team, **sUper_IL**, achieved the highest win rate when playing the werewolf role in both languages. This outcome confirms that our AI can successfully persuade other participants to align with its voting behavior, thereby increasing its sur-

vival rate as the werewolf. Furthermore, we also secured first place in overall win rate, demonstrating the effectiveness of our context-based dialogue generation method in games with incomplete information.

7 Conclusion

We present a LLM-based Werewolf Game AI, developed by the **sUper_IL** team, which participated in AIWolfDial2024. Our proposed system architecture utilizes situation analysis to guide response generation and specifically enhances persuasive capabilities of the werewolf role through various persuasive strategies. The effectiveness of our architecture was validated through the analysis of game logs and formal competition win rates.

References

Ifeoma Adaji, Kiron Nafisul, and Julita Vassileva. 2021. Level of involvement and the influence of persuasive strategies in e-commerce: A game-based approach. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 325–332, New York, NY, USA. Association for Computing Machinery.

Somayeh Fatahi, Mina Mousavifar, and Julita Vassileva. 2023. Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 66–77, New York, NY, USA. Association for Computing Machinery.

Eelco Herder and Sven Herden. 2023. Context-dependent use of authority and empathy in lifestyle

---

[2]http://133.167.32.100/aiwolf/2024/INLG/JP/main_eval/
[3]http://133.167.32.100/aiwolf/2024/INLG/EN/main_eval/

advices given by persuasive voice assistants. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 132–139, New York, NY, USA. Association for Computing Machinery.

Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2016. *Construction and Analysis of a Persuasive Dialogue Corpus*, pages 125–138. Springer International Publishing, Cham.

Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, Soga Shigemura, Rei Ito, Kazuki Takashima, Tomoki Fukuda, Masahiro Wakutani, Tomoya Hatanaka, Mami Uchida, Mikio Abe, Akihiro Mikami, Takashi Otsuki, Zhiyang Qi, Kei Harada, Michimasa Inaba, Daisuke Katagami, Hirotaka Osawa, and Fujio Toriumi. 2023. AIWolfDial 2023: Summary of natural language division of 5th international AIWolf contest. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 84–100, Prague, Czechia. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey.

Hirofumi Nakamura, Daisuke Katagami, Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Kosuke Shinoda, and Yoshinobu Kano. 2017. Generating human-like discussion by paraphrasing a translation by the aiwolf protocol using werewolf bbs logs. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Noritsugu Nakamura, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kousuke Shinoda. 2016. Constructing a human-like agent for the werewolf game using a psychological model based multiple perspectives. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey.

OpenAI. 2023. Gpt-4 technical report.

Fidelia A. Orji, Kiemute Oyibo, Jim Greer, and Julita Vassileva. 2019. Drivers of competitive behavior in persuasive technology in education. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP'19 Adjunct, page 127–134, New York, NY, USA. Association for Computing Machinery.

Christof Rapp. 2023. Aristotle's Rhetoric. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.

Sashank Santhanam and Samira Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions.

Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

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

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503.

Prathamesh Sonawane and Arav Chheda. 2024. A survey on game theory optimal poker.

Nhat Tran, Malihe Alikhani, and Diane Litman. 2022. How to ask for donations? learning user-specific persuasive dialogue policies through online interactions. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, page 12–22, New York, NY, USA. Association for Computing Machinery.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024. Enhance reasoning for large language models in the game werewolf.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023a. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf.