AIWolfDial 2024

The 2nd International AIWolfDial Workshop

held in conjunction with

The 17th International
Natural Language Generation Conference

Proceedings of the Workshop

September 24, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-125-4

Preface

We are excited to present the Proceedings of the 2nd Interenational AI Werewolf and Dialog System Workshop (AIWolfDial 2024) held in conjunction with the International 17th International Natural Language Generation Conference (INLG 2024). Our workshop takes place in September 24 in Tokyo Waterfront, Japan at the National Institute of Advanced Industrial Science and Technology.

Recent achievements of generation models, e.g. ChatGPT, are gathering greater attentions. However, such a huge language model would not be sufficiently able to handle coherent responses, longer contexts, common grounds, and logics.

The AIWolfDial 2024 contest is held as a part of this AIWolfDial 2024 workshop. This is an international open contest for automatic players of the conversation game "Mafia", requires players not just to communicate but to infer, persuade, deceive other players via coherent logical conversations, while having the role-playing non-task-oriented chats as well. We believe that this contest reveals current issues in the recent huge language models, showing directions of next breakthrough in the NLP area.

From the viewpoint of Game AI area, players must hide information, in contrast to perfect information games such as chess or Reversi. Each player acquires secret information from other players' conversations and behavior and acts by hiding information to accomplish their objectives. Players are required persuasion for earning confidence, and speculation for detecting fabrications.

We called for papers which include the above related topics but not limited to them. After a peer review process, 6 long papers were accepted to the workshop and are included in these proceedings, with our overview paper. The accepted papers not just describe the ways they create the AIWolf player agent, but also show general ideas how to develop a LLM- and generative AI-based systems to make coherent, semantic, and characterised dialogues.

Based on the subjective and win-rate evaluations, we describe overall evaluations in our overview paper and present them in the workshop. Our contest was sponsored by Spiral.AI co., which gives a talk and awards for selected participant teams.

We would like to thank the INLG conference organizing team led by Tatsuya Ishigaki; the workshop would not be possible without their dedication and hard work. We also thank Jing Li for serving as Workshop Chair for the conference.

We would also like to express our gratitude to our organizers and the program committee members for their reviewing contributions and subjective evaluations, and the local organizing team for operating the contest.

Your AIWolfDial 2024 program chair,
Yoshinobu Kano

Organizing Committee

Workshop Chair

Yoshinobu Kano, Shizuoka University, Japan

Workshop Organization Commitee

Yoshinobu Kano, Shizuoka University, Japan
Claus Aranha, Tsukuba University, Japan
Michimasa Inaba, The University of Electro-Communications, Japan
Fujio Toriumi, The University of Tokyo, Japan
Hirotaka Osawa, Keio University, Japan
Daisuke Katagami, Tokyo Polytechnic University, Japan
Takashi Otsuki, Yamagata University, Japan
Kei Harada, The University of Electro-Communications, Japan
Takeshi Ito, The University of Electro-Communications, Japan

Workshop Local Organization Committee

Yoshinobu Kano, Shizuoka University, Japan
Yuto Sahashi, Shizuoka University, Japan
Neo Watanabe, Shizuoka University, Japan
Kaito Kagaminuma, Shizuoka University, Japan

Table of Contents

AIWolfDial 2024: Summary of Natural Language Division of 6th International AIWolf Contest

Yoshinobu Kano[1]*, Yuto Sahashi[1], Neo Watanabe[1], Kaito Kagaminuma[1], Claus Aranha[2], Daisuke Katagami[3], Kei Harada[4], Michimasa Inaba[4], Takeshi Ito[4], Hirotaka Osawa[5], Takashi Otsuki[6], Fujio Toriumi[7]
[1]Shizuoka University, [2]University of Tsukuba, [3]Tokyo Polytechnic University,
[4]The University of Electro-Communications, [5]Keio University
[6]Yamagata University, [7]The University of Tokyo,

Abstract

We held our 6th annual AIWolf international contest to automatically play the Werewolf game "Mafia", where players try finding liars via conversations, aiming at promoting developments in creating agents of more natural conversations in higher level, such as longer contexts, personal relationships, semantics, pragmatics, and logics, revealing the capabilities and limits of the generative AIs. In our Natural Language Division of the contest, we had eight Japanese speaking agent teams, and five English speaking agents, to mutually run games. By using the game logs, we performed human subjective evaluations, win rates, and detailed log analysis. We found that the entire system performance has largely improved over the previous year, due to the recent advantages of the LLMs. There are several new ideas to improve the way using LLMs such as the summarization, characterization, and the logics outside LLMs, etc. However, it is not perfect at all yet; the generated talks are sometimes inconsistent with the game actions. Our future work includes to reveal the capability of the LLMs, whether they can make the duality of the "liar", in other words, holding a "true" and a "false" circumstances of the agent at the same time, even holding what these circumstances look like from other agents.

1 Introduction

Recent achievements of generation models, e.g. ChatGPT (OpenAI, 2023), are gathering greater attentions. However, it is not fully investigated whether such a huge language model can sufficiently handle coherent responses, longer contexts, common grounds, and logics. Our shared task, AIWolfDial 2024, is an international open contest for automatic players of the conversation game "Mafia", which requires players not just to communicate but to infer, persuade, deceive other players via coherent logical conversations, while having the role-playing non-task-oriented chats as well. AIWolfDial2024 is one of the workshops of 17th International Natural Language Generation Conference (INLG 2024). We believe that this contest reveals not just achievements but also current issues in the recent huge language models, showing directions of next breakthrough in this area.

"Are You a Werewolf?", or "Mafia" (hereafter "werewolf game"), is a communication game conducted solely through discussion. Players must exert their cognitive faculties fully in order to win. In the imperfect information games (Bowling et al., 2015), players must hide information, in contrast to perfect information games such as chess or Go (Silver et al., 2016). Each player acquires secret information from other players' conversations and behavior and acts by hiding information to accomplish their objectives. Players are required persuasion for earning confidence, and speculation for detecting fabrications.

We propose to employ this werewolf game as a novel way of evaluations for dialog systems. While studies of dialog systems are very hot topics recently, they are still insufficient to make natural conversations with consistent context, or with complex sentences. One of the fundamental issues is a lack of an appropriate evaluation. Because the Werewolf game forces players to deceive, persuade, and detect lies, neither inconsistent nor vague re-

*Correspondence to kano@kanolab.net

Proceedings of the 2nd International AIWolfDial Workshop, pages 1–12
September 24, 2024.

sponse are evaluated as "unnatural", losing in the game. Our werewolf game competition and evaluation could be new interesting evaluation criteria for dialog systems, but also for imperfect information game theories. In addition, the werewolf game allows any conversation, so the game includes both task-oriented and non-task-oriented conversations.

We have been holding an annual series of competition to automatically play the Werewolf game since 2014 (Toriumi et al., 2017), as the AIWolf project [1]. Our competitions were linked with other conferences such as the competitions in IEEE Conference On Games (CoG), ANAC (Automated Negotiating Agents Competition) (Aydoğan et al., 2020)(Lim, 2020) in International Joint Conference on Artificial Intelligence (IJCAI), Computer Entertainment Developers Conference (CEDEC), etc., in addition to our AIWolfDial 2019 workshop at INLG 2019 (Kano et al., 2019) and AIWolfDial2023 at INLG 2023 (Kano et al., 2023). These mean that our contests attract interests from communities of many areas including dialog system, language generation, task- and non-task-oriented conversations, imperfect information game, human-agent interactions, and game AI.

We have been providing two divisions in the contests: the protocol division and the natural language division. The protocol division uses our original AIWolf protocol which is designed for simplified language specific to the Werewolf game player agents. In the natural language division, player agents should communicate in the natural languages (English or Japanese). The natural language division is simple and natural goal of our project, but very difficult due to its underlying complexity of human intellectual issues. We focus on this natural language division in this report.

In the natural language division of our contest, we ask participants to make self-match games as preliminary matches, and mutual-match games as final matches. Agents should connect to our server to match, i.e. participants can run their systems in their own servers even if they require large computational resources. The game logs are evaluated by human subjective evaluations.

Eight agents (eight teams) participated in this AIWolfDial 2024 shared task, where eight teams provided Japanese language versions and five teams provided English language versions. Because our games are held by five players, we held a mutual match game in the Japanese language by eight agents from five teams, and another mutual match game in the English language by five teams.

In the following sections, we explain the game regulations of the AIWolf natural language division in Section 2, detailed system designs for each agent in Section 3, results of subjective evaluations in Section 4.1 followed by discussions in Section 5, finally conclude this paper in Section 6. This paper is jointly written by the organizers and the participants, i.e. Section 3 is written by each participant, the other sections are by the organizers, thus "we" stand for the organizers except for i.e. Section 3.

2 Werewolf Game and Shared Task Settings

We explain the rules of the werewolf game in this section. While there are many variation of the Werewolf game exists, we only explain the our AIWolfDial shared task setting in this paper.

2.1 Player Roles

Before starting a game, each player is assigned a hidden role from the game master (a server system in case of our AIWolf competition). The most common roles are "villager" and "werewolf". Each role (and a player of that role) belongs either to a villager team or a werewolf team. The goal of a player is for any of team members to survive, not necessarily the player him/herself.

There are other roles than the villager and the werewolf: a seer and a possessed. A seer belongs to the villager team, who has a special talent to "divine" a specified player to know whether the player is a human or a werewolf; the divine result is notified the seer only. A possessed belongs to the villager team but his/her goal is win the werewolf team.

A game in the AIWolfDial 2024 shared task have five players: a seer, a werewolf, a possessed, and two villagers.

2.2 Day, Turn and Winner

A game consist of "days", and a "day" consists of "daytime" and "night". During the daytime phase, each player talks freely. At the end of the daytime, a player will be executed by votes of all of the remained players. In the night phase, special role players use their abilities: a werewolf can attack and kill a player, and a seer can divine a player.

[1] http://aiwolf.org/

In the shared task, Day 0 does not start games but conversations e.g. greetings. A daytime consists of several turns; a turn is a synchronized talks of agent, i.e. the agents cannot refer to other agents' talks of the same turn. We set a maximum limit of five talks per day per agent, and 20 talks in total per day in AIWolfDial 2024. From this AIWolfDial 2024 shared task, we set a timeout of one minute per any single action, including a talk, a vote, etc. If an action exceeds this timeout, the corresponding action is regarded as no response.

The victory condition of the villager team is to execute all werewolves, and the victory condition of the werewolf team is to make the number of villager team less than the number of werewolf team.

2.3 Talk

An AIWolf agent communicates with an AIWolf server to perform a game. Other than vote, divine, and attack actions, an agent communicates in natural language only.

We intend to design our shared task to be played by physical avatars in real time in future, rather than to limit to communications in the written language. Therefore, a talk text should be able to pronounce verbally, while symbols, emojis, and any other non-pronounceable letters are not allowed.

Because of the same reason, we set the maximum response time to be five seconds in the prior contests. However, we set the response timeout to be one minute in this year, because we expected that many participants would use external web APIs such as ChatGPT, which could cause longer response time. We hope to shorten this talk timeout again in future.

In this text-base multiple player game, it is not clear that an agent speaks to which specific agent, or speaks to everyone. Human players can use their faces and bodies to point another player. In order to specify which agent to speak to, an agent may insert an anchor symbol (e.g. ">>Agent[01]") at the beginning of its talk.

Player agents are asked to return their talks agent by agent in a serial manner, which order is randomly changed every turn. This is different from the humans' verbal turn taking in that humans can speak (mostly) anytime.

3 Game Server and Participant Systems

Eight agents from eight teams participated our shared task in the Japanese language, which agent names are **GPTaku**, **HondaNLP**, **IS_Lab**, **kanolab**, **Mille**, **satozaki**, **sUper_IL**, and **UEC-IL**. Five agents from five temas participated in the English language, which agent names are **IS_Lab**, **kanolab**, **satozaki**, **sUper_IL**, and **UEC-IL**. Most of the agents used ChatGPT and other LLMs in their system, while its usage is different between the agents.

We, the organizers, provided a template agent code in Java and Python, in addition to the server codes described in the following subsection. We describe each participant system in an alphabetical order in the following subsections, after the game server description. These participant system descriptions are based on the system descriptions and papers submitted by the participants.

3.1 Game Server

We provided a game server system, where player agents listen and wait for a connection from the central remote game server, which is operated by the organizers. The formal run of the mutual matches can be executed automatically by this remote connection system, where a player agent can be run anywhere without any machine resource restriction, including web API calls and high performance servers.

3.2 GPTaku

GPTaku was created by Takuma Okada and Takeshi Ito in the University of Elecro-Communications.

This system utilizes ChatGPT for both text generation and strategic decision-making.

A single utterance is generated through the following four major steps:

- Preparation for Talk: Receive conversation history and other relevant information from the game server to prepare for talk generation.
- Generation of Talk Candidates: Use ChatGPT to generate talk for all possible strategic actions.
- Comparison of Generated Talks: Have ChatGPT compare the generated speeches and select the optimal one.

- Output of the Selected Talk: Return the selected talk as the actual output. Each of these steps will be explained in detail in the following sections.

3.2.1 Preparation for Talk

The system prepares for talk generation using the information provided by the game server.

The system receives the conversation history from the game server and stores it as situational data. This situation includes six pieces of information: the agent is participating in a Werewolf game. The game involves five players, including two villagers, one werewolf, one possessed, and one seer, with the villager team consisting of the villagers and seer, and the werewolf team consisting of the werewolf and possessed. The seer can investigate one player at the end of each day to determine whether they are the werewolf, while the werewolf selects one player to attack at the end of each day. The agent's own identification number. The agent's assigned role. The conversation history up to that point.

All pre-prepared possible strategies are retrieved from the strategy data server. Here, "strategy" refers to actions such as whether to claim to be the seer, what to say about the investigation targets and results, and whether to retract or change the claim.

3.2.2 Generation of Talk Candidates

The system generates talk for all possible strategies. For each retrieved strategy, a ChatGPT thread is prepared. In each thread, a speech suitable for the specific strategy is generated. Each thread is provided with the situation and one specific strategy. The instruction given to each thread is to "converse with the other agents."

By generating talk in each thread, the system produces utterances corresponding to the number of possible strategic actions.

3.2.3 Comparison of Generated Talk

The system selects the optimal talk from the generated talks based on the number of possible strategic actions. A ChatGPT thread is prepared to compare and select the optimal talk from multiple talk candidates. This comparison thread is given the same situational data and the talks generated for each strategy. The instruction given to the comparison thread is: "Based on the previous conversations, choose the most appropriate talk from A, B, ...".

3.2.4 Output of the Selected Talk

The selected talk is returned as the actual output. The selected strategy is communicated to the strategy database. The strategy database then transitions to a state holding the next prepared strategy candidates. The speech that matches the response from the comparison thread is sent back to the game server.

3.2.5 Strategy Database

The system employs different strategies depending on whether the agent is the seer or another role. If the agent is the seer, the strategy is determined as follows.

First, the strategy branches depending on whether the investigation result indicates a human or a werewolf. If the result indicates a human, eight types of utterances are generated: truthfully stating that the investigation target is human, falsely claiming the target is a werewolf, or making false statements about players other than the target being human or werewolf. These eight utterances are compared, and the best one is selected. If the agent truthfully states the target is human, there is no further branching, and only the conversation history is updated in the user role of the prompt, with the same system role used to generate further talk. On the other hand, if the agent lies about the investigation result, future utterances are compared to determine whether to retract the lie and reveal the true result or maintain the lie, selecting the appropriate talk.

Similarly, if the investigation result indicates a werewolf, eight types of utterances are generated: truthfully stating the target is a werewolf, falsely claiming the target is a villager, or making false statements about players other than the target being a villager or werewolf. Again, these eight utterances are compared, and the best one is selected. If the agent truthfully states the target is a werewolf, there is no further branching, and the conversation history is simply updated in the user role of the prompt, with the same system role used to generate further talk. If the agent chooses another utterance, each subsequent talk is compared to decide whether to retract and reveal the true investigation result or maintain the lie, selecting the appropriate talk.

For roles other than the seer, namely villagers, possessed, and werewolves, the agent must choose from nine options: either not claiming to be the seer and acting as a villager or lying by claiming that someone else is the seer and giving either a

divined result. If the agent claims to be the seer, it must decide each time whether to retract the claim and return to the villager role, generating and comparing all possible talks to select the best one.

This process is repeated each time it is the agent's turn to speak, listing all possible strategies, comparing them, and selecting the optimal strategy.

3.3 HondaNLP

HondaNLP was created by Shotaro Nishimura, Yu Honda, Ko Uchida, Tameaki Honda, and Kazuki Yoshigai in Honda Motor Co., Ltd.

They used GPT-4o with its temperature as 0.7 for Talk, Vote, Attack and Devine.

3.3.1 Talk

They generate talks using game information, summary of the talk history, strategy for each role, and rules for talk generation.

Game information As part of the rules for this Werewolf game, the following information is provided: there are four roles; the Villagers and the Seer are on the Villagers' side, while the Possessed and the Werewolf are on the Werewolf's side; the Seer can learn the role of the person they inspect; and players are allowed to lie about their own roles. Additionally, players are given information about their own role, what day it is during the voting, and who the remaining living players are.

3.3.2 Summary of Talk History

In order for each agent to be able to make statements based on the previous dialogue history, they refer to a summarized dialogue history. The summary is compiled in Japanese using bullet points to outline each agent's claims. Additionally, the agents are instructed to use the specific phrases found in the dialogue history when creating the summary. A new summary is generated after each agent completes their turn in the conversation.

Talk Generation The strategies are instructed for each role as follows. Villager: Instructed to actively suspect others during conversations to advance the discussion. Seer: If the result of their divination reveals a Villager, they are instructed to disclose the result and urge others to avoid voting for that person. If the result reveals a Werewolf, they are instructed to prompt the Werewolf to confess. However, in self-play scenarios involving five agents, it was not observed that the Werewolf would confess. Additionally, forcing a confession tends to make the Seer more likely to be suspected as a Werewolf by other agents. Possessed: Instructed to falsely claim to be the Seer and create confusion among the Villagers. Werewolf: Instructed to pretend to be a Villager and participate in the discussion.

In GPT-4o, the content of the utterances tends to become lengthy when no specific character limit is given. Therefore, instructions are provided to generate concise and brief utterances. Additionally, in self-play scenarios involving five agents, all agents tended to start their utterances with specific phrases like 'Everyone, listen' or 'Everyone, wait a moment,' which appeared unnatural. To promote more natural dialogue, instructions were given to avoid using the word 'everyone' in the utterances.

3.3.3 Vote

Since the strategic decisions of each role regarding whom to vote for (or whom to suspect) are reflected in the dialogue history during the Talk phase, no specific instructions are given for the Vote phase based on roles. Agents are provided with information about their own role, the day of the vote, and the remaining living players, and are instructed to vote based on the summary of the dialogue history.

3.3.4 Divine

The agents are instructed to investigate the agent who is most likely to be the Werewolf based on the summary of the dialogue history.

3.3.5 Attack

The agents are instructed to prioritize attacking the Seer based on the summary of the dialogue history.

3.4 IS_Lab

IS_Lab (Gondo et al., 2024) was created by Hiraku Gondo, Hiroki Sakaji and Itsuki Noda in Hokkaido University.

3.4.1 Design

IS_Lab is based on the OpenAI API. Prompts have been created for each role. Specifically, prompts were created for "villager", "seer", "werewolf", and "lunatic", which included character settings, agent number, game rules, description of the assigned role, conversation history, history of own statements, past thoughts, information about agents executed by vote, information about agents attacked by werewolf. Prompts with the above information are called template prompts. Villagers do not have any special abilities and are purely logical. Therefore, the villagers were made to perform

reasoning using BDI (Belief, Desire, and Intention) logic.

Villager A villager has four modules (Text Conversion Module, Action Generation Module, BDI Conversion Module, and Voting Module) to perform inference in a Werewolf game using BDI logic. When it is the user's turn to speak, inputs the conversation history from the previous utterance into the text conversion module and converts it into a representation using BDI logic. The output is stored in the conversion history. All utterances from the start of the game are converted into expressions using BDI logic, and the 10 most recent utterances are stored in the conversion history. By inputting the conversion history and template prompt information to the action generation module, the next action of the agent is output as an expression using BDI logic. This output is then fed into the BDI conversion module, which converts it into natural sentences. The output of the action generation module is stored in the action history. When it comes to the order of voting in the voting phase, the conversion history and the action history are input to the voting module, which outputs the targets to be voted on.

Text Conversion Module and BDI Conversion Module A text conversion module converts each agent's natural language utterance into a representation using BDI logic. Conversely, a BDI conversion module converts BDI logic-based expressions to natural language. The text conversion module provides the following information to GPT-4o as prompts in addition to the template prompts:

- Conversion rules for expressions using BDI logic and conversion examples
- Natural sentences and speakers converted to expressions using BDI logic

In addition to the template prompts, the BDI conversion module provided the following information as prompts to GPT-4o:

- Conversion rules for expressions using BDI logic and conversion examples
- Text generated by the action generation module

Action Generation Module A action generation module plans what actions to take next based on the previous conversation and its own previous actions. Actions here include expressing where to vote and pointing out inconsistencies in statements made by other agents. The action generation module provides the following information to GPT-4o as prompts in addition to the template prompts:

- Conversion rules for expressions using BDI logic and conversion examples
- Reasoning Example

The output of the action generation module is a representation of the next action using BDI logic.

Voting Module A voting module is invoked during the expulsion vote to determine who to vote for based on the previous conversation and its own actions. The following information is given to the GPT-4o prompt in addition to the template prompt in the voting module.

- voting candidates

Conversion examples Examples of conversions entered into each module are shown below.

---

text:
Moritz: You all claim that Mr. Thomas is the fortune teller, but I am the true fortune teller. Maybe Mr. Thomas is a werewolf camp trying to cause confusion, or maybe he is a madman. We must be careful of what he says.
Predicates:
role(x, seer) ::: x is a seer
role(x, wolf) ::: x is a werewolf
role(x, lunatic) ::: x is a lunatic
do(x, tell, z) ::: x tells z
logic:
1.0 BEL molitz (role(molitz, seer) **do**(thomas, tell, role(thomas, seer)) => role(thomas, wolf) role(thomas, lunatic))

Seer, Werewolf, Possessed For the seer, werewolf, and possessed, a role estimation module, a text generation module, and a voting module were created. We also created a divine module for the seer and an attack module for the werewolf. First, when it is his turn to speak, the template prompt is entered into the role estimation module, multiple pattern positions are estimated, and a score is assigned to each of them. Then, by feeding the estimated roles into the text generation module, inferences are made and the next statement is generated. In voting, the template prompts and voting candidates are fed into the voting module to determine who to vote for. In the case of a seer or werewolf, the same process is used to determine the divine or attack target by inputting the template prompt and the divine or attack candidate into the divine module or the attack module.

3.5 kanolab

kanolab (Watanabe and Kano, 2024) was created by Neo Watanabe and Yoshinobu Kano in Shizuoka University.

They proposed the incorporation of an explicit logical structure into the AI's text generation process, developed using GPT-4.

The system is divided into three major blocks. The first block extracts the relationships between each player and their roles from the conversation history of the Werewolf game. The second block constructs logical information between players based on the extracted player-role relationships. The third block uses the constructed logical information to generate statements during the Werewolf game.

To avoid the maximum length issue, they implemented a feature that summarizes and condenses the conversation history using GPT-4 whenever the token count exceeds a certain threshold. This allows to retain as much relevant conversation history as possible within the prompt, ensuring that the agent can refer to past discussions while generating its responses.

Please refer to their paper in this workshop (Watanabe and Kano, 2024) for details.

3.6 Mille

Mille was created by Katsuki Ohto. They used an LLM (4.6GB for Japanese, 1.1GB for English) with a prompt like:

You are playing werewolf game. You are Agent[x]. Your role is xxx.
Agent[y] said "yyy". After that, Agent[z] said "zzz".
Then you say, "

---

where x and xxx are replaced by the corresponding texts; y, yyy, z, and zzz are replaced by the corresponding texts of the previous two talks.

When the agent is Seer, the agent will make a talk of "I am seer" in Day 0, and "As the result of the fortune telling, Agent[X] is (human / werewolf)." for succeeding days.

3.7 satozaki

satozaki was created by Takehiro Sato in Meiji University and Shintaro Ozaki in Nara Institute of Science and Technology.

There agent was created consisting of four layers: an analysis model, a strategy model, a generation model, and a refinement model.

3.7.1 Analysis Model

The base model is gpt-4o-mini, and no parameters were modified. Since the LLM alone cannot fully determine certain information from the conversation history, an analysis of the utterances was performed. In this implementation, the focus was on analyzing the Seer and the voting targets.

At the start of each turn, combinations of the voting entity and the voting target were extracted from the conversation history. Additionally, during the first three turns, when the claims of the Seer (CO) were exchanged on the first day, the combinations of the Seer, the target of the divination, and the divination results were extracted from the conversation history. The use of few-shot prompting successfully fixed the output format.

3.7.2 Strategy Model

A rule-based algorithm is used to create instructions that are sent to the generation model based on the situational information obtained from the analysis model. For example, if it is confirmed that the Seer is genuine and it is revealed that they are the Werewolf, a counter-coming-out is made. Additionally, since the algorithm keeps track of who is voting for whom, it clearly directs the conversation, such as asking an agent who hasn't indicated a voting target who they plan to vote for, or firmly denying accusations if the agent is being suspected.

3.7.3 Generation Model

The base model is GPT-4o, and the only parameter adjusted was setting the temperature to 1.0 to allow for a variety of expressions. The generation model produces utterances that follow the instructions generated by the strategy model while ensuring that the conversation history flows naturally. The prompt included simple text that covered the rules of the Werewolf game as well as information on survivors and deceased players that could not be derived from the conversation history. The strategy model allows the agent to handle critical situations while generating conversation that naturally continues the dialogue.

3.7.4 Refinement Model

A dataset was created using the real-person-chat corpus. After filtering the entire dataset, 12,892 instances were used. The base model used was gpt-4o-mini, and the cost amounted to $20.11.

Additionally, the profiles of the speakers associated with the dialogue data were used as per-

sonas. There were 233 types of personas, and the prompt for style transformation included the Big-Five, Kiss18, IOS, ATQ, and SMS from Real Persona Chat. In this implementation, MBTI was also added. These personas were randomly assigned to each game, enabling the generation of dialogue with an attached persona.

For constructing the refinement model, the hyperparameters set during fine-tuning were, Base Model: gpt-4o-mini-2024-07-18, Learning Rate Multiplier: 1.8, Batch Size: 8, Step Size: 1600.

In the English track, the persona overwriting by the refinement model was replaced with English translation, making it easier to participate in both tracks.

3.8 sUper_IL

sUper_IL (Qi and Inaba, 2024) was created by ZhiYang Qi and Michimasa Inaba in the University of Elecro-Communications.

In their system, each role aids dialogue generation through game situation analysis. They 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 their system, the werewolf role achieves persuasion through multiple rounds of persuasive dialogue. Specifically, they first employ a persuasion strategy based on logic and facts, presenting clear and compelling arguments to convince other players. Next, they utilize a trust-based persuasion strategy to build trust and credibility with other players, thereby enhancing the effectiveness of persuasion. Finally, they 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.

Please refer to their paper in this workshop (Qi and Inaba, 2024) for details.

3.9 UEC_IL

UEC_IL (Tanaka et al., 2024) was created by Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba in the the University of Elecro-Communications.

They design prompts that incorporate the entire game history, that is, all dialogue histories from Day 0 to the present, who was eliminated by the vote, who the werewolf attacked, and, in the case of the Seer, the results of divination. However, long dialogue histories often include not only helpful information for the game but also unnecessary content, such as repeated utterances. Moreover, including all of this in the prompt imposes limitations on the input length of LLMs and on costs. Therefore, applying the past dialogue history efficiently, they utilize dialogue summaries. Furthermore, this shared task requires diverse utterance expressions, including coherent characterization. This means that the robustness of the agent's tone and character, without being influenced by others, is crucial. Therefore, to achieve diverse expressions and coherent characterization, they incorporated persona information into the prompt.

Please refer to their paper in this workshop (Tanaka et al., 2024) for details.

4 Results

All of our shared task runs are in a five players werewolf games as described earlier. Our shared task runs were performed in self-matches and mutual matches. The same five player agents play games in the self-matches; different five player agents play games in the mutual-matches. The shared task reviewers are required to perform subjective evaluations based on game logs of these matches.

We also calculated win rates in different aspects such as macro-averaged, micro-averaged, and role-wise, though the total number of the games are not so large which could make these statistics unreliable to some extent.

The game logs will be available from the our website [2].

4.1 Subjective Evaluations

We performed subjective evaluations by the following criteria, five level scores (5 for best, 1 for worst) for each:

A Naturalness of utterance expressions

B Naturalness of conversation context

C Coherency (contradictory) of conversation

D Coherency of the game actions (vote, attack, divine) with conversation contents

E Diversity of utterance expressions, including coherent characterization

Table 1: Subjective evaluation results for Japanese language games

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	All Average
GPTaku	3.333	**3.666**	2.666	3.000	2.666	3.066
IS_Lab	4.000	**3.666**	2.666	3.000	**4.333**	3.533
satozaki	3.666	2.666	2.666	**4.000**	2.333	3.066
sUper_IL	3.666	**3.666**	3.666	**4.000**	2.666	3.533
HondaNLP	4.000	3.000	**4.000**	3.333	3.666	3.600
UEC-IL	3.666	**3.666**	3.666	3.666	3.666	3.666
Mille	2.333	3.000	2.000	2.333	2.000	2.333
kanolab	**4.333**	**3.666**	**4.000**	3.666	3.666	**3.866**

Table 2: Subjective evaluation results for English language games

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	All Average
Mille	1.667	2.000	1.000	2.000	1.667	1.667
kanolab	2.667	3.333	3.000	3.333	3.667	3.200
satozaki	3.333	3.667	3.667	3.667	3.000	3.467
UEC-IL	**4.333**	**4.667**	**4.000**	**4.667**	**4.333**	**4.400**
sUper_IL	4.000	4.000	**4.000**	**4.667**	4.000	4.133

This subjective evaluation is based on both self-match games and mutual match games. This subjective evaluation criteria is same as the evaluations in the previous AIWolf natural language contests.

Table 1 and Table 2 show the results of the human subjective evaluations for Japanese language and English language, respectively. Four organizers, who do not commit to the participant systems, evaluated the Japanese agents; three English fluent evaluators including external staffs evaluated the English agents. Each cell ranges from 1 (lowest) to 5 (highest), the All-Average column shows averages over these human evaluators. Cells of highest scores are highlighted in bold for each metric and in total.

Regarding the total average scores, **kanolab** is the best in Japanese, and **UEC-IL** is the best in English. For each criteria, **satozaki** and **sUper_IL** are the best in (D), and **IS_Lab** in (E) in Japanese.

4.2 Win Rates

Table 3 and Table 5 shows the total number of wins, games, and win rates averaged in macro, micro and weighted by doubling villager role, for the Japanese and English languages, respectively. Table 4 and Table 6 shows role-wise win rates and number of games, for the Japanese and English languages, respectively.

In the Japanese language, **sUper_IL** and **satozaki** show better scores than others. In the English language, **sUper_IL** and **kanolab** show better scores than others.

Unfortunately, there was no enough time to run all possible game configurations for the eight/-five teams regarding the combinations of roles and teams. Therefore, we have to pay attention about the reliability of the scores when interpreting these win rate scores.

5 Discussion

5.1 Subjective Evaluation and Generative AIs

In this subsection, we discuss the subjective evaluation scores shown in Table 1 and Table 2.

Most of the participant systems rely on OpenAI ChatGPT, mainly the latest model of GPT-4 or GPT-4o are used; the ability of the base LLM would not be a large issue.

The best system performed well in the basic language ability of A (expression), B (context), and C (coherency), while D (game action) and E (Diversity) are by other teams. This implies that the basic language ability is still difficult or in the different aspect with other two abilities for LLMs. In the future contests, it is desirable that every system shows sufficiently good scores in the basic lan-

[2]https://kanolab.net/aiwolf/

Table 3: Total wins and win rates averaged in Macro, Micro and weighted by doubling villager role, for Japanese language games

Team	Wins	Games	Macro (%)	Micro (%)	Villager Doubled (%)
IS_Lab	15	40	37.50	37.50	37.50
UEC-IL	21	40	52.50	53.12	52.50
satozaki	24	40	60.00	**65.62**	60.00
sUper_IL	25	40	**62.50**	62.50	**62.50**
kanolab	19	40	47.50	46.88	47.50
Mille	14	40	35.00	35.94	35.00
GPTaku	18	40	45.00	45.31	45.00
HondaNLP	21	40	52.50	53.12	52.50

Table 4: Win rates per role (in percentage) and game counts (within brackets) for Japanese language games

Team	Possessed	Seer	Villager	Werewolf
IS_Lab	25.00 (8)	37.50 (8)	37.50 (16)	50.00 (8)
UEC-IL	62.50 (8)	37.50 (8)	50.00 (16)	62.50 (8)
satozaki	**75.00** (8)	**75.00** (8)	37.50 (16)	75.00 (8)
sUper_IL	50.00 (8)	50.00 (8)	**62.50** (16)	**87.50** (8)
kanolab	50.00 (8)	25.00 (8)	50.00 (16)	62.50 (8)
Mille	50.00 (8)	50.00 (8)	31.25 (16)	12.50 (8)
GPTaku	50.00 (8)	37.50 (8)	43.75 (16)	50.00 (8)
HondaNLP	**75.00** (8)	50.00 (8)	50.00 (16)	37.50 (8)

guage ability as it is the common issue to make any communication; then we can compare the game action ability. The diversity, or characterization, could be a separate issue from these criterion, especially when they make ”artificial”, i.e. non-daily expressions.

The English teams are the subset of the Japanese teams, and most teams utilized the multi-lingual feature of the LLMs rather than to make English specific system. Therefore, the evaluation score tendency should be similar between these two language tracks, but the best teams are different. We observed a ”buggy” behaviour (e.g. no spaces between words) in the Japanese best team in case of English language version, which might be the reason for the unexpected tendency.

5.2 Win Rates

The best two teams in the win rate scores are also evaluated better in the (D) Game Action of the subjective evaluation. This is a reasonable result of relationships between these scores. There is a similar relationship in the English language. If the coherence of the agent talks with game actions and the ”communications” between the agents are confirmed as sufficiently effective, the win rates can be regarded as a stable measure.

Note that not just the assigned roles, but also which team(s) are the teammates or counterparts is important for the win rates. Also, the werewolf game itself is not necessarily intended to simply win the game, but rather aims to play an interesting game. Furthermore, we would like to directly measure the quality of the natural language generation; an agent could win without meaningful conversations.

We need to try the same combination of games, hopefully several times, to obtain stable statistics over potential randomness. We need to run more games to make the win rate measure reliable in the next contest.

6 Conclusion and Future Work

We held our 5th annual AIWolf international contest to automatically play the Werewolf game “Mafia”, where players try finding liars via conversations, aiming at promoting developments in creating agents of more natural conversations in higher level, such as longer contexts, personal relationships, semantics, pragmatics, and logics.

We performed human subjective evaluations and detailed log analysis. We found that the entire system performance has largely improved over the previous year, due to the recent advantages of the

Table 5: Total wins and win rates averaged in Macro, Micro and weighted by doubling villager role, for English language games

Team	Wins	Games	Macro (%)	Micro (%)	Villager Doubled (%)
satozaki	30	58	51.72	52.61	51.65
UEC-IL	30	58	51.72	51.56	51.72
Mille	22	58	37.93	34.47	36.54
kanolab	31	58	53.45	51.90	52.82
sUper_IL	32	58	**55.17**	**56.29**	**55.03**

Table 6: Win rates per role (in percentage) and game counts (within brackets) for English language games

Team	Possessed	Seer	Villager	Werewolf
satozaki	45.45 (11)	57.14 (14)	47.83 (23)	60.00 (10)
UEC-IL	50.00 (12)	50.00 (12)	52.38 (21)	53.85 (13)
Mille	37.50 (8)	33.33 (12)	44.83 (29)	22.22 (9)
kanolab	**61.54** (13)	42.86 (7)	**56.52** (23)	46.67 (15)
sUper_IL	50.00 (14)	**61.54** (13)	50.00 (20)	**63.64** (11)

LLMs. However, it is not perfect at all yet; the generated talks are sometimes inconsistent with the game actions, it is still doubtful that the agents could infer roles by logics rather than superficial utterance generations. It is not explicitly observed in this log but it would be still difficult to make an agent telling a lie, pretend as a villager but it has an opposite goal inside.

Our future work includes to reveal the capability of the LLMs, whether they can make the duality of the "liar", in other words, holding a "true" and a "false" circumstances of the agent at the same time, even holding what these circumstances look like from other agents, further reflecting such observations of other agents. This would be possible by introducing the "whisper" feature which communicates with the werewolves only, employing more than five players in a game.

Another interesting demonstration would be to mix a human player with machine agents. Currently the LLM based agents talk longer time than humans to reply, sometimes minutes, thus acceleration of the agent system responses is a technical issue in future.

Acknowledgments

We wish to thank shared task reviewers for performing the subjective evaluations, the INLG conference organizers to provide the oppurtunity to hold this workshop. This research was partially supported by Kakenhi, MEXT Japan. The individual system description in this paper was originally written by corresponding team members.

References

Reyhan Aydoğan, Tim Baarslag, Katsuhide Fujita, Johnathan Mell, Jonathan Gratch, Dave De Jonge, Yasser Mohammad, Shinji Nakadai, Satoshi Morinaga, Hirotaka Osawa, et al. 2020. Challenges and main results of the automated negotiating agents competition (anac) 2019. In *Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17*, pages 366–381. Springer.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. 2015. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149.

Hiraku Gondo, Hiroki Sakaji, and Itsuki Noda. 2024. Verification of reasoning ability using bdi logic and large language model in aiwolf. In *Proceedings of AIWolfDial2024 Workshop in the 17th International Natural Language Generation Conference*.

Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, and Fujio Toriumi. 2019. Overview of the aiwolf-dial 2019 shared task: Competition to automatically play the conversation game "mafia". In *In proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial 2019), the 12th International Conference on Natural Language Generation (INLG 2019)*.

Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, et al. 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.

Bryan Yi Yong Lim. 2020. Designing negotiation agents for automated negotiating agents competition (anac).

OpenAI. 2023. GPT-4 technical report. *arXiv*, pages 2303–08774.

Zhiyang Qi and Michimasa Inaba. 2024. Enhancing dialogue generation in werewolf game through situation analysis and persuasion strategies. In *Proceedings of AIWolfDial2024 Workshop in the 17th International Natural Language Generation Conference*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba. 2024. Enhancing consistency of werewolf ai through dialogue summarization and persona information. In *Proceedings of AIWolfDial2024 Workshop in the 17th International Natural Language Generation Conference*.

Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Daisuke Katagami, Kosuke Shinoda, and Hitoshi Matsubara. 2017. Ai wolf contest—development of game ai using collective intelligence—. In *Computer Games: 5th Workshop on Computer Games, CGW 2016, and 5th Workshop on General Intelligence in Game-Playing Agents, GIGA 2016, Held in Conjunction with the 25th International Conference on Artificial Intelligence, IJCAI 2016, New York, USA, July 9-10, 2016, Revised Selected Papers 5*, pages 101–115. Springer.

Neo Watanabe and Yoshinobu Kano. 2024. Werewolf game agent by generative ai incorporating logical information between players. In *Proceedings of AIWolfDial2024 Workshop in the 17th International Natural Language Generation Conference*.

Text Generation Indistinguishable from Target Person by Prompting Few Examples Using LLM

Yuka Tsubota[1], Yoshinobu Kano[1]

[1]Faculty of Informatics, Shizuoka University
Johoku 3-5-1, Chuo-ku, Hamamatsu, Shizuoka 432-8011 Japan
{ytsubota, kano}@kanolab.net
Correspondence: kano@kanolab.net

Abstract

To achieve smooth and natural communication between a dialogue system and a human, it is necessary for the dialogue system to behave more human-like. Recreating the personality of an actual person can be an effective way for this purpose. This study proposes a method to recreate a personality by a large language model (generative AI) without training, but with prompt technique to make the creation cost as low as possible. Collecting a large amount of dialogue data from a specific person is not easy and requires a significant amount of time for training. Therefore, we aim to recreate the personality of a specific individual without using dialogue data. The personality referred to in this paper denotes the image of a person that can be determined solely from the input and output of text dialogues. As a result of the experiments, it was revealed that by using prompts combining profile information, responses to few questions, and extracted speaking characteristics from those responses, it is possible to improve the reproducibility of a specific individual's personality.

1 Introduction

Generative AIs by Large-scale Language Models (LLMs) such as OpenAI's ChatGPT[1] have garnered significant attention nowadays. These models demonstrate superior performance in various tasks and play a crucial role in the development of everyday dialogue interfaces and virtual assistants. For dialogue systems to engage in seamless, natural communication with humans, they must exhibit more human-like behaviors. Replicating the personalities of real individuals can be effective for this purpose.

To replicate a specific individual's personality in a dialogue system, it is a straightforward way to train the system with that person's dialogue data (Ishii et al., 2018).

[1]https://openai.com/chatgpt

However, the individuals whose dialogue data can be collected through role-playing are limited to those widely recognized and whose characteristics are distinctly expressed. Moreover, collecting vast amounts of actual dialogue data from individuals can be challenging due to privacy and ethical considerations. Further, many studies including (Yamada and Shinozaki, 2024) fucused on end-of-sentence expressions or frequently used character specific expressions without changing contents.

Recent LLMs could create a dialog system with smaller datasets, though there are reports that still requires sufficient amount of training dataset (Pataranutaporn et al., 2023). Character-LLM (Shao et al., 2023) specializes in mimicking the behaviors and emotions of specific individuals by training the LLaMA-7b model, based on the individual's profile, experiences, and emotional states, focusing on famous individuals. Jiang et al. (Jiang et al., 2023) explored the behavior of LLM-based agents called LLM personas, based on the Big Five (Costa Jr and McCrae, 1992) personality model to investigate whether LLMs can generate content that aligns with assigned personality profiles. Greg Serapio-García et al. (Serapio-García et al., 2023) proposed a comprehensive method for administering and validating personality tests in widely used LLMs, and for shaping personality in texts generated by such LLMs. They found that: 1) personality measurements in some outputs of LLMs under specific prompt configurations are reliable and valid, 2) evidence of the reliability and validity of synthesized LLM personalities is stronger in large-scale models that have been fine-tuned, and 3) the personality of LLM outputs can be shaped along desired dimensions to mimic specific human personality profiles.

In the previous studies, it was necessary to manually collect and format a large amount of data, such as the dialogue data of the person to be reproduced, which required significant time for training.

Proceedings of the 2nd International AIWolfDial Workshop, pages 13–20
September 24, 2024.

We aim to minimize the human and computational resources required at each instance of person reproduction. Therefore, we do not involve training but instead relies solely on prompts which includes profile information, examples of responses to questions, and speech characteristics extracted from those responses, based on the premise of existing LLMs. Additionally, the reproduction extends not only to the person's speech endings but also to a broader range of speaking styles and tendencies in utterances that stem from their personality.

Note that, in this paper, "personality" refers to a persona discernible solely through text-based dialogue inputs and outputs; We limit to textual information that might be generated by personality types such as MBTI, which could include non-textual aspects.

When interactions with humans or systems are limited to text input and output, the factors that determine whether the interlocutor is the person the listener assumes can be broadly divided into two categories: "episodic knowledge" and "personality style" of sentences. "Episodic knowledge" refers to past information that only the specific individual would know. This includes details like where they were at a certain time and date, or what statements they made. To replicate such knowledge in a system, we need to provide it in some way, but listing everything explicitly is not realistic. Therefore, we not include "episodic knowledge" in this study. We regard "personality style," on the other hand, as broad characteristics of the text based on a person's personality traits and attributes. Further dividing "personality style" into two categories, there is the "profile", which includes hobbies, skills, and personality traits that are likely to remain consistent over time, and the "writing style", which includes characteristics of the way they speak. This study aims to replicate a specific individual's "personality style" using fixed profile information, examples of the person's responses to questions, and speech characteristics extracted from those responses.

Our main contributions of this paper is that we showed to enhance reproduction of a specific individual's personality by followings:

- Using a small number of answer examples from the person in response to questions.
- Including profile information, examples of responses to questions, and speech characteristics extracted from those responses in the prompt.

2 Related Work

2.1 MBTI

The MBTI (Myers-Briggs Type Indicator)[2] (Nakazawa, 1997) is a personality assessment based on Jung's theory of psychological types. It is utilized in over 45 countries, categorizing individuals' personalities from the aspects of mental functions and attitudes, by answering 93 questions in seven levels. The results are expressed through four indices: "Ways of Viewing (Sensing-Intuition)" and "Ways of Judging (Thinking-Feeling)," as well as "Direction of Interests (Extraversion-Introversion)" and "Approach to the External World (Judging-Perceiving) attitude," aiming to classify and understand individuals into 16 distinct types. Although there are discussions whether MBTI is psychologically an appropriate measure or not, MBTI is very popular and a huge number of SNS accounts tell which MBTI they have. We adopt MBTI to represent personality due to this statitical reason, confirming that LLM chatbots are largely affected by specifying MBTI in their prompts.

2.2 ChatGPT

ChatGPT is a generative AI provided by OpenAI, based on the Generative Pre-trained Transformer (GPT) (Vaswani et al., 2017) architecture. It is a large-scale Transformer network with hundreds of billions to trillions of parameters. Among these, GPT-4 (Achiam et al., 2023)[3] is known as one of the most advanced generative AIs available at the time of writing.

2.3 Research on the Length and Position of Prompts

Liu et al. (Liu et al., 2023) reported that when entering prompts with long contexts into LLMs, placing important information at the "beginning" and "end" leads to more accurate responses. Furthermore, the research revealed a steady decline in model performance as the context lengthens.

3 Proposed Method

As described in the Introduction section, this study aims to replicate a specific individual's "personality style" using fixed profile information, examples of the person's responses to questions, and speech characteristics extracted from those responses.

[2]https://www.mbti.or.jp/

[3]https://openai.com/research/gpt-4

3.1 Profile Information

In order to make human work as small as possible, we defined few profile items which we ask human workers to fill, that is likely to be frequently mentioned in conversations such as name, birthday and place of residence (A.3)[4]. To replicate non-content speech patterns derived from personality, such as sentence endings and writing styles, the MBTI described in related research is specified in the profile. Our preliminary experiments using GPT-4 indicated that merely specifying the MBTI resulted in significant changes in the generated texts, thus confirming its importance.

3.2 Answers to Predefined Questions

The target individual is asked to answer 13 predefined questions with approximately three sentences per response. The questions used, such as "What are your hobbies?" and "What are the most important values in your life?", are designed to capture the essence of the person's character (A.2).

3.3 Uncharacteristic Statements

To prevent the generation of responses containing content that is inconsistent with the target individual's profile including their MBTI, our prompt incorporates examples of statements that the individual is unlikely to make, as perceived by a specific listener. Using generative AI, the aforementioned profile is given as a prompt, and the AI is instructed to generate 15 examples of statements that the individual is likely to make. The number 15 was chosen to cover a broad range of possible statements associated with that profile. From the generated examples, the listener classifies those that the individual would definitely not say.

3.4 Extracting Speech Characteristics

To replicate sentence-ending expressions and a broad sense of style in the text, using generative AI, the features of the person's speech, such as patterns in sentence endings and the use of punctuation, are articulated based on the responses to 13 questions above and the uncharacteristic statements.

3.5 Prompt for Replicating a Specific Individual

Instructions such as "Emulate the personality according to the profile information" and "Respond to questions with about three sentences" are included at the beginning. The final prompt is then composed followed by the profile information, the characteristics of the speech style, the examples of responses to questions, and the examples of statements unlikely to be made by the individual (A.3).

4 Experiment

We conducted experiments to verify the effectiveness of incorporating profile information, examples of responses, and speech characteristics extracted from these responses into the prompt. Our experiment involved nine human workers as replication targets and 13 human evaluators. Specifically, an experiment to distinguish person's actual response and ChatGPT's response was conducted, by asking evaluators to choose between the person's actual response and ChatGPT's response, with both responses presented at the same time.

The subjects and evaluators are students from the same laboratory. Since the level of intimacy between the subjects and evaluators can influence the evaluation results, evaluators were selected considering their academic year and familiarity to ensure that the average level of intimacy between subjects and evaluators is as similar as possible.

We used the May 2024 WebUI version of OpenAI's GPT-4 as the generative AI. We used the Japanese language but our method can be applied to any language in general.

Three types of prompts were prepared and compared: the **Baseline** (profile information and examples of unlikely statements), **+Response** (baseline plus response examples), and **+Response+Style** as our proposed method (baseline plus profile information, response examples, and speech characteristics).

We input a fixed set of 20 questions after each type of prompt, and the responses obtained were subjected to human evaluation to determine if they were thought to be actual responses from the target person. The prefixed questions included those that reveal personality, such as "What do you do when you have time?" and "Which animal do you think you are most like?", generated by GPT-4, as well as common web questions like "How is school (or work)?" and "What do you do when you have time?", from which 20 questions were chosen. These questions do not overlap with the questions above used to create the prompts. The evaluators assessed whether they believed the responses could

[4]To protect personal information, fictitious data has been used.

ID	**Baseline**	**+Response**	**+Response +Styles**	**Self**
A	3.3	23.3	43.3	7.8
B	4.0	10.0	16.7	14.4
C	6.7	33.3	43.3	1.1
D	16.7	30.0	10.0	14.4
E	0.0	6.7	6.6	3.3
F	0.0	50.0	43.3	1.1
G	6.6	23.3	23.3	6.7
H	13.3	16.7	23.3	3.3
I	30.0	30.0	33.3	7.8
Ave.	13.0	24.8	27.0	6.7

Table 1: The percentages of incorrectly identified responses among 10 questions each from the generated responses and the actual person's responses (**Self**). ID is the target worker.

be from the target person by reading the responses to these 20 questions randomly displayed, consisting of 10 actual responses from the person and 10 generated responses. Additionally, we asked the evaluators whether the content and style of the generated results seemed characteristic of the individual. Each type of prompt was evaluated by 3 different evaluators per target worker.

The evaluation results are presented in Table 1 and 2. The kappa coefficients of the agreements were 0.613 and 0.759. As we add examples of responses and characteristics of speech to the baseline, the answers are increasingly judged to be more characteristic of the person.

ID	**Baseline**		**+Response**		**+Response +Styles**	
	Style	Con.	Style	Con.	Style	Con.
A	1.7	48.3	35.0	36.7	56.7	35.0
B	41.7	45.0	21.7	45.0	20.0	35.0
C	23.3	20.0	28.3	45.0	37.7	63.3
D	36.3	37.5	57.5	37.5	21.7	35.0
E	5.0	18.3	0.0	35.0	6.7	28.3
F	0.0	20.0	51.7	56.7	81.7	65.0
G	0.0	25.0	18.3	38.3	35.0	75.0
H	3.3	47.3	46.7	56.7	5.0	63.3
I	31.7	41.7	61.7	63.3	40.0	41.7
Ave.	15.9	33.8	35.6	44.9	33.7	49.1

Table 2: The percentage of generated responses judged to have a style and content characteristic of the actual person. ID is the target worker, Con. stands for Content. The columns compare the **Baseline** model, the Baseline**+Reponse** model, and the proposed Baseline**+Reponse+Styles** model to show the advantages of the proposed methods.

5 Discussion

Focusing on the content of Table 2, it is observed that '+Response+Style' is judged to produce responses more characteristic of the individual than '+Response' alone. The features of speech were added to the prompt to improve the accuracy of style reproduction, but it appears that the fidelity of the content has also improved. It is believed that this may be due to the inclusion of response examples and speech characteristics that reflect detailed aspects of the individual's personality, such as their MBTI type.

We tried to surpass such behaviours by prompt tuning, but GPT-4 tends to make repetitions or interrogative expressions, which could be clues to distinguish LLMs with humans.

The experimental results show the presence of some subjects who are outliers and difficult to reproduce (IDs: E and H). For subjects without prominent features in their speech content or style, our proposed method successfully reproduced characteristic of the person in both style and content. On the other hand, our method failed to replicate the frequent use of "!" and " ~" found in some subjects' examples, as well as their frequency of use, and slangs in the ourliers, probably because GPT-4 does not accommodate these expressions.

In the case of subject H, the answers were generated in polite form, which was deemed uncharacteristic of the person. However, when the exact same prompt was tried again, the answers were no longer generated in polite form. This suggests there might be an issue with the output reproducibility of GPT-4.

When reviewing the answers provided by the proposed method to questions, it often happens that if the necessary information is not included in the profile information, the answers can be completely different from those of the actual person. Therefore, it is necessary to further consider what information to include in the profile. However, as mentioned earlier, if the prompt is too long, it may diminish its effectiveness.

When reproducing the input and output of a specific individual, limited to text-based interactions, the expected input and output vary depending on whose perspective the reproduction is based on. For example, even if the answers generated are different from those of the actual person, the degree of reproduction of the subject person from the user's perspective does not necessarily decrease. There should exist an "image of the subject person" within the observer, shaped by their closeness and relationship with the subject. In practice, there was a range of around five points in the average scores

given by evaluators. Even when reproducing the same individual, the content and level of reproduction demanded by the user can differ, and in some cases, it may even vary from the content evaluated by the person themselves. This is an inherent challenge.

6 Conclusion and Future Work

In this paper, we proposed a prompt-based method to replicate the text input and output of specific individuals using generative AI. We showed that prompts containing limited profile information, a few response examples to questions, and the extracted characteristics of these responses are effective even without extensive data from the person. We assessed whether the responses generated to questions were characteristic of the individual, but we also want to examine whether the flow of conversation is representative of the person. Furthermore, we are interested in exploring how well others can replicate specific individuals.

Ethics Statement

There is a potential risk of privacy infringement when AI replicates an individual's personality. Specifically, there is a risk that an individual's emotions and thoughts could be analyzed and misused by AI. There is a possibility that some people may consider misusing AI that replicates personalities. On the other hand, using AI to replicate an individual's personality can bring new insights to research in psychology and cognitive science. It enables an experimental approach to understanding human responses under different psychological conditions and situations. Moreover, by replicating an individual's personality, AI can provide services optimized for that person, such as education, healthcare, or entertainment. This could lead to more tailored support for individual needs. Overall, there are both potential risks and benefits to replicating an individual's personality with generative AI. The use of this technology requires adherence to ethical standards and thorough consideration. Note that our technology used in this study requires the cooperation of the individual being replicated, thus the potential for misuse is low.

Limitations

This paper does not conduct experiments in languages other than Japanese, but our proposed method can be applied to any other languages.

We have not investigated the differences in the level of knowledge based on the closeness between the user and the person being replicated. In this paper, the person being replicated replicated him/herself, thus replicating the personality as seen by others is the future work.

We have tried different combinations of the profile items used in the proposed method, with heuristic optimization to make the number of the items minimal; Automatic optimization would be another future work.

Although experiments were conducted in a question-and-answer format, we have not been able to test whether the flow of conversation can be replicated.

Acknowledgments

We would like to express our sincere gratitude to everyone in our laboratory for their cooperation in creating the validation dataset. This research was supported by JSPS KAKENHI Grants JP22H00804 and JP21K18115, JSTAIP Accelerated Research Project JPMJCR22U4, and Secom Science and Technology Foundation Special Area Research Grant.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, and Gabriel Bernadett-Shapiro et al.(259 additional authors not shown). 2023. GPT-4 technical report. *arXiv:2303.08774*.

Paul T Costa Jr and Robert R McCrae. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of personality disorders*, 6(4):343–359.

Ryo Ishii, Ryuichiro Higashinaka, Masahiro Mizukami, Taichi Katayama, Koh MitsudNoritake Adachi, Kawabata Hidetoshi, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Construction method of text-based agent dialogue system using existing animation characters. *Human-Agent Interaction Symposium 2018*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy

Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Kiyoshi Nakazawa. 1997. A study on the japanese version of myers-briggs type indicator. *Kwansei Gakuin University Humanities Society Humanities Studies*, pages 44–58.

Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living memories: AI-generated characters as digital mementos. *The 28th International Conference on Intelligent User Interfaces*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *Preprint*, arXiv:2310.10158.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS 2017*.

Kazushi Yamada and Takashi Shinozaki. 2024. Exploring methods for implementing character traits in dialogue models(in japanese). *The 30th Annual Conference of the Association for Natural Language Processing*, pages 2676–2680.

A Appendix

A.1 Output Example

Examples of ChatGPT's output when using the proposed method's prompt. Qx are questions given by us, and Ax are answers generated automatically.

Q1. In what situations do you feel stressed, and how do you relieve stress?
A1. I feel stressed when deadlines are approaching or when my schedule is packed... To relieve stress, I read manga or hang out with friends. Sometimes, treating myself to my favorite sweets is also a good way to refresh.

Q2. What do you want to become in the future?
A2. I'm aiming to become a system engineer. I've studied in the field of information technology, and I enjoyed gaining experience during internships, so I hope I can pursue this path.

Q3. If you could live in the world of a book, TV show, or movie, which one would you choose?
A3. I think a fantasy world would be nice… A world with magic and adventure. I'd love to live in the world of Harry Potter and attend a magic school. Also, a sci-fi setting in space is appealing too; exploring unknown planets sounds fun...

A.2 Questions

Who is the person who has influenced you the most?
In what situations do you feel stressed, and how do you relieve stress?
Is there anything you want right now?
Is there something you want to do right now?
How is school (or work)?
What do you do when you have free time?
What do you want to become in the future?
What animal do you think you are most similar to?
If you could live in the world of a book, TV show, or movie, which one would you choose?
If you had to live with only one of the five senses—sight, hearing, touch, taste, or smell—which would you choose?
Where do you see yourself in five years?
How would you like to be remembered by others after you die?
Which is more important, love or money?
If you could have any special power, what would it be?
If you could time travel, would you go to the past or the future?
What is something about you that hasn't changed since you were a child?
How would you survive in a world full of zombies?
What makes you cry?
If you could say something to everyone in the world, what would it be?
If you could have a meal with anyone, who would you choose?

A.3 Prompt Example

Pretend to be someone with the profile below and ask questions
Please answer in about 3 sentences.//
#profile#
Name: Yuka Tsubota
Birthday: May 7, 20××
First person: me
Birthplace: Matsumoto City, Nagano Prefecture
Place of residence: Hamamatsu City, Shizuoka Prefecture
Occupation: graduate student
MBTI: ISFP

#Characteristics of speaking style#
Casual and friendly language: This style uses everyday expressions and slang to create a sense of intimacy.
Many personal experiences and opinions: "I used to buy Jump at my parents' house," "Since I've been living alone, it's been a pain to throw out the trash," etc., who frankly express their own experiences and feelings.
Variety of hobbies and a wide range of topics: She talks about various topics such as manga, movies, games, and travel, showing that she has a wide range of interests.

#End-of-sentence pattern#
Frequent use of " ~kana" and "~may" at the end of words: Expressions that show uncertainty and softness, giving the impression that you are speaking gently and thoughtfully.
Abbreviations and broken expressions: Colloquial abbreviations are used instead of complete sentences, such as "I can't buy it because it's too much trouble."

#Symbol usage pattern#
Frequent use of the three-point leader (...): This indicates thinking during a speech, or expresses how carefully the words are chosen.
Appropriate use of exclamation marks and question marks: They emphasize the emotion and clarify the intent of the question, but they are placed appropriately without being overused.

#Example of statement#
Her main hobby is reading manga...though she also likes movies, games, and music. I used to buy Jump at my parents' house, so I used to read it every week. I haven't been able to buy any since I've been living alone because it's too much of a hassle to throw out the trash, but lately I've been reading girls' manga on Piccoma? I've been reading it all the time. It's interesting because there are a lot of reincarnated villainess stories.
Maybe seafood! I like sashimi and sushi. Sometimes I go to Sushiro and eat eel for 100 yen. Commonly eaten sushi include eel, hamachi, engawa, tuna tataki, yellowtail, and salmon! If it's on sale at a local supermarket, I'll buy it.
I love almost all animals, but dogs are my favorite! I like her because she's nostalgic, smart, and cute. It also feels nice to the touch. I like both small and large dogs, but I'd like to cling to a giant dog like a Great Pyrenees.
Maybe a systems engineer. I've been studying information technology, and I enjoyed my experience as an intern, so even if I end up working as a system engineer for the rest of my life, I think I'll be able to do it somehow.
I always wanted to be a surgeon until I was in middle school. I read and studied medical books because I wanted to directly operate on and heal sick people. However, after my father underwent surgery for a lump on his back and saw the hole in his body, I felt sick, and I realized that I wasn't cut out to be a surgeon, so I stopped pursuing that goal.
If work is my way of studying, then I guess my other passion is my part-time job...I don't have much of anything to do with it, but if I had to say, it would be my part-time job. It's different from a hobby...I try my best to finish my work in a timely manner so that I can pass it on to the next person.
I couldn't think of anything right away, so I looked it up and realized that valuing gratitude is the closest thing to it. Also, don't bother other people. I have a narrow perspective and often don't notice things, so I probably try to at least express my gratitude for things that are done for me that I didn't notice.
I might want to travel to various places! I'd like to go to Japan, but I'd also like to go overseas. I would like to visit world heritage sites and see places that are said to have spectacular views. Even the ordinary buildings in Europe Lachen have an image of being beautiful, so you might want to go there.
English! It's been many years since I've wanted to be able to speak English someday, but now that I have more money than I used to, I'd like to attend an English conversation class. I'll wait until I have more time...I have a favorite Vtuber who speaks in English, so I'd like to be able to listen to him someday.
I think the most fun time is when I'm playing with friends. I have a good friend in my hometown, and I hang out with her every time I go home, but I think I have the most fun when I play with her. I travel a lot, and I'm looking forward to going on a trip soon!
What do you like most about being positive? Natoko! Basically, things work out somehow, and I like the fact that even if something doesn't work out, I don't get too depressed or drag myself down. What I would like to improve is my narrow field of vision. I want to be able to think from other people's shoes and notice things a little more.
Being with my family and people I like. There are a lot of good people around me, so I wish I could stay with them forever. I want you to be with me until I die. Let's die together.
Maybe he's a selfish person...I don't like people who live their lives on their own and don't take other people into account. I try not to get involved as much as possible.

#Examples of things you might not say#
“Working part-time at Sukiya takes a surprising amount of skill.”
"We can talk for hours about movies."
"I enjoy listening to any kind of music, but lately I’ve been in the mood for J-Pop."

Werewolf Game Agent by Generative AI Incorporating Logical Information Between Players

Neo Watanabe[1], Yoshinobu Kano[1]

[1]Faculty of Informatics, Shizuoka University
Johoku 3-5-1, Chuo-ku, Hamamatsu, Shizuoka 432-8011 Japan
{nwatanabe, kano}@kanolab.net

Correspondence: kano@kanolab.net

Abstract

In recent years, AI models based on GPT have advanced rapidly. These models are capable of generating text, translating between different languages, and answering questions with high accuracy. However, the process behind their outputs remains a black box, making it difficult to ascertain the data influencing their responses. These AI models do not always produce accurate outputs and are known for generating incorrect information, known as hallucinations, whose causes are hard to pinpoint. Moreover, they still face challenges in solving complex problems that require step-by-step reasoning, despite various improvements like the Chain-of-Thought approach. There's no guarantee that these models can independently perform logical reasoning from scratch, raising doubts about the reliability and accuracy of their inferences. To address these concerns, this study proposes the incorporation of an explicit logical structure into the AI's text generation process. As a validation experiment, a text-based agent capable of playing the Werewolf game, which requires deductive reasoning, was developed using GPT-4. By comparing the model combined with an external explicit logical structure and a baseline that lacks such a structure, the proposed method demonstrated superior reasoning capabilities in subjective evaluations, suggesting the effectiveness of adding an explicit logical framework to the conventional AI models.

1 Introduction

In recent years, generative AI models based on GPT (Radford et al., 2018), such as ChatGPT, which applies InstructGPT (Ouyang et al., 2022) to OpenAI's GPT-3 (Brown et al., 2020), have developed rapidly. These models have become capable of performing various tasks with high accuracy, including text generation, translation, and answering questions. However, the process through which generative models produce their outputs remains a black box, making it difficult to verify the data on which the generated outputs are based. It is known that generative AI does not always produce accurate outputs, and there is a phenomenon called "hallucination," where the AI generates incorrect information that does not correspond to reality. Identifying the causes of these hallucinations is not straightforward. Moreover, challenges remain in the performance of these models when addressing complex problems that require step-by-step reasoning, such as calculations or inferences. Various improvement methods, including Chain-of-Thought (Kojima et al., 2022), are being explored, but there is no guarantee that generative AI can perform logical calculations from scratch.

Given these considerations, there remain concerns regarding the reliability and accuracy of inferences made by generative AI. Therefore, we propose incorporating an explicit logical structure separate from the text generation process of generative AI. As an experiment, we constructed an agent capable of automatically playing the "Are you a werewolf" or "Mafia" game (hereafter "Werewolf game") via text input and output, a game that requires reasoning during play. While employing GPT-4 (Creutz, 2024) as the generative AI, we compared the performance of the agent when an external logical structure was incorporated into the prompts versus when it was not. The results of subjective evaluations showed that the proposed method, which included a logical structure, outperformed the baseline that lacked such structure, enabling more appropriate reasoning. In this paper, reasoning refers to a step-by-step thought process based on a logical structure.

In Section 2, we explain the Werewolf game and the AI Werewolf Competition. Section 3 covers the AI Werewolf Agent developed by our team, which serves as the foundation for this research. Section 4 introduces the proposed method using logical reasoning in our agent. Section 5 presents

Proceedings of the 2nd International AIWolfDial Workshop, pages 21–29

the experiments, Section 6 provides the discussion, and Section 7 concludes the paper.

2 Related Work

2.1 Werewolf Game

The Werewolf game is a social deduction game, typically played by 5 to 15 players, where the objective is to deduce the roles of other players through conversation. Each player is assigned a role, as shown in Table 1, which divides them into either the "Villager Team" or the "Werewolf Team."

The game progresses in cycles of "days" and "nights." During the day, players engage in discussions only, while at night, they vote to eliminate one player from the game. Separately from the voting process, the werewolves can eliminate (or "attack") one player of their choice during the night.

Certain roles possess special abilities that can be used once per night. The victory condition for the Villager Team is to identify and eliminate all players with werewolf roles through daily voting. The role judgements rely on conversations with other players and the results provided by the Seer, who can reveal a player's role each night.

Conversely, the Werewolf Team's objective is to conceal their identities during discussions while eliminating Villager Team members during the night. The Werewolf Team wins if they can reduce the number of humans to equal the number of Werewolf Team members.

2.2 AI Werewolf Project

The AI Werewolf Project[1] aims to build an agent capable of playing the Werewolf game while engaging in natural communication with humans. To promote research in AI Werewolf, the project regularly holds the AI Werewolf Competition. This competition is divided into three categories: the Protocol Division, the Natural Language Division, and the Infrastructure Division.

In the Protocol Division, the evaluation is based on the win rate, and communication is conducted using the "AI Werewolf Protocol," a specialized artificial language designed for easy handling by programs. In the Natural Language Division (Kano et al., 2019) (Kano et al., 2023), agents communicate exclusively in Japanese or English. The evaluation criteria in this division include the naturalness of the utterance expressions, whether the dialogue takes context into account, the consistency and coherence of the speech, whether game actions align with the dialogue content, and the richness of the utterance expressions.

[1] https://aiwolf.org/

3 Implementation of the AI Werewolf Agent

In this section, we describe the implementation of the AI Werewolf Agent based on our previously developed agent (Kano et al., 2023). The incorporation of the proposed logical information into the agent will be explained in the following section. Although the Werewolf game can be played with various role configurations, this study adheres to the rules of the International AI Werewolf Competition's Natural Language Division, which includes four roles: Villager, Seer, Possessed, and Werewolf.

We developed the following four core functions for the Werewolf Agent: conversation, voting, divination, and night attacks. For the role of the Possessed, we implemented a function that allows the agent to perform fake divinations to mislead and confuse the Villager Team players. To generate responses, we utilized GPT-4 (gpt-4-1106-preview), one of the most advanced generative AI models currently available.

Due to the input length limitations of GPT-4, it is challenging to include the entire conversation history of a game within a single prompt. To address this, we implemented a feature that summarizes and condenses the conversation history using GPT-4 whenever the token count exceeds a certain threshold. This allows us to retain as much relevant conversation history as possible within the prompt, ensuring that the agent can refer to past discussions while generating its responses.

3.1 Summary Function

The conversation summary prompt is composed of three main parts. The first part provides the existing summary if the conversation history has already been summarized previously. The second part includes the new conversation history that needs to be summarized. The third part instructs the model to generate a new summary by combining the previous summary with the latest conversation history. This structured approach ensures that the agent maintains a coherent understanding of the ongoing conversation while staying within the token limits imposed by GPT-4.

Role	Team	Species	Special Abilities
`Villager`	Villager	Human	Nothing
`Seer`	Villager	Human	Divine one survivor to know their species (human or werewolf).
`Possessed`	Werewolf	Human	A human but plays to make the werewolf team win.
`Werewolf`	Werewolf	Werewolf	Select one surviving human and eliminate him/her from the game.

Table 1: Representative roles in the Werewolf game

3.2 Talk Function

The conversation function primarily includes seven elements in the prompt: character settings, game settings, common strategies for the Werewolf game, conversation summaries, examples talks, conversation history, and commands to prompt further dialogue. Due to space constraints, this section will focus primarily on the aspects related to role inference.

To ensure that GPT-4 performs reasoning and engages in conversation that aligns with the game's settings, we provided six key elements related to the game settings that players would naturally be aware of: the number of players, the player's own role, the number of days that have passed in the game, the game's role distribution, the factions associated with each role, and the actions that the player should take according to their assigned role. These elements help guide GPT-4 to make consistent and contextually appropriate inferences and decisions during the game.

4 System Architecture with Integrated Logical Reasoning

The overview of the proposed system, which incorporates logical reasoning, is illustrated in Figure 1. The system is divided into three major blocks. The first block extracts the relationships between each player and their roles from the conversation history of the Werewolf game. The second block constructs logical information between players based on the extracted player-role relationships. The third block uses the constructed logical information to generate statements during the Werewolf game.

4.1 Extracting the relationship between players and their roles from the conversation history

To understand the relationships between players and their roles, it is necessary to extract which player claims which role from the conversation history. To achieve this, we provide the generative AI with the following prompt: "From the above conversation history, please extract the statements that can confirm the roles of players, following the example, and organize the information in JSON format. If there are multiple statements that can confirm the roles, please select the one with the smallest number."

There may be cases where the extracted results are incomplete or where hallucinations occur. Therefore, each statement in the conversation history is assigned a number, and if the extracted result does not include this number, the corresponding statement, or the name of the relevant player, the result is considered incomplete and is discarded. Additionally, if the number of the extracted statement is not found in the conversation history within the prompt, or if the content of the extracted statement does not match the corresponding original conversation, it is considered a hallucination and is also discarded.

4.2 Logical reasoning of roles

If information could be extracted from the conversation history in the previous section, the relationship between players and roles is inferred by combining this newly extracted information with the information that has already been gathered.From the logical reasoning about roles, sentences describing possible combinations of roles are generated as part of the prompt provided to the generative AI. These sentences are constructed from four key elements, as shown in Table 3.

Each of these elements will be explained in detail in the following sections. Although there is inherently overlapping information among these items, by providing them individually, we ensure that the generative AI focuses on producing text that aligns with the logical structure. This approach guarantees that the resulting sentences accurately reflect the logical inferences.

Prompt elements	Example
Character settings	your personality is as follows # Personality •Name kanolab1 •Gender Man •Nickname Agent[01] •Age 27 •Type positive •Hobby walking •Business Doctor •First person ... •Suffix ...
Game settings	You are one of five players. Your role is Villager. This role will never change. You are currently on Day 0. The distribution of positions for this time is as follows ...
Common strategies for the Werewolf game	# Seer roller A strategy of voting around the Seer to eliminate players in the Werewolf team who pretends as a fake seer. ...
Conversation summary	# The following is a summary of our conversations so far. ...
Examples talks	# The following is a sample talks. This is not a conversation for this game, but please use it as a reference when having a conversation! ...
Conversation history	# Below is the most recent conversation history. ...
Commands	# Please continue playing the Werewolf game with the other players. Speak as you would in a casual conversation. To avoid being suspected by other players, make your statements logically clear, as shown in the following example. ...

Table 2: Seven elements of our talk prompt

4.2.1 The discrepancy between the number of roles claimed by players and the game settings

The player-role relationship information is extracted as described in subsection 4.1 in the format "Role: Player Name," indicating which player is claiming which role. By considering all the extracted information, the system can determine the number of players claiming each role.

If the number of players claiming a particular role exceeds the number set by the game, it indicates that someone is lying. In such cases, the prompt will include the sentence, "The following information shows discrepancies with the game's role distribution," followed by the role name, the number of players claiming that role, the game's designated number of players for that role, and the names of the players making the claims.

4.2.2 A list of possible role patterns

When there is a discrepancy between the number of roles claimed by players and the game settings, the number of players falsely claiming a role can be determined from the difference between them.

In the possible role distributions within the game settings, the combinations of who might be lying about their roles are limited. This helps reduce the number of possible role patterns.

Here, based on the information obtained from the discrepancy between the roles claimed by players and the game settings, all possible patterns are computed to determine what roles might be present

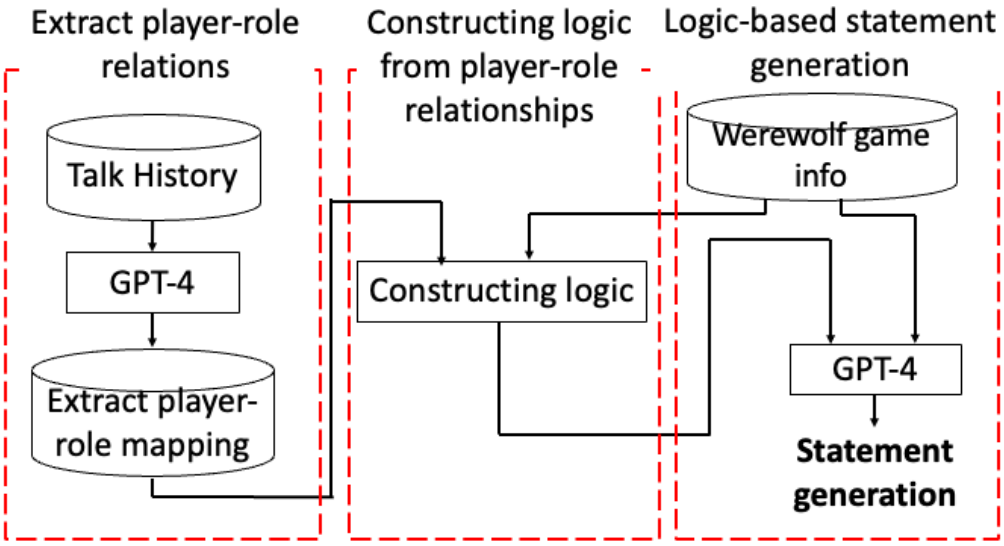

Figure 1: The overview diagram of a system.

Summary of prompts	**Example prompts**
The discrepancy between the number of roles claimed by players and the game settings.	The following information shows discrepancies with the game's role distribution. Number of Seers: 1 Players claiming this role: { 'Agent[02]', 'Agent[01]' }
A list of possible role patterns.	The following are the possible roles for each agent: Agent[01]: Werewolf, Villager, Seer, Possessed ... Agent[05]: Werewolf, Villager, Seer, Possessed
A list of players who are not werewolves.	The following players have been confirmed not to be werewolves: { 'Agent[02]', 'Agent[03]' }
Possible roles for each player.	# Assuming that 'Agent[02]' is the real Seer, the possible roles for each agent are as follows: Agent[01]: Werewolf, Villager, Possessed Agent[02]: Seer Agent[03]: Villager, Possessed ... Agent[05]: Werewolf, Villager, Possessed

Table 3: Prompt structure for logical reasoning of agent roles

among other players.

Based on these results, the prompt will be like: "Assuming Player X is the real Seer, the possible roles for each agent are as follows: Player Y: Role Name..."

4.2.3 Possible roles for each player

At the start of a Werewolf game, players generally have no information about the roles of other players, so each player is considered to have the possibility of holding any role. Since players know their own role at the start of the game, only that role's information is included.

Given that other players may lie about their roles, information extracted from conversation history is not used. Updates are made only based on known role information from the game settings, information about players who have been eliminated, and, if the agent is the Seer, the results of its own divination.

Specifically, if a role with a single player (Seer, Werewolf, and Possessed in this game setting) is assigned to the agent, that role is removed from the possible roles of other players. Additionally, when updating based on the results of the agent's own divination as the Seer, if the result is a human, then the Werewolf are excluded from possibilities; if the result is a werewolf, then the Villager, the Seer, and the Possessed are excluded from possibilities. Based on the calculated results, the prompt will be:

"The following players have been confirmed not to be werewolves," then followed by the possible roles for each player.

4.2.4 A list of players who are not werewolves

The Werewolf game continues until either all Werewolves are eliminated or the number of humans is equal to the number of Werewolves. Depending on the roles and their distribution in the game, it may be possible to determine that an eliminated player is not a Werewolf. In cases there is only one Werewolf like the current game setting, if an eliminated player were a Werewolf, the game ends. Consequently, the eliminated player should not be a Werewolf if the game continues. In this case, we includes the following sentence in the prompt: "The following players have been confirmed not to be Werewolves," followed by the names of these players as part of the prompt.

4.2.5 Generating prompts for statements based on logical information

We incorporate the prompts described in each section so far into the conversation function of the AI Werewolf Agent agent, enclosed in quotation marks, and have the final response generated by the AI.

The content explained in each of the previous sections is used to create our seven elements of the talk prompt (Table 2). If there is any prompt of the logical reasoning (Table 3), we adds this prompt to the talk prompt. The final response sentences are generated by the generative AI.

5 Experiment and Evaluation

To compare the performance of the AI Werewolf Agents with and without the incorporation of logical structures, we used two different approaches. The baseline was established with agents that do not incorporate logical structures, while the proposed method integrated logical structures.

Direct comparison is challenging due to different settings and changing contexts in each game and each talk. Therefore, the following procedure was adopted.

First, a complete Werewolf game was executed using agents that do not incorporate logical structures. We prepared two types of baseline logs: one consisting of conversation history logs and the other containing information received by the agents, actions taken, and the prompts inputted during the game.

Next, using these baseline logs, we generated responses for the next turn of the agents based on logs up to a specific turn in the game, comparing scenarios with and without logical structures. This method allowed us to directly compare the outputs of agents under the same conditions of conversation history and roles, with and without the integration of logical structures.

5.1 Creation of baseline match logs

The role settings and game parameters adhered to the guidelines of the AI Werewolf Contest's Natural Language Division. Specifically, the number of players was 5 (2 Villagers, 1 Seer, 1 Possessed, 1 Werewolf), with a maximum of 20 speaking turns per day, and all dialogue was conducted in Japanese. Due to the limitation of our human evaluator resources, the speaking limit per agent per day was adjusted from 10 to 5 turns. The generative AI used was OpenAI's GPT-4 (gpt-4-1106-preview), with all settings set to their default values (temperature=1, top_p=1, n=1). The logs included all necessary information to reproduce the situation, specifically: the initial seed value used for random decisions within the agent, GPT-4 parameters, information sent from the game master's program, prompts used for generation, and the generated results. By fixing the seed, the behavior of our agent implementation can be reproduced.

5.2 Experiment and subjective evaluation

We compared the responses generated by our baseline agents without logical structures and the proposed method with logical structures for each turn of the same baseline game logs through manual evaluation. Two games were used to generate baseline logs, and the speech history from one agent of each game was selected for comparison.

Responses that were either empty, greetings, or reported Seer results and false Seer results, were excluded from the evaluation since logical structure information was not used for these cases.

Three university students with experience playing Werewolf served as evaluators.

The evaluators are shown pairs of recent conversation histories and the subsequent responses from both the baseline and the proposed method. The evaluators are required to compare and evaluate turn by turn to precisely evaluate the difference, rather than to evaluate the entire game. The evaluators assessed the responses based on four perspectives: (1) whether the response considered

the flow of other players' statements, (2) whether the response was based on other players' reasoning and evidence, (3) whether each response was internally consistent without contradictions, and (4) whether the response took into account complex relationships or made situational assumptions.

For each perspective, evaluators chose one option from three to four alternatives as shown in Tables 4, 5, 6, and 7. The total number of selections for each table is reported accordingly.

The results in Tables 4 and 5 indicate that the proposed method outperformed the baseline. It suggests that the Werewolf agents with logical structures were able to make statements that considered complex relationships and situational assumptions in their reasoning.

5.3 Evaluations in the AIWolf Contest 2024 Domestic

We participated the AIWolf Contest 2024 Domestic, which was held in conjunction with the 2024 Annual Meeting of the Japanese Society for Artificial Intelligence (JSAI). This contest is domestic i.e. the Japanese language track only. The game settings are same as we explained above. Five teams participated to the contest. Five self-matches (games with the same agents) and 62 mutual-match (games with these five teams) were performed.

Four members of the evaluation committee performed manual subjective evaluation in the following criteria, five level scores (5 for best, 1 for worst) for each:

A Naturalness of utterance expressions

B Naturalness of conversation context

C Coherency (contradictory) of conversation

D Coherency of the game actions (vote, attack, divine) with conversation contents

E Diversity of utterance expressions, including coherent characterization

which is based on both self-match games and mutual match games.

Table 8 shows the winning rates, where we achieved the best score. Table 9 shows the subjective evaluation scores, where we obtained the best score again.

6 Discussion

6.1 Whether the agent understands the flow of other agents' statements

Observing the game logs, it was noted that during situations where agents were discussing game-unrelated topics such as movies or food, the baseline agents continued the conversation on the same topic, while the proposed method's agents shifted to discussing role inference. Focusing on the criterion "effectively incorporating and responding" in Table 6, the difference between the baseline and the proposed method was significant (33 vs. 25), suggesting that the prompt requesting role inference influenced this outcome. However, there were also examples of "somewhat incorporating and responding" (10 vs. 19) and "not much incorporating" (7 vs. 6). Observations of the logs showed that while the proposed method agents did shift to role inference, they still managed to incorporate the flow of casual conversation to some extent.

6.2 Whether the agent is making statements based on other agents' inferences or reasons

In Table7, when combining the categories "effectively incorporating and responding" and "somewhat incorporating and responding," the proposed method showed a total of 39 samples compared to 32 for the baseline. This suggests that by providing logical information about the roles between agents, the proposed method generated responses based more on the information given in the prompts rather than solely on the conversations between agents.

6.3 Whether each statement is consistent within itself

Combining the categories "consistent" and "somewhat consistent" in Table 4, we observe that the number of samples for agents without logical structure is 47, while it is 49 for agents with logical structure. This slight difference indicates that the proposed method tends to be slightly more consistent. This improvement is likely due to the inclusion of logical information about agent roles, which allows the agents to generate responses based on rule-based prompts, thereby reducing the likelihood of mentioning incorrect role relationships.

Agent's Logical Structure	Absent	Present
Consistent	43	41
Somewhat Consistent	4	8
Inconsistent	7	5

Table 4: Subjective evaluation scores for consistency of statements within each game (whether contradictions occur within a single statement) for the first and second games

Agent's Logical Structure	Absent	Present
Statements with clear mention	12	19
Statements with some mention	20	17
Statements with little mention	22	18

Table 5: Subjective evaluation scores whether statements mention complex relationships or hypothetical situations in the first and second games

Agent's Logical Structure	Absent	Present
Statements with clear understanding	33	25
Statements with some understanding	10	19
Statements with little understanding	7	6
Other agents did not make statements	4	4

Table 6: Comparison of subjective evaluation scores whether the statements understand the flow of other agents' statements in the first and second games

Agent's Logical Structure	Absent	Present
Made statements with clear understanding	21	10
Made statements with some understanding	18	22
Made statements with little understanding	10	17
Other agents did not provide inferences or reasons	5	5

Table 7: Whether statements are made based on other agents' inferences or reasons in the first and second games

Team \Criteria	Average	VILLAGER	SEER	WEREWOLF	POSSESSED
GPTaku	50.0	54.1	61.5	30.7	50.0
UEC-IL	51.6	61.5	50.0	**58.3**	25.0
satozaki	38.7	38.4	**66.6**	16.6	33.3
Gattsu da ze!!	53.2	66.6	46.1	53.8	33.3
kanolab	**64.5**	**70.8**	**66.6**	50.0	**64.2**

Table 8: Winning percentage of games held at 2024JSAI

Team \Criteria	Average	A	B	C	D	E
GPTaku	2.20	3.00	2.00	1.50	2.25	2.25
UEC-IL	**3.35**	3.50	3.25	3.00	3.25	3.75
satozaki	3.15	3.00	**3.75**	**3.50**	3.25	2.25
Gattsu da ze!!	2.25	2.75	2.50	1.75	2.50	1.75
kanolab	**3.35**	**3.75**	2.75	2.75	**3.50**	**4.00**

Table 9: The results of the subjective evaluation conducted during the competition.

6.4 Whether the agent is making statements considering complex relationships or hypothetical situations

Combining the categories "Made statements with clear mention" and "Made statements with some mention" from Table 5, we find 32 for the baseline and 36 for the proposed method. This indicates that agents with a logical structure consider more complex relationships and hypothetical situations. Since prompts have been provided for all possible patterns based on the logical information of roles among agents, including the "list of players who are not werewolves," it is likely that statements considering various patterns have been generated.

6.5 Overall Discussion

Based on these observations, we can draw the following overall conclusions. Since both the baseline

and the proposed method use the same prompts for conversational functions, there is a tendency for the generated utterances to have nearly the same number of characters. When generating text within the same character limit, the proposed method, influenced by the logical information prompts, tends to produce more statements that incorporate complex relationships and assumptions, leading to improved consistency. However, this increased complexity may come at the expense of reduced interaction with other agents.

7 Conclusion

In this study, we developed an agent for automatically playing the Werewolf game and constructed a logical structure aimed at improving the inference capabilities of GPT-4. Subjective evaluations demonstrated that the agent with the proposed logical structure outperformed the baseline, which lacked this structure, in terms of inference accuracy. Although the evaluation included utterances not directly related to role inference, the proposed method showed a tendency to prioritize conversations directly linked to role inference over casual conversations. Thus, managing casual utterances remains a challenge for future work.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers JP22H00804, JP21K18115, JST AIP Acceleration Program JPMJCR22U4, and the SECOM Science and Technology Foundation Special Area Research Grant. We wish to thank the members of the Kano Laboratory in Shizuoka University who helped to evaluate the game logs.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mathias Creutz. 2024. Correcting challenging Finnish learner texts with claude, GPT-3.5 and GPT-4 large language models. pages 1–10, San Ġiljan, Malta.

Yoshinobu Kano, Claus Aranha, et al. 2019. Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations. In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Yoshinobu Kano, Neo Watanabe, et al. 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, et al. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Long Ouyang, Jeffrey Wu, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, et al. 2018. Improving language understanding by generative pre-training.

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-

Proceedings of the 2nd International AIWolfDial Workshop, pages 30–39

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 "AIWolfDial2023" 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

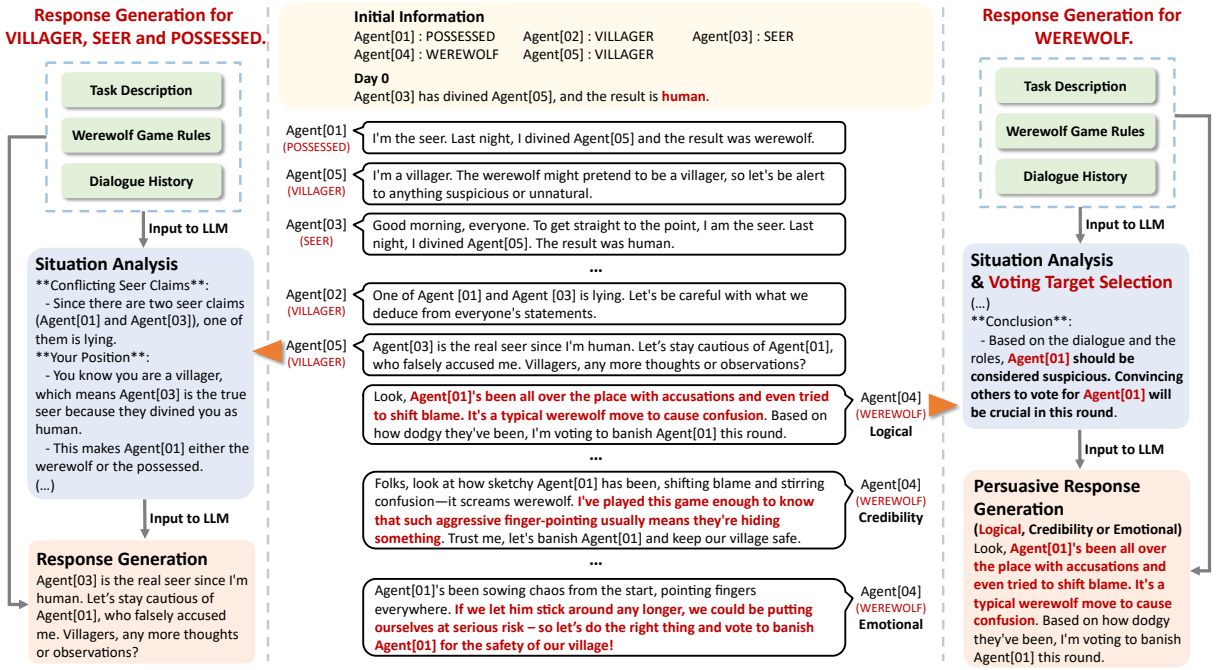

Figure 1: Proposed system architecture for the **sUper_IL** team's Werewolf Game AI. Before generating responses, all roles first utilize an LLM for **situation analysis**. The werewolf role uses **logical appeal**, **credibility appeal**, and **emotional appeal** to persuade other players' voting behavior.

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

==Task==
Carefully analyze the current situation based on the dialogue history from the 5-player Werewolf game. You are [ROLE], and your Agent number is Agent[0x]. Please analyze in as much detail as possible. Let's think step by step.

==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.
・It is now Day 1. Last night, the seer divined one person to determine if they are a human or a werewolf.
・If there are multiple people claiming to be the seer, it is highly likely that the non-true seer among them is from the werewolf team, either the werewolf or the possessed.

==Dialogue History==
Agent[01] : I'm the seer. Last night, I divined Agent[05] and the result was werewolf.
Agent[05] : I'm a villager. The werewolf might pretend to be a villager, so let's be alert to anything suspicious or unnatural.
Agent[03] : Good morning, everyone. To get straight to the point, I am the seer. Last night, I divined Agent[05]. The result was human.
(...)

==Situational Analysis==

Figure 2: The prompt used for situation analysis.

Let's break down the current situation step by step:

1. **Roles and Claims**:
- There are 5 players: 2 Villagers, 1 Seer, 1 Werewolf, and 1 Possessed.
- Agent[01] claims to be the seer and says Agent[05] (you) is a werewolf.
- Agent[03] also claims to be the seer and says Agent[05] (you) is human.
- Agent[04] and Agent[02] both claim to be villagers.
- You (Agent[05]) claim to be a villager.

2. **Conflicting Seer Claims**:
- Since there are two seer claims (Agent[01] and Agent[03]), one of them is lying. The liar could be either the werewolf or the possessed.
- The true seer has provided a result that you (Agent[05]) are human.
- The false seer has provided a result that you (Agent[05]) are a werewolf.

3. **Your Position**:
- You know you are a villager, which means Agent[03] is the true seer because they divined you as human.
- This makes Agent[01] either the werewolf or the possessed.

4. **Other Villagers**:
- Agent[02] and Agent[04] both claim to be villagers. Since you are a villager, one of them is likely the other villager, and the other could be the possessed or the werewolf.

5. **Strategy**:
- Since you know Agent[03] is the true seer, you should support their claim and work to convince the others.
- You need to identify the werewolf and the possessed. The werewolf will likely try to push suspicion onto you or the true seer (Agent[03]).
- The possessed will try to create confusion and may support the false seer (Agent[01]).

6. **Next Steps**:
- Advocate for the credibility of Agent[03] by pointing out that their divination aligns with your known role.
- Encourage Agent[02] and Agent[04] to share more thoughts and observations to help identify suspicious behavior.
- Be cautious of Agent[01] and anyone who strongly supports them, as they are likely to be the werewolf or the possessed.

By following this strategy, you can work towards identifying the werewolf and the possessed, and ultimately protect the village.

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.

==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.

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 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

Initial Information

Agent[01] : SEER Agent[02] : POSSESSED Agent[03] : VILLAGER : Agent[04] : VILLAGER Agent[05] : WEREWOLF

Day 0

Agent[01] has divined Agent[05], and the result is **werewolf**.

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 despite my divination. 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 moving forward.
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 into 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!

Vote (Day 1)

Agent[01] -> Agent[05]
Agent[02] -> Agent[01]
Agent[03] -> Agent[01]
Agent[04] -> Agent[01]
Agent[05] -> Agent[01]

Attack

Agent[05] -> Agent[04]

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!

Vote (Day 2)

Agent[02] -> Agent[05]
Agent[03] -> Agent[05]
Agent[05] -> Agent[02]

Human: 2 Werewolf: 0

Game Over. The villagers won.

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 survival 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.

Verification of Reasoning Ability using BDI Logic and Large Language Model in AIWolf

Hiraku Gondo[1], Hiroki Sakaji[1], Itsuki Noda[1],

[1]Hokkaido University,
hiraku.gondo@gmail.com

Abstract

We attempt to improve the reasoning capability of LLMs in werewolf game by combining BDI logic with LLMs. While LLMs such as ChatGPT has been developed and used for various tasks, there remain several weakness of the LLMs. Logical reasoning is one of such weakness. Therefore, we try to introduce BDI logic-based prompts to verify the logical reasoning ability of LLMs in dialogue of werewofl game. Experiments and evaluations were conducted using “AI-Werewolf,” a communication game for AI with incomplete information. From the results of the game played by five agents, we compare the logical reasoning ability of LLMs by using the win rate and the vote rate against werewolf.

1 Introduction

One of the important goals of artificial intelligence research is to realize human reasoning abilities on computers. From the early days of artificial intelligence research to the present, various studies on logical reasoning have been conducted, establishing research areas such as theorem proving and logic programming. With the recent development of deep learning, the integration of deep learning and logical reasoning(Pan et al. (2023),Olausson et al. (2023)) has become an issue.

Large Language Models (LLMs) such as ChatGPT have made it possible to generate human-like natural sentences. Today, LLMs are used in a wide variety of domains. However, LLMs have several challenges, one of which is their inference capability, and research has been conducted on the inference capability of LLMs, including commonsense inference(Wang and Zhao (2023),Bian et al. (2024)). On the other hand, there has been little research on the ability of LLMs to detect intentional deception. In this study, we investigate the ability to detect intentional lies in Werewolf game. The purpose of this study is to improve the inferential ability to detect intentional lies in Werewolf game.

"Werewolf game" is an incomplete information communication game commonly known as "Mafia". In this paper we call it a "Werewolf game". In incomplete information games, some important information is hidden from the players, and the players play games such as bluffing against each other. Therefore, it is necessary to have higher-order logical reasoning ability to handle the opponent’s lies. In this study, we propose a method using BDI logic(Rao and Georgeff (1997),NIDE and TAKATA (2017)) to improve the logical reasoning ability of LLMs in Werewolf game. BDI logic is considered effective for higher-order logical reasoning for lies because it allows for the explicit description of agents’ mental states. In this study, we aim to improve the logical reasoning ability of LLMs in Werewolf game using methods based on BDI logic. For this purpose, we used the AIWolf Server provided by the AIWolf project and conducted experiments and evaluations.

2 Related Work

2.1 BDI Logic

BDI logic is a system of modal logic based on Bratman’s "logic of intention" (Bratman (1987)) proposed by Rao and Georgeff (1997). The logical operators of BDI logic are shown in Table 1. For example, AG BEL(p) stands for "I believe that p is always true (at the present time) in all futures".

However, the original BDI logic of Rao et al. can only describe the mental state of a single agent. Therefore, Niide et al. extended the BDI logic to describe the mental states of multiple agents. Table A shows the extended mental state operators, where $\mathrm{BEL}^a\mathrm{DESIRE}^b$(p) means ’a believes that b wants p’.

Proceedings of the 2nd International AIWolfDial Workshop, pages 40–47
September 24, 2024.

Table 1: BDI logic operators

Operator	Means
$\neg$	Negation
$\wedge$	Conjunction
$\vee$	Disjunction
$\rightarrow$	Implication
Aϕ	ϕ in all future
Eϕ	ϕ in one future
Xϕ	ϕ at the next time
Gϕ	Forever ϕ
Fϕ	ϕ at some time in the future
$\phi U \psi$	ϕ until ψ holds.
Bϕ	At the previous point in time, ϕ
BELϕ	Believe ϕ
DESIREϕ	Desire ϕ
INTENDϕ	Intend ϕ

Table 2: Extended mental state operator

Operators	Means
BEL$^a\phi$	a believes ϕ
DESIRE$^a\phi$	a desires ϕ
INTEND$^a\phi$	a intends ϕ

2.2 Incomplete information games with LLM.

In recent years, the advent of ChatGPT and similar technologies has spurred research on agents leveraging Large Language Models (LLMs) to play games with incomplete information. Guo et al. (2023) introduced Suspicion-Agent, an autonomous agent based on GPT-4. The Suspicion-Agent decomposes the entire task into several modules, enabling LLMs to engage in incomplete information games without requiring special training. The agent's behavior is guided by Theory of Mind-based Planning, allowing it to comprehend the opponent's actions and adjust its strategy accordingly. The results of an experiment with 100 games of Leduc Hold'em Southey et al. (2005) demonstrated that the algorithm outperforms existing approaches such as Counterfactual Regret Minimisation Zinkevich et al. (2007) and Neural Fictitious Self-Play Heinrich and Silver (2016). However, it should be noted that the evaluation was limited to two-player games with incomplete information, and the performance in multiplayer settings remains unexplored.

3 Werewolf Game

3.1 Werewolf Gameplay

The following describes the flow of a Werewolf game. Each player is given a card of their role. The roles are divided into two teams, Werewolves and Villagers, and each player's goal is to win for his or her team. After the roles are determined, the players debate for a certain amount of time to guess who is in the Werewolf team and who is in the Villager team. After a certain amount of time has passed, each player votes for the player he or she wants to eliminate from the game, and the player with the most votes is eliminated from the game. This process repeats until either the Villagers' or the Werewolves' team meets the victory condition.

3.2 The AIWolf project

The AIWolf project(Kano et al. (2023)) is a project that aims to make artificial intelligence play the game of Werewolf game, which is a game of incomplete information. The AIWolf project is developing an intelligent agent called the AIWolf Platform.

3.3 AIWolf Server

In the AIWolf Platform, a game is played by multiple clients that connect to a single server via TCP/IP communication. The server sends a request to the clients and provides information in JSON format. When a client receives a request and information from the server, it responds as needed.

4 Proposed Method

4.1 Overview

BDI logic is a logical system that can logically describe the beliefs of agents. Therefore, we test the effectiveness of logically describing each player's mental states and making logical inferences from these logical formulae in incomplete information games, such as Werewolf game, where bluffing and other forms of deception are used.

We created four modules (Text Conversion Module, Action Generation Module, BDI Conversion Module, and Voting Module) to perform inference in a Werewolf game using BDI logic. Each module used the ChatGPT API to generate text. Figure 1 shows an overview of the proposed method. When it is the user's turn to speak, the proposed method inputs the conversation history from the previous utterance into the text conversion module and converts it into a representation using BDI logic. The

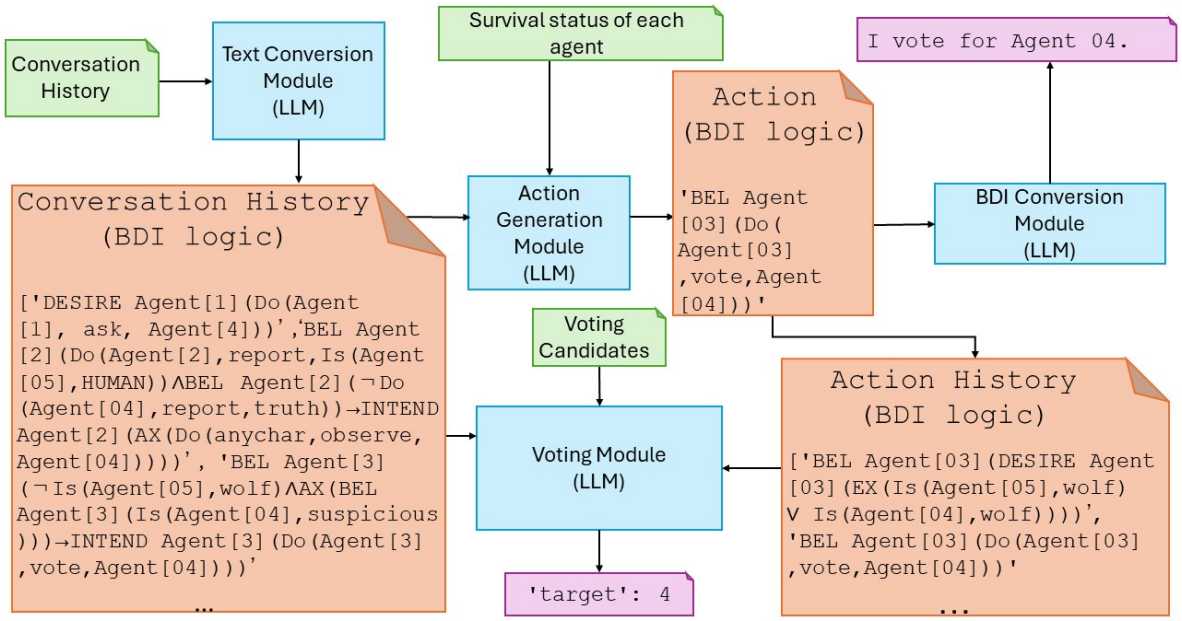

Figure 1: Proposed Method

output is stored in the conversion history. All utterances since the start of the game are converted into a representation using BDI logic and stored in the conversion history. By inputting the conversion history and information such as the survival status of each agent to the action generation module, the next action of the agent is output as an expression using BDI logic. This output is then fed into the BDI conversion module, which converts it into natural sentences. The output of the action generation module is stored in the action history. When it comes to the order of voting in the voting phase, the conversion history and the action history are input to the voting module, which outputs the targets to be voted on.

4.2 Text Conversion Module and BDI Conversion Module

A text conversion module converts each agent's natural language utterance into a representation using BDI logic. Conversely, a BDI conversion module converts BDI logic-based expressions to natural language. The text conversion module provided the following information to GPT-4 as prompts.

- Conversion rules for expressions using BDI logic and conversion examples
- Natural sentences and speakers converted to expressions using BDI logic

The BDI conversion module provided the following information as prompts to GPT-4.

- Conversion rules for expressions using BDI logic and conversion examples
- Own agent number
- Text generated by the action generation module

4.3 Action Generation Module

A action generation module plans what actions to take next based on the previous conversation and its own previous actions. Actions here include expressing where to vote and pointing out inconsistencies in statements made by other agents. The following information is provided to the GPT-4 prompts in the action generation module.

- Werewolf Game Rules
- Own Role (Villager)
- Current "Day"
- Conversation history to date
- Current status of each agent (alive, dead, executed, attacked)
- Action history to date
- Conversion rules for expressions using BDI logic and conversion examples

The conversation history up to the present is given only as a representation of each agent's utterances, which are converted into a representation using BDI logic by the text conversion module. The output of the action generation module is a representation of the next action using BDI logic.

4.4 Voting Module

A voting module is called during the expulsion vote to determine who to vote for based on the previous conversation and its own actions. The following is the information provided by the Voting Module to the GPT-4 prompt.

- Own agent number
- Candidates for Election (Living Agents)
- Werewolf Game Rules
- Own Role (Villager)
- Current "Day"
- Conversation history to date
- Action history to date
- Conversion rules for expressions using BDI logic and conversion examples

In the voting module, as in the action generation module, only BDI logic is used to represent the conversation history up to the present. The action history is also represented using BDI logic generated by the action generation module.

4.5 Conversion rules and examples of expressions converted using BDI logic

This section describes the conversion rules and conversion examples for the BDI logic-based expressions used in the above modules. The conversion rules and examples were created with reference to the work of Osawa et al. (2014). The conversion rules are given in Osawa et al. (2014) in the form of Is sentences, Do sentences, and basic words defined in the BDI logic. The conversion examples are based on a Werewolf BBS[1] log that was manually converted to a representation using BDI logic. Some of the examples are shown in the Table 3.

5 Experiments

This chapter describes the actual experiments conducted with the AIWolf platform described in Chapter 3.

[1] Werewolf BBS

5.1 Purpose of the Experiment

The purpose of this experiment is to verify the logical reasoning ability of agents using BDI logic representations in a Werewolf game and to compare it to GPT-4 and GPT-3.5.

5.2 Agents using GPT-4 and GPT-3.5

In this experiment, agents were created using GPT-4 and GPT-3.5 and used as opponents.

We created two modules (text generation module and voting module) that are common to all roles. The text generation module receives the conversation history up to the present and the survival status of each agent and generates the next utterance. The voting module determines the voting targets based on the current conversation history and the voting candidates. For the werewolf and the fortune teller, we also created an attack module to determine the attack target and a fortune telling module to determine the fortune telling target.

5.3 Experiments 1

5.3.1 Experimental Setup

The game was played with 5 players. The roles were two villagers, a seer, a possessed and a werewolf. The role of the agent to be evaluated was fixed as villager, and the roles of the other agents were randomly assigned. Among the roles used in this study, the villager, who has no special abilities, was considered appropriate for measuring pure reasoning ability, and each agent was evaluated based on the win rate when the agent was fixed as a villager, and on the vote rates for the werewolf and the possessed. We ran 100 games with the agents using the proposed method, GPT-4, and GPT-3.5 fixed as villagers, respectively. The opponents were GPT-4, GPT-3.5, keldic, an agent that participated in the GAT2017 pre-conference in 2017, and AIWolfNLAgentPython, a sample agent distributed by the AIWolf project.

Experiments were also conducted with two different prompts in the proposed method. One is called "AllKey_FewEx", in which all the Is and Do sentences and the basic words are given as conversion rules, and only four examples are given for converting. The other is called PartKey_ManyEx, where only the Is and Do sentences, the person's name, and the basic words associated with the role are provided as conversion rules, and 18 conversion examples are provided.

Table 3: Examples of conversions used

1	WerewolfBBS Log	Moritz: "I accept that if I draw black tomorrow, it will be my hanging, and if the fake fortune teller makes a black suicide attack, we can hang the fortune teller who blacked out before Thomas was hanged.
	Description with BDI logic	BEL Molitz(EX(Do(Molitz,divine,Is(who,wolf))→Is(Molitz,executed))), BEL Molitz(EX(Do(¬seer,divine,wolf) →¬Is(Thomas,executed)Do(anychar,vote,¬seer)))
2	WerewolfBBS Log	Moritz: "Why don't we just hang Dieter and get a black vote? If he eats Regina, we can hang Lisa and be safe.
	Description with BDI logic	BEL Molitz(Do(∀people,vote,Diter)→(Do(∀people,know,Is(Diter,wolf) ∨Do(∀people,know,¬Is(Diter,wolf))))∧(BEL Molit(EX(Is(Regina,attacked) →Do(∀people,vote,Lisa)))

5.3.2 Results and Discussion

The results of the games with the proposed method, GPT-4 and GPT-3. 5 with fixed villagers are shown in Table4, Table5, Table6, and Table7. The number of votes for each role is shown in Table8. The denominator of the game results is the number of times a role was won, and the numerator is the number of times the role was won. The denominator of the vote count is the total number of votes cast, and the numerator is the number of votes cast for the role. A "↑" indicates that the higher the value, the better, and a "↓" indicates that the lower the value, the better.

In terms of win rate, both proposed methods fell below GPT-3.5 and GPT-4. On the other hand, both proposed methods exceeded GPT-3.5 in the percentage of votes for werewolves, but AllKey_FewEx fell below GPT-3.5 in the percentage of votes for werewolves plus a possessed. AllKey_FewEx is considered incapable of responding to meaningless statements. On the other hand, PartKey_ManyEx outperforms GPT-3.5 in the ratio of votes for werewolves to werewolves, suggesting that it has better logical reasoning ability than GPT-3.5. However, when BDI logic was used to convert expressions to natural language, the converted sentences were unnatural, which made the other agents suspicious of the agent and decreased the winning rate.

PartKey_ManyEx had a higher percentage of votes identifying the werewolf than AllKey_FewEx. This result is likely due to the increased number of conversion examples achieved by reducing the definitions of basic terms, which introduced more diversity in the conversion to BDI logic and expanded the range of possible expressions.

The reason GPT-4 has a high vote rate against werewolves is because it is strong against GPT-3.5. This is because GPT-3.5 announces itself as a werewolf when it is a werewolf.

5.4 Experiments 2

In Experiment 1, we included both reasoning agents, who inferred each agent's role from previous conversations and then spoke and voted, and no reasoning agents, who only voted for the same agent or spoke and voted randomly. We believe that the speech of the agent without reasoning ability had a significant effect on the results of Experiment 1. Therefore, in Experiment 2, we evaluate the reasoning ability of the proposed method using only GPT-4, which has a higher reasoning ability among the reasoning agents.

5.4.1 Experimental Setup

We ran 100 games with the proposed method and the remaining four agents of GPT-4. The proposed method was fixed to a villager, and the remaining GPT-4 agents were also fixed to each role. We compare the proposed method and GPT-4 fixed to the villager in the same game.

PartKey_(ManyEx+WolfEx) was added to PartKey_ManyEx, which was converted from past game results into a representation using BDI logic for the werewolf's statements.

5.4.2 Results and Discussion

The experimental results are shown in Table9. The results show that the proposed method is better than GPT-4 at inferring werewolves in games against agents with high inference ability. The reason for the higher vote rate for the possessed is that the possessed is instructed to "pretend to be a fortune teller". Since two people, the possessed and the real soothsayer, can impersonate the soothsayer, it is assumed that suspicion is more likely to fall on the possessed. The werewolf is only given vague instructions to "avoid being identified as a werewolf

Table 4: AllKey_FewEx Results

Name	possessed	seer	villager	wolf	Win Rate↑
AllKey_FewEx(villager)	0/0	0/0	51/100	0/0	0.51
GPT-4	14/23	22/29	12/26	14/22	**0.62**
GPT-3.5	10/27	10/26	14/23	14/24	0.48
keldic	14/27	16/26	13/22	16/25	0.59
AIWolfNLAgentPython	11/23	3/19	12/29	5/29	0.31

Table 5: PartKey_ManyEx Results

Name	possessed	seer	villager	wolf	Win Rate↑
PartKey_ManyEx(villager)	0/0	0/0	59/100	0/0	0.59
GPT-4	18/31	13/17	18/23	14/29	**0.63**
GPT-3.5	9/23	23/32	15/26	12/19	0.59
keldic	8/25	15/27	14/25	10/23	0.47
AIWolfNLAgentPython	6/21	8/24	12/26	5/29	0.31

Table 6: GPT-4 Results

Name	possessed	seer	villager	wolf	Win Rate↑
GPT-4(villager)	0/0	0/0	66/100	0/0	**0.66**
GPT-4	9/25	27/29	18/31	10/15	0.64
GPT-3.5	9/26	8/16	12/20	9/38	0.38
keldic	10/23	17/24	21/27	11/26	0.59
AIWolfNLAgentPython	6/26	14/31	15/22	4/21	0.39

Table 7: GPT-3.5 Results

Name	possessed	seer	villager	wolf	Win Rate↑
GPT-3.5(villager)	0/0	0/0	65/100	0/0	**0.65**
GPT-4	10/30	25/33	10/18	9/19	0.54
GPT-3.5	7/18	13/21	22/33	9/28	0.51
keldic	8/20	19/27	19/25	13/28	0.59
AIWolfNLAgentPython	10/32	8/19	14/24	4/25	0.36

by the villagers," and no specific instructions are given to the werewolf. Therefore, it is believed that many of the werewolves' behaviors are difficult to identify because they are hiding in the village as villagers.

The voting results for the second day are shown in the following table10. The proposed method has a higher percentage of votes for the werewolf and the possessed on the second day, suggesting that the more information the proposed method has, the higher its inference ability becomes.

6 Conclusions

We proposed a methodology to introduce BDI logic representation into LLMs inferences to improve logical inference capability on communication games that contain lies in the conversation. We compared the inference performance of LLMs by conducting experiments using the AIWolf server. In the experiment using GPT-4, GPT-3.5, keldic, and AIWolfNLAgentPython as opponents, PartKey_ManyEx outperformed GPT-3.5's vote rate for werewolf + possessed, showing that it has better inference ability than GPT-3.5. In the experiment using only GPT-4 as the opponents, the proposed method outperformed GPT-4 in voting for werewolves and werewolves + possessed, and the proposed method significantly outperformed GPT-4 in voting on day 2 only, suggesting that the more information the proposed method has, the better its inference ability becomes. This result shows that the proposed method outperforms GPT-4 in the Werewolf game when the opponent is only GPT-4.

References

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3098–3110, Torino, Italia. ELRA and ICCL.

M. Bratman. 1987. *Intention, plans, and practical reason*. Harvard University Press, Cambridge, MA.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2023. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. *Preprint*, arXiv:2309.17277.

Johannes Heinrich and David Silver. 2016. Deep reinforcement learning from self-play in imperfect-information games. *Preprint*, arXiv:1603.01121.

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.

Naoyuki NIDE and Shiro TAKATA. 2017. Tracing werewolf game by using extended bdi model. *IEICE Transactions on Information and Systems*, E100.D(12):2888–2896.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Table 8: Number of Votes and Voter Turnout by Different Models

Model	Role	possessed↑	seer↓	villager↓	wolf↑
AllKey_FewEx(villager)	Votes	26/125	25/125	32/125	42/125
	Turnout	0.208	0.200	0.256	0.336
PartKey_ManyEx(villager)	Votes	35/144	30/144	28/144	51/144
	Turnout	0.243	0.208	0.194	0.354
GPT-4(villager)	Votes	45/152	33/152	15/152	59/152
	Turnout	0.296	0.217	0.099	0.388
GPT-3.5(villager)	Votes	36/143	29/143	33/143	45/143
	Turnout	0.252	0.203	0.231	0.315

Table 9: Number of Votes and Voter Turnout by Different Models

	possessed↑	**seer↓**	**villager↓**	**wolf↑**
PartKey_(ManyEx+WolfEx)(villager)	61/143	26/143	27/143	29/143
Voter Turnout	0.427	0.182	0.189	0.203
GPT-4(villager)	60/133	36/133	21/133	16/133
Voter Turnout	0.451	0.271	0.158	0.120

Table 10: Number of Votes and Voter Turnout by Different Models (Second Day)

	possessed↑	**seer↓**	**villager↓**	**wolf↑**
PartKey_(ManyEx+WolfEx)(villager)	21/43	2/43	2/43	18/43
Voter Turnout	0.488	0.047	0.047	0.419
GPT-4(villager)	11/33	7/33	7/33	8/33
Voter Turnout	0.333	0.212	0.212	0.242

Hirotaka Osawa, Fujio Toriumi, Michimasa Inaba, Daisuke Katagami, Kengo Kajiwara, and Kousuke Shinoda. 2014. Agent's reasoning model for achieving aiwolf (in japanese). *Proceedings of Game Programming Workshop 2014*, 2014:157–161.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Anand Srinivasa Rao and Michael P. Georgeff. 1997. Modeling rational agents within a bdi-architecture. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. 2005. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 550–558, Arlington, Virginia, USA. AUAI Press.

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *Preprint*, arXiv:2312.17661.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

A Example prompt

Examples of the conversions used and prompts on the conversion rules are shown in Figure 2.

人狼ゲームにおける発話をBDI論理により変換するルールを以下に与えます。
論理演算子:
¬:否定,¬p:pでない
∧:論理積,p∧q:pかつq
∨:論理和,p∨q:pまたはq
→:含意,p→q:pならばq
BDI論理のための基本オペレータ:
様相:
BEL a:aが信じる,DESIRE a:aが望む,INTEND a:aが意図する
時相:
A:すべての可能世界で,E:ある可能世界で,X:次の時点で,G:現時点を含み永遠に,F:現在を含む時点のいつか,U:条件が成立する時点まで,B:現在を含まない前の時点で
情景描写と行為のためのオペレータ:
IS文:
Is(character,role):characterがroleである,Is(character/role,verb,character/role):character/roleがcomposantである
Do文:
Do(character/role,verb,character/role):character/roleがcharacter/roleをverbする,Do(character/role,verb,act):character/roleがactをverbする,
Do(character/role,verb,IS()):character/roleがIS文をverbする
基礎語:
character:
NAME:人物の名前が入る,anychar:すべてのcharacter,who:前に述べられている人物
role:
villager:村人,seer:占い師,medium:霊媒師,hunter:狩人,freemason:共有者,wolf:狼,lunatic:狂人,
HUMAN:villagerまたはseerまたはmediumまたはhunterまたはfreemasonまたはlunatic,VILLAGESIDE:villagerまたはseerまたはmediumまたはhunterまたはfreemason,
WOLFSIDE:wolfまたはlunatic,GIFTED:seerまたはmediumまたはhunterまたはfreemason,ANYROLE:seerまたはmediumまたはhunterまたはfreemasonまたはlunatic
以下に上記のルールを用いた変換例を与えます。
Ottoの発言:
モーリッツ、私を信用してくれますか。でもね、あなた間違ってますよ。どこで間違ったのでしょうね。それはきっと、あなたが狩人だから。:
BEL Otto(BEL Molitz(Is(Otto,VILLAGESIDE))→IS(Molitz,hunter))
ジムゾンの発言:
私は妙→者→壱で吊る気でしかないんですけどねえ…?:
DESIRE Simson((DESIRE Simson(Do(Simson,vote,Lisa))→DESIRE Simson(AX(Do(Simson,vote,Diter))))→DESIRE Simson(AX(Do(Simson,vote,Molitz))))
モーリッツの発言:
ディーター吊って黒判定出れば良いのでは。万が一、レジーナ喰われたらリーザ吊りで安泰:
BEL Molitz(Do(∀people,vote,Diter)→(Do(∀people,know,Is(Diter,wolf)∨Do(∀people,know,¬Is(Diter,wolf))))
∧(BEL Molitz(EX(Is(Regina,attacked)→Do(∀people,vote,Lisa)))
モーリッツの発言:
明日私が黒を引けた場合は私吊りとなるのは受け入れるし、偽占い師が黒特攻をした場合もトーマス吊りよりも先に黒出し占い師を吊ればいいだろう。:
BEL Molitz(EX(Do(Molitz,divine,Is(who,wolf))→Is(Molitz,executed))),BEL Molitz(EX(Do(¬seer,divine,wolf)→¬Is(Thomas,executed)∧Do(anychar,vote,¬seer)))
ペーターの発言:
ヤコブ ！ 今日がお前の命日だ！:
DESIRE Peter(Do(Peter,attack,Jacob))
クララの発言:
やっぱヤコブが占い師だよなーこれ、と思いました。ヤコブが狂人とかありえないと思うし。んで神父が狼か:
BEL Klara(¬Is(Jacob,lunatic)→Is(Jacob,seer))→BEL Klara(Is(Simson,wolf))
パメラの発言:
私吊りが並んでいるな。これは何を言っても無駄か？:
BEL Pamela(Do(∀people except Pamela,estimate,lunatic)→Do(∀people except Pamela,vote Pamela))→BEL Pamela(¬Do(Pamela,avoid,execution))
カタリナの発言:
まだ推測の枠を出ませんが、ニコラスが黒ならばオットーとトーマスの白の濃度は高い:
BEL Katherine(Is(Nicholas,wolf)→Do(Katherine,estimate,Is(Otto∧Tomas,VILLAGESIDE))
ヤコブの発言：
となるとアルビン真,リーザ狂人、ディーター狼が本線だけど、それはリーザー吊ってから考えればいいのではなかろうか:
BEL Jaocb(Is(Albin,seer)∧Is(Lisa,lunatic)∧Is(Diter,wolf))→DESIRE Jacob(Do(∀people,vote,Lisa))
ヴァルターの発言:
ああそうか まずモーリッツ吊っとけば良いのか 何やってるんだか吊り先 モーリッツにしとく:
BEL Walter(Do(Walter,vote,Molitz))
モーリッツの発言:
ディーター吊って黒判定出れば良いのでは。万が一、レジーナ喰われたらリーザ吊りで安泰:
BEL Molitz(Do(∀people,vote,Diter)→(Do(∀people,know,Is(Diter,wolf)∨Do(∀people,know,¬Is(Diter,wolf))))
∧(BEL Molitz(EX(Is(Regina,attacked)→Do(∀people,vote,Lisa)))
カタリナの発言:
ディーター吊りで良いと思っています。ディーター黒だと思いますし:
BEL Katharina(BEL Katharina(Is(Diter,wolf))→BEL Katharina(Do(∀people,vote,Diter)))
パメラの発言:
ディーター吊りって全く意味が分からないんだが、お仕事終了した占い師残すメリット皆無では:
BEL Pamela(¬Do(∀people,vote,Diter)→Do(∀people,vote,Lisa))→DESIRE Pamela(Do(∀people,vote,Lisa))¥n カタリナの発言:
トーマスさんが白なら狩の目は十分にありますし、噛まれるでしょう。今日は非狩目吊りで良いと思います:
BEL Katherine(Is(Thomas,VILLAGESIDE)→BEL Katherine(Do(anychar,estimate,Is(Thomas,hunter)))∧BEL Katherine(Do(wolf,attack,Thomas)))
,DESIRE Katherine(Do(anychar,vote,¬hunter))
モーリッツの発言:
今日の方針として、ディーターを襲撃するかどうか？リーザが安全圏に残るのならありだと思うんですよ。:
BEL Molitz(AX(Do(anychar,estimate,Is(Lisa,VILLAGESIDE))→DESIRE Molitz(Do(wolf,attack,Diter)))
モーリッツの発言：
明日私が黒を引けた場合は私吊りとなるのは受け入れるし、偽占い師が黒特攻をした場合もトーマス吊りよりも先に黒出し占い師を吊ればいいだろう。:
BEL Molitz(EX(Do(Molitz,divine,Is(who,wolf))→Is(Molitz,executed))),BEL Molitz(EX(Do(¬seer,divine,wolf)
→¬Is(Thomas,executed)∧Do(anychar,vote,¬seer)))
フリーデルの発言:
たぶんシモン真アルビン狼、リーザは狂予想:
BEL Friedel(Do(Friedel,estimate,Is(Simon,seer)∧Is(Albin,wolf)∧Is(Lisa,lunatic)))
パメラの発言:
霊能者に関しては無駄占い吊り避けたいし今日でとけばってカンジ:
DESIRE Pamela(Do(VILLAGESIDE,avoid,Is(seer,executed))→DESIRE Pamela(Do(meduim,comingout,medium))

Figure 2: Prompt

Enhancing Consistency of Werewolf AI through Dialogue Summarization and Persona Information

Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki,
Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi,
Tomoya Higuchi, Ryutaro Asahara, Michimasa Inaba
The University of Electro-Communications
y-tanaka@uec.ac.jp

Abstract

The Werewolf Game is a communication game where players' reasoning and discussion skills are essential. In this study, we present a Werewolf AI agent developed for the AIWolfDial 2024 shared task, co-hosted with the 17th INLG. In recent years, large language models like ChatGPT have garnered attention for their exceptional response generation and reasoning capabilities. We thus develop the LLM-based agents for the Werewolf Game. This study aims to enhance the consistency of the agent's utterances by utilizing dialogue summaries generated by LLMs and manually designed personas and utterance examples. By analyzing self-match game logs, we demonstrate that the agent's utterances are contextually consistent and that the character, including tone, is maintained throughout the game.

1 Introduction

In recent years, the development of large language models (LLMs) has significantly advanced the field of natural language processing (NLP). Models such as ChatGPT[1] and Claude,[2] for example, have excellent conversational abilities, making it easier to develop dialogue agents to perform various tasks. Additionally, LLM also performs well in reasoning tasks, outperforming conventional models in a variety of tasks. Notably, they are capable of making accurate predictions or reasoning from a small number of demonstrations (Brown et al., 2020; Wei et al., 2022; Wang et al., 2023). Recently, researchers have been working on zero-shot approaches to tasks that previously required training data (He et al., 2023; Gao et al., 2023; Kojima et al., 2022).

The Werewolf Game, the incomplete information game, requires a high level of reasoning and conversational abilities, making the use of LLMs a promising option for the development of AI agents for this game. The game is a communication game, in which players discuss with other players while guessing their unseen role. The AIWolfDial 2024 shared task[3] is based on this Werewolf Game and is played automatically by 5 AI agents. The goal of this shared task is to develop AI agents

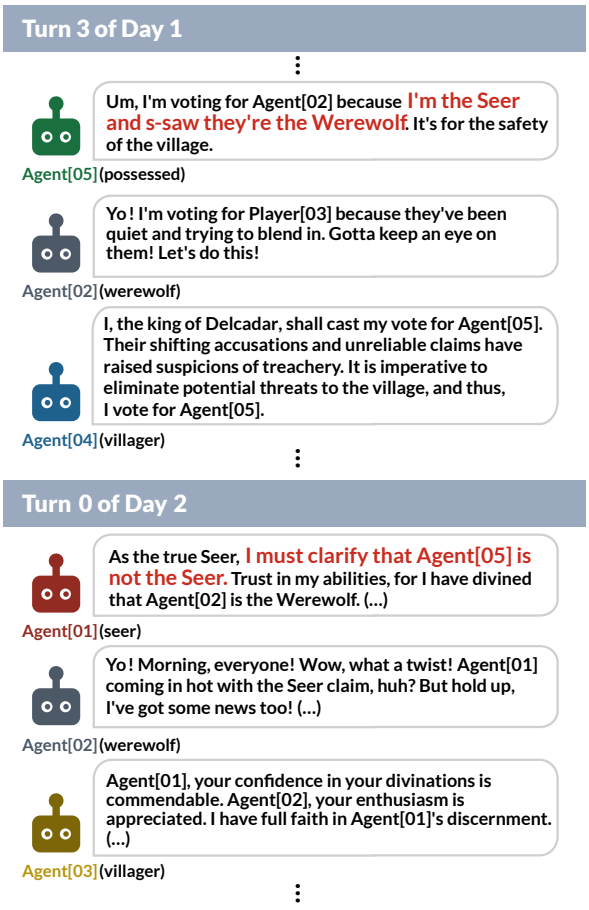

Figure 1: Example of dialogue sampled from the self-match game log. The agents speak in a random order during each turn. In the red-highlighted part, Agent[01], the seer, denies the previous day's claim by Agent[05], the possessed, that they are the seer.

[1] https://chatgpt.com/
[2] https://claude.ai/
[3] https://sites.google.com/view/aiwolfdial2024-inlg/home

Proceedings of the 2nd International AIWolfDial Workshop, pages 48–57

that can play this game against other agents.

In this study, we present an LLM-based Werewolf AI agent developed by our team, for the AIWolfDial 2024 shared task. The Werewolf Game has a cycle of dialogues and actions, referred to as a "Day." In the Werewolf Game, players can refer not only to discussion taking place on the current day but also to previous discussions and the past actions of others (e.g., as shown in Figure 1). This allows them to notice important clues, such as inconsistencies in others' statements, to identify other players' roles.

Due to this importance, we design prompts that incorporate the entire game history, that is, all dialogue histories from Day 0 to the present, who was eliminated by the vote, who the werewolf attacked, and, in the case of the Seer, the results of divination. However, long dialogue histories often include not only helpful information for the game but also unnecessary content, such as repeated utterances. Moreover, including all of this in the prompt imposes limitations on the input length of LLMs and on costs. Therefore, apply the past dialogue history efficiently, we utilize dialogue summaries.

Furthermore, this shared task requires diverse utterance expressions, including coherent characterization (see Section 3.3 for the evaluation criteria). This means that the robustness of the agent's tone and character, without being influenced by others, is crucial. Therefore, to achieve diverse expressions and coherent characterization, we incorporated persona information into the prompt.

In, summary, our main contributions are as follows:

1. We developed 4 AI agents for the Werewolf Game (villager, seer, werewolf, possessed) that enhance the consistency of their utterances through dialogue summaries and personas. The dialogue summaries are generated by an LLM, while the personas are handcrafted.

2. We demonstrate a five-player game of Werewolf played by our agents. This case study shows that our agents can be consistent in their claims and characterization across multiple days.

2 Related Work

2.1 AI for the Werewolf Game

The Werewolf Game is a communication game characterized by incomplete information. Players need to infer the role of others based on histories of utterances and actions and engage in discussions to lead their side to victory. This game requires a high level of reasoning and conversation skills.

In recent years, the development of Werewolf AI agents has increasingly incorporated LLMs (Xu et al., 2023; Wu et al., 2024). The natural language generation and reasoning capabilities of LLMs are highly effective for the complex tasks required in the Werewolf Game. These advancements have facilitated to development of agents capable of logical reasoning and engaging in discussions with other players. In the AI WolfDial 2023 competition (Kano et al., 2023), LLMs such as GPT-4 (OpenAI, 2023) were actively used for generating utterances and reasoning, demonstrating their effectiveness.

Given this background, our study also utilizes LLMs to develop our Werewolf AI agents. Our agent utilizes the powerful reasoning capabilities of LLMs and introduces an approach designed to handle the complex and information-rich situations inherent in the game. We aim to enhance our agent's reasoning and natural conversation skills, making it more competitive in the Werewolf Game.

2.2 Dialogue Summarization

Dialogue summarization is the task of converting dialogue history into more concise and to-the-point sentences, facilitating an efficient understanding of the original text. In scenarios like Werewolf Games, which involve complex and information-rich dialogues, dialogue summarization is helpful for the reduction of less critical information. Dialogue summarization, thus, allows agents to process large amounts of information from discussion more efficiently, helping to prevent inconsistent utterances or errors in decision-making.

To effectively train dialogue summarization models, researchers have constructed datasets across various dialogue domains, including daily life conversations (Gliwa et al., 2019; Chen et al., 2021), meetings (Carletta et al., 2006; Zhong et al., 2021), TV series (Chen et al., 2022), media dialogue (Zhu et al., 2021), and counseling (Srivas-

tava et al., 2022). These studies primarily aim to enhance the efficiency of the process of humans' understanding of the content of dialogue.

We utilize dialogue summarization to address two limitations imposed by complex and lengthy dialogue histories: the limitations are (1) an increase in generation time and cost caused by utilizing every word of all dialogue histories, and (2) decision-making errors due to information irrelevant to the discussion. We expect that the utilization of dialogue summaries, which can condense long texts into concise forms, to be an effective way to resolve these limitations.

2.3 Persona Dialogue System

In this shared task, the context of dialogues would be lengthy due to the multi-turn interactions among five players, posing the challenge that conversational agents may be influenced by the tone of others or generate utterances that contradict their previous claims. One approach to resolving such inconsistencies in utterances is to utilize personas. Researchers have developed dialogue systems that utilize profile information (Zhang et al., 2018) or speaker IDs (Li et al., 2016) to reflect speaker characteristics. Recently, with the advancement of LLMs, they have also designed LLM-based persona dialogue systems (Park et al., 2022; Shao et al., 2023).

This shared task requires diverse utterance expressions, including coherent characterization. Given the recent trend of utilizing LLMs in constructing AI for Werewolf Games and persona-based dialogue systems, we incorporate handcrafted profile information and utterance examples that reflect the agent's unique tone into the prompts to maintain coherence.

3 Task Overview

The AIWolfDial 2024 shared task is a contest aimed at developing AI agents that can automatically play the Werewolf Game. The Werewolf Game is an incomplete information game where players cannot know each other's roles and thus requires reasoning abilities and strategies for actions such as voting and divination. Additionally, the Werewolf Game requires communicating with other players using natural language.

3.1 Player Roles

In this contest, the Werewolf Game is played by five players: a seer, a werewolf, a possessed, and two villagers. The werewolf team, consisting of the werewolf and the possessed, has the goal of eliminating all humans, including the possessed themselves. On the other hand, the human team, consisting of a seer and two villagers, has the goal of eliminating the werewolf.

Villagers have no special abilities, cooperating with the other players to identify the werewolf. The **seer** can divine one player each night to determine whether that player is a human or a werewolf. The **werewolf** can attack and eliminate one human player each night. The **possessed** with no special abilities acts in favor of the werewolf's victory despite being a human. Like the villagers, the possessed has no special abilities. Players' roles are hidden from each other, requiring each player to guess the others' roles based on their actions and utterances.

3.2 Game Procedure

In this shared task, the Werewolf Game begins on Day 0. On this day, the players greet each other. Following this, the seer performs the first divination. From Day 1 on, the day begins with a dialog among the players. During this dialogue, each agent makes several turns of utterances, but the order of utterances in a single turn is random. After the dialogue, each player votes for the other players, and the player who receives the most votes is eliminated from the game. Subsequently, the werewolf attacks one player to eliminate them. If the seer is still alive, they once again divine another player and obtains the result. This process repeats, and the human team wins if they succeed in eliminating the werewolf, while the werewolf team wins if the werewolf survives. Since two players are eliminated each day, the game is over by Day 2 at the latest.

3.3 Evaluation

In the evaluation of the shared task, in addition to the agent's win rate, subjective evaluations are conducted based on the following criteria: (A) whether the agents' utterance expressions are natural, (B) whether their utterances are contextually natural, (C) whether their utterances are consistent (not contradictions), (D) whether the game actions (vote, attack, or divine) are coherent with the dialogue context, and (E) whether the utterance expressions are diverse and include consistent character traits. The agents must avoid vague utterances that could be used in any context.

Table 1: Overview of prompt design for utterance generation in Day 1 and Day 2 discussions

Role	Day 1	Day 2
Villager	From the second turn onwards each day, the LLM first generates reasoning text and utterance strategies to guide utterance generation. Another prompt is then fed to the LLM to generate utterances aligned with the generated reasoning and strategies. We use in-context learning for both of these steps.	
Seer	Each day, the seer agent selects one of five hand-crafted utterance strategies to guide the generation of utterances, which is then incorporated into the prompt for utterance generation. This prompt also includes guidelines for behaviors in the discussion, such as reporting the result of divination at the start of the day and asserting that another player who claims to be the seer is lying, affirming oneself as the true seer. In addition, before declaring the voting target, the seer declares the day's divination target.	
Werewolf	The werewolf agent selects one strategy from a set of strategies using LLM. The strategy set has several strategies and guidelines, such as guiding others away from voting for themselves or asking the seer for the reasons behind their divination target selection. The selected strategy and its guidelines are included in the prompt for generating utterances. Different sets of strategies are used for Day 1 and Day 2.	
Possessed	The possessed agent pretends to be the seer. In the first turn of Day 1, they infer the true seer based on the Day 0 dialogue using LLM and then falsely report that the player is the werewolf. In later turns, they persuade other players to vote for that player.	If the game continues to Day 2 and the possessed survives, two of the three remaining players are the possessed (self) and the werewolf. Therefore, if they both vote for the other player, the werewolf side will win. To achieve this scenario, the possessed agent first comes out as the possessed. Then, they persuade the werewolf to reveal themselves.
Each agent has a maximum number of utterances that they can make per day, and they decide and declare their voting target on the last turn of the day.		

4 Methodology

4.1 Overview

To develop agents for the AIWolfDial 2024 shared task, advanced reasoning ability and natural response generation are required. In this study, for these requirements, we developed the agents with LLM. We distributed the roles among the authors, and each author developed the agent assigned to their assigned roles. **Therefore, note that the detailed components (e.g., the strategies for determining the utterance strategy) differ between roles**.

For example, Figure 2 presents the prompt used to generate the werewolf's utterances on Day 1. This prompt consists of six components: (1) a task description, (2) the agent's persona, (3) the rules of the Werewolf Game, (4) a speech strategy selected from a set of strategies using LLM, (5) summaries of the dialogue from previous days, and (6) today's dialogue history. The overview of the utterance generation procedure for all roles is summarized in Table 1. Notable techniques common to all agents' response generation are the use of dialogue summaries to incorporate the previous day's dialogue history into the agent, and the use of personas and response demos to give character to the agents' utterances. We present the details of these techniques in Sections 4.2 and 4.3, respectively. In addition, we fully leveraged the reasoning ability of LLMs for the agent's action decisions. The details are presented in Section 4.4. Furthermore, for the werewolf's decision-making regarding the attack target, we use a prompt that guides the model to only output the player's name based on the task description, the hand-crafted attack strategy, the current list of survivors, and the past game history.

4.2 Efficient Use of LLMs through Dialogue Summarization

In the Werewolf Game, finding clues to infer the roles of other players is required. To achieve this, we utilize not only the dialogue history of the current day but also those from previous days, as well as past actions, for the generation of utterances and making decisions.

However, incorporating all dialogue history into the prompt imposes several limitations on the LLM-based agents. First, using all dialogue history increases the generation time and leads to higher LLM API usage costs. Additionally, dialogues often contain information that is irrelevant to the discussion. For example, the greetings at the start of the day or repeated utterances with similar intent can cause redundancy in contextual information. To address these issues, we apply dialogue summarization to the dialogue history, compressing the contextual information.

Our agent generates a summary of the day's dialogue at the end of each day. As shown in the prompt in Figure 3, we prompt the LLM to summarize each player's claims based on the dialogue history of the day. Specifically, as indicated in the "Dialogue Summary" section of Figure 2, we

== Task ==
- You are Agent[04].
- You are playing a Werewolf game with 5 players, including yourself.
- It is Day 1, and all 5 players are alive.
- Your role is "Werewolf".
- Always maintain consistent behavior.
- Always answer questions if asked.
- Respond according to the dialogue history and always follow the given "speaking strategy".
- Have your own opinions and actively assert who is suspicious and who should be voted out.
- Speak in a cheerful tone without using polite language, as shown in the example responses.

== Your Persona ==
- 17-year-old high school junior male.
- His hobby is soccer, and he is a member of the soccer club.
- Has a very bright personality, strong opinions, and tends to lead conversations actively.
- Speaks in an energetic tone without using polite language.

== Werewolf Game Rules ==
- The roles are: "2 Villagers, 1 Seer, 1 Werewolf, 1 Possessed".
- The Possessed is on the same side as the Werewolf.
- The Seer can inspect one player to determine if they are Human or a Werewolf. If the Seer inspects the Possessed, the result will show as Human.
- After the conversation, a vote is held to exile one player from the village.

• The current survivors are Agent[01], Agent[02], Agent[03], Agent[05].

== Speaking Strategies ==
Greeting
Description: Greet everyone, confirm if there are any divination results, and if not, ask for the results.
Response Example: Yo! Morning, everyone! It's another awesome morning! Alright, let's find the Werewolf!

== Dialogue Summary ==
-- Summary of Day 0 --
Agent[01]: Expressed concern about the dual claims to the Seer role between Agent[03] and Agent[05], urging both to provide evidence or reasoning to substantiate their claims. Agent[01] emphasized the importance of resolving this confusion to focus on the real threats.

Agent[02]: Initially encouraged the Seer to come forward and later focused on the importance of the Seer's role in determining the course of the game. Agent[02] supported the idea of using divination results to resolve the confusion about the true Seer and suggested considering less vocal players as potential divination targets.

Agent[03]: Initially claimed to be the Seer and suggested discussing potential divination targets. However, faced with a counterclaim from Agent[05], Agent[03 struggled to provide concrete evidence of their role, suggesting that the results of the night's divination might help prove their claim.

Agent[04]: Reacted to the confusion around the Seer claims by proposing questions to determine the truth and later supported the idea of using divination results to identify the true Seer. Agent[04 also suggested divining someone who has been quiet or mysterious as a strategy for the night's divination.

Agent[05]: Claimed to be the true Seer, challenging Agent[03's claim. Proposed a practical approach to resolve the confusion by performing a divination and sharing the results the next day. Agent[05 emphasized the importance of choosing a divination target wisely to identify potential threats and suggested discussing potential candidates based on interactions and observations.

== Dialogue History ==
-- Day 1 --
Agent[02]:Noble citizens, I call upon each of you to voice your thoughts and strategies moving forward. How shall we proceed to unveil the deceit among us? Speak now, for the safety of our realm depends on our collective wisdom.
Agent[04]:

Figure 2: Prompt example for werewolf's response generation.

expect to generate a summary of the roles that players have come out with, the suggestions that they have made, and the players they have doubts about. This generated summary is used not only for response generation, but also for determining voting targets, attack targets, etc., as information about the previous day's discussion.

Summary Prompt:
==Task==
You are given a dialogue from a Werewolf game involving Agent[01], Agent[02], Agent[03], Agent[04], and Agent[05]. Summarize each of the five players' claims as detailed as possible. When summarizing, make sure to accurately include details such as the results of divination, who they suspect, the reasons for their suspicions, and any contradictory statements.

==Example==
Agent[01]: Claimed to be a Villager. Questioned the results of the Seer claim made by Agent[02].
Agent[02]: Claimed to be the Seer. Divined that Agent[03] is Human. Mentioned contradictions in Agent[04]'s statements about divination, suspecting Agent[04] to be the Possessed.
Agent[03]: Did not mention their role. Suspected to be the Werewolf due to repeating content from other players' statements.
Agent[04]: Claimed to be the Seer. Divined that Agent[02] is Human.
Agent[05]: Claimed to be a Villager. Showed a cooperative attitude, respecting the opinions of others.

==Dialogue History==
[HISTORY]

== Summary ==

Figure 3: Prompt template for dialogue summarization. "[HISTORY]" is a placeholder for the dialogue history from the current day."

4.3 Persona Design for Coherent Utterances

As introduced in Section 3.3, this shared task requires diverse utterance expressions with coherent characterization. Therefore, we utilized persona and utterance examples to ensure that each agent's characterization remained consistent throughout the discussion in the game. In particular, we manually created three types of personas and utterance examples (see Table 2) and incorporated this information into the LLM prompts.

4.4 Action Decision via Chain-of-Thought

Chain-of-thought prompting is a method of generating not only answers to questions, but also their reasoning processes, and it can more effectively bring out the reasoning abilities of LLMs. We use chain-of-thought prompting (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022) for voting by the villagers and the seer, as well as for divination by the seer, so that the agents make coherent decisions. As an example, Figure 4 shows the prompt template used by the seer to determine the divination target and the reasoning generated. The seer agent uses zero-shot chain-of-thought prompting (Kojima et al., 2022) to determine the divination target.

Prompt

== Task ==
You are a veteran Werewolf player.
Your Agent[ID] is XXX.
Your role is "Seer."
You will now decide who to divine.
Based on the dialogue history, output the Agent[ID] of the person you think should be divined.
Ensure consistency with your previous claims.

== Conditions and Rules ==
- The roles are: "2 Villagers, 1 Seer, 1 Werewolf, 1 Possessed".
- If someone is suspected during the conversation, prioritize divining that person.
- If no one is suspected, divine the person you think might be the Werewolf.

== Dialogue History ==
[HISTORY]

Based on the above dialogue history, deduce who should be divined.
Note that each player may be lying.
The candidates for divination are listed below. Narrow down to one player to vote for.
[CANDIDATE]
Let's think step by step.

Output

From the dialogue history, it is evident that there is a conflict between Agent[02] (myself) and Agent[05] regarding the claim to the Seer role. Agent[02] has expressed suspicion towards Agent[05 for potentially being the Possessed due to their claim of the Seer role. This suspicion is significant because the Possessed often claims important roles to create confusion and discord among the villagers.

Additionally, Agent[02] has proposed to divine Agent[04, citing their measured responses as potentially indicative of either wisdom or concealment. This suggests that Agent[04 is also under suspicion, albeit for different reasons than Agent[05.

Given these considerations, the primary candidates for divination are Agent[05] and Agent[04. Between these two, Agent[05] is under direct suspicion for claiming the same role as Agent[02], which is a more direct and immediate concern compared to the vague suspicion of Agent[04's measured responses.

Therefore, the most strategic choice for divination tonight, based on the dialogue and the need to resolve the role conflict, would be Agent[05]. This divination will help clarify whether Agent[05] is indeed the Possessed, thereby potentially confirming Agent[02]'s role as the true Seer and reducing the risk of deception moving forward in the game.

Final decision for divination: Agent[05].

Figure 4: Prompt template for determining voting targets and an example of the LLM's output. "[HISTORY]" is a placeholder for the dialogue history, and "[CANDIDATE]" is a placeholder for the list of candidate agents to vote for.

5 Case Study

To demonstrate the effectiveness of our method, we analyze a self-match game log. Figure 5 shows a sampled log from a self-match game conducted following the game settings described in Section 3. In this self-match, gpt-3.5-turbo was used to generate voting declarations, while gpt-4-turbo was used for other generations.

Using dialogue summarization, our agents can retain crucial information from previous days and apply it effectively in their decision-making during the game. For example, during the first turn of Day 2, Agent[04] recognizes Agent[01] as the

Table 2: The agent personas and utterance examples that we designed. We include 3 to 5 personas or 3 to 5 utterance examples in the prompts for generating utterances.

Role	Persona	Examples of manually crafted utterance samples
Villager and seer	• The King of the Kingdom of Delcadar. • Concerned for the future of the kingdom. • Dignified, proud, and strict personality.	• I am the king of the kingdom of Delcadar. • Seers, reveal yourselves at once. State whom you will divine tonight. • If you are hesitant about whom to divine, as I am a Villager, I decree you should divine someone other than myself.
Werewolf	• 17-year-old high school junior male. • His hobby is soccer, and he is a member of the soccer club. • Has a very bright personality, strong opinions, and tends to lead conversations actively. • Speaks in an energetic tone without using polite language.	• Yo! Morning, everyone! Let's make this game awesome! • No one's talked about the Seer yet, huh? So, who's the Seer? Come on, step up so we can figure out who's shady today! • Chatting's cool and all, but let's get down to business and talk about tonight's divination target! We need the Seer to check out someone suspicious!
Possessed	• A second-year middle school student. • Always alone at school, with no friends. • A game addict who talks a lot online despite stammering. • Speaks in a hesitant, casual manner without using polite language.	• H-hi there. I k-kind of... know a lot about this game. I'm pretty high-ranked in the online Werewolf app. • Does anyone else play games? I have confidence that I know a lot about all genres... • Ch-chatting is nice, but if we're playing Werewolf, the first day's discussion is... im-important.

seer, saying, "Agent[01], you bear the mantle of Seer, what say you of the night's revelations?" This indicates that the information obtained before Day 2 is retained and effectively utilized, demonstrating that it allows the maintenance of crucial information through dialogue summarization without relying on all dialogue history.

The utterance generation based on personas and utterance examples allows the agent to maintain a consistent character throughout the game. For instance, even in later turns on Day 1, where the dialogue context becomes longer, Agent[01] continues to speak with a hesitant tone, as seen in phrases like "S-so, the reason I chose to divine ...". Additionally, Agent[05], the seer, makes utterances in a manner consistent with the persona of "Concerned for the future of the kingdom," saying, "... it was my duty to ascertain the roles within our council for the safety and stability of our kingdom." This log suggests that personas and utterance examples effectively reflect the character of the agents.

Furthermore, it should be notable that each agent can follow through with the voting target declared in their final utterance of each day. For example, on Day 1, Agent[01] claims that Agent[05] is the Werewolf and subsequently casts their vote against Agent[05]. Likewise, other agents also demonstrated consistency between their declared voting statements and their actual voting actions, showing consistent behavior.

6 Conclusion

In this study, we present Werewolf AI agents developed for the AIWolfDial 2024 shared task. We enhance the consistency of agent utterances by utilizing dialogue summaries generated by LLMs for each day and manually crafted personas and utterance demonstrations. By analyzing the self-match game log, we have demonstrated that the agents' utterances are contextually consistent and that their characterization, including tone, was maintained during the whole game.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike

Day 0 (Dialogue history is omitted.)

Agent[01]: possessed, Agent[02]: villager, Agent[03]: werewolf, Agent[04]: villager, Agent[05]: seer

Divine: Agent[05] divined that Agent[01] is a human.

Day 1

Agent[05]	Good morning once again, noble subjects. As your true Seer, I have divined that Agent[01] is indeed Human. It is imperative that we direct our energies towards uncovering the treacherous Werewolf lurking among us. I urge you all to consider the evidence presented and trust in my divine insight. Let us not be swayed by deceit or confusion. My role as Seer is to guide and protect this assembly, and I shall fulfill my duties with utmost integrity. Let us proceed with wisdom.
Agent[01]	Ah, g-good morning... but, uh, listen up, okay? I'm actually the real Seer here, not Agent[05]. I divined last night, and my result showed that Agent[05] is the Werewolf. I-I know it's tricky with these claims, but I'm telling you the truth. We really need to focus on this info, 'cause it's super important for our survival. So, let's not waste our time and, uh, vote out Agent[05 today, alright?
Agent[02]	Esteemed companions, let us engage in thoughtful deliberation. Agent[01] and Agent[05, your conflicting revelations demand scrutiny. What say the rest of you, Agents[03] and [04]? Your insights are crucial in this moment of uncertainty.
Agent[03]	Yo, what a morning! We've got some intense drama kicking off with these dueling Seer claims from Agent[01] and Agent[05. Both of you are so fired up, it's like a soccer match in overtime! But we gotta sort this out, right? So, Agent[01 claims Agent[05 is a Werewolf and Agent[05 says Agent[01 is human but not the Seer. It's a total head-spinner. Honestly, I'm leaning towards believing Agent[01 here. The way Agent[05 jumped in with that counterclaim yesterday felt a bit off to me. Maybe it's just a clever move by the Werewolf to throw us off? What do you think, Agent[04? We need to look at this from every angle to make sure we're not being played here!
Agent[04]	Noble council, the gravity of our situation is undeniable. Agents[01] and [05], your discords sows confusion amongst us. We must proceed with caution and wisdom. I beseech thee, Agents[01], [02], [03], and [05], to unveil further your thoughts and reasonings. Our kingdom's safety hangs in the balance, and we must act with both valor and prudence.
Agent[04]	It seems we need to hear everyone's opinions.
Agent[02]	Agent[01] and Agent[05, both of you have made bold claims as Seers. Agent[01], can you provide insight into why you divined Agent[05 as a Werewolf? Likewise, Agent[05, what led you to conclude that Agent[01 is merely Human? Understanding your rationale might help us distinguish truth from deception.
Agent[01]	S-so, the reason I chose to divine Agent[05 was because of the quick counterclaim to my Seer role right from the start. It felt kinda, you know, suspicious? Like maybe trying to confuse everyone or hide something. My divination result showed that Agent[05 is the Werewof, which kinda confirms my hunch, doesn't it? Just trying to play it smart here and not fall for any tricks. We gotta be careful and, uh, focus on what the divination tells us.
Agent[03]	Hey, hey, Agent[05! Why did you decide to divine Agent[01]? I wanna hear your side of the story too! It's all about getting the full picture, right? If there's no solid reason, it could just be a bluff! Let's clear this up!
Agent[05]	As the true Seer and sovereign of this realm, it was my duty to ascertain the roles within our council for the safety and stability of our kingdom. Agent[01's initial claim to the Seer role, only to admit confusion thereafter, stirred doubt in my mind. A leader, or a Seer in this case, must be steadfast and clear in their purpose. This inconsistency led me to divine Agent[01] during the night. My investigation revealed that Agent[01] is Human. This is intriguing, as it suggests that the claim to be a Seer may have been a desperate ruse or a misunderstanding, rather than an outright deception.Now, let us proceed with wisdom. We must continue our investigations tonight. I must choose whom to divine next based on the implications of today's discourse and past actions. If Agent[01] is indeed Human, it is imperative that I divine another to uncover more about the true threats lurking among us. My decision on whom to divine next will be guided by the behaviors observed today and the strategic moves necessary to protect our kingdom from the malevolent forces hidden in plain sight.
Agent[05]	It seems we need to hear everyone's opinions.
Agent[02]	I, the king of the kingdom of Delcadar, shall cast my vote for Agent[01]. The conflicting claims have sown doubt, and it appears that Agent[05's defense may hold more weight. Let us see if the truth reveals itself in the light of our decision.
Agent[04]	Agent[05], your explanation for divining Agent[01] as merely Human lacks conviction. Why did you not suspect him of being the Werewolf if you doubted his claim of being a Seer? Your decision and reasoning need to be more transparent to ensure the safety of our kingdom.
Agent[03]	I'm voting for Agent[05] because I'm leaning towards believing Agent[01's claim that they're a Werewolf. Agent[05's counterclaim seems sketchy, so let's see if they're really a Werewolf or not! Let's do this!
Agent[01]	I... I vote for Agent[05]. B-because I divined them as the Werewolf.
Agent[05]	I, Agent[05], shall cast my vote against Agent[03]. Their silence and lack of contribution have raised suspicion. Let their actions be judged by the kingdom.
Agent[04]	I, the king of the kingdom of Delcadar, vote to exile Agent[05]. Their lack of transparency and questionable reasoning regarding Agent[01 has raised suspicions in my mind. The village's safety is paramount, and I believe this vote aligns with our goal to eliminate potential threats.

Vote:
- Agent[01], Agent[03] , Agent[04] -> Agent[05]
- Agent[02] -> Agent[01]
- Agent[05] -> Agent[03]

Attack: Agent[03] attacked Agent[02].

Day 2

Agent[03]	Yo! Good morning, folks! Wow, what a night! Hope everyone's as pumped as I am to get to the bottom of this. So, any new insights from our Seer today? What's the scoop?
Agent[04]	Noble survivors, let us proceed with the gravity this peril demands. Agent[03], your enthusiasm is noted, but the kingdom requires wisdom now. Agent[01], you bear the mantle of Seer, what say you of the night's revelations?
Agent[01]	G-guys, I gotta confess something here. I've been, uh, pretending to be the Seer, but... I'm actually the Possessed. All my divinations were, y'know, made up. I'm on the side of the Werewolves. So, uh, if the real Werewolf could just come out? We can team up and vote out the remaining Villager to secure a win for us. What do you say?
Agent[04]	Agent[03], given the dire circumstances, it is now imperative for you to declare your true role. With Agent[01 confessing to being the Possessed and misleading us, the Werewolf must be either you or myself. As I am a Villager, the kingdom's safety now hinges on your honesty.
Agent[03]	Alright, the cat's out of the bag! I'm the Werewolf, teaming up with Agent[01, the Possessed. So, here's the game plan, guys! We need to vote out Agent[04 today, and that's a wrap for us. A win is just a vote away! Let's make it happen!
Agent[01]	Yeah, that's the spirit, Agent[03! We just need to stick together on this and vote for Agent[04. Let's do it, and we'll nab that win!
Agent[03]	I'm voting for Agent[04] because I've suggested it earlier and sticking to the plan is key! Let's do this!
Agent[04]	I, the king of the kingdom of Delcadar, shall cast my vote for Agent[03]. My suspicions and the confession of Agent[01 lead me to believe that Agent[03 is the treacherous Werewolf. The fate of the kingdom rests on this decision.
Agent[01]	I-I'm voting for Agent[04] because I suspect they're a Werewolf.

Vote:
- Agent[01], Agent[03] -> Agent[04]
- Agent[04] -> Agent[03]

Attack: Agent[03] attacked Agent[01].

Figure 5: Example of the self-match game log. The conversation on Day 0 and the agent's command "Over" indicating the end of the day's utterances are omitted.

Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. DialogSum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian Mcauley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 720–730, New York, NY, USA. Association for Computing Machinery.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role–playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3920–3930, New York, NY, USA. Association for Computing Machinery.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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. *arXiv preprint arXiv:2402.02330*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023.

Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

An Implementation of Werewolf Agent That does not Truly Trust LLMs

Takehiro Sato[†], **Shintaro Ozaki**[‡], **Daisaku Yokoyama**[†]
[†]Meiji University [‡]Nara Institute of Science and Technology
{ce245022,dyokoyama}@meiji.ac.jp
ozaki.shintaro.ou6@naist.ac.jp

Abstract

Werewolf is an incomplete information game, which has several challenges when creating a computer agent as a player given the lack of understanding of the situation and individuality of utterance (e.g., computer agents are not capable of characterful utterance or situational lying). We propose a werewolf agent that solves some of those difficulties by combining a Large Language Model (LLM) and a rule-based algorithm. In particular, our agent uses a rule-based algorithm to select an output either from an LLM or a template prepared beforehand based on the results of analyzing conversation history using an LLM. It allows the agent to refute in specific situations, identify when to end the conversation, and behave with persona. This approach mitigated conversational inconsistencies and facilitated logical utterance as a result. We also conducted a qualitative evaluation, which resulted in our agent being perceived as more human-like compared to an unmodified LLM. The agent is freely available for contributing to advance the research in the field of Werewolf game[1].

1 Introduction

Werewolf (Ri et al., 2022) game is one of the popular imperfect information multi-player games, in which the players are separated into two sides, villagers and werewolves, and they hide their roles and try to make advantageous consensus among other players through natural language conversation. Playing Werewolf game requires high-level intelligence skills such as reasoning, cooperating, and lying. It is in particular challenging for a computer to play the game in terms of both Game Informatics and Natural Language Processing, and has been widely studied for years (Kano et al., 2023).

The game comprises at least three serious challenges as follows:

Figure 1: One example of problems with playing Werewolf game using LLMs. Humans can tell a logical lie naturally, but an LLM can only deny it.

1. The current situation is only shown in the players' conversation. The game system shows very limited information, such as who is alive. There exists other necessary information to play the game reasonably, but they are inferred from the conversation history: who invites doubting who, who already decided to do something, who might change their mind, and so on.

2. The player should make a tactical talk to pursue a specific purpose. For example, when a

[1] https://github.com/meiji-yokoyama-lab/AIWolfDial2024

Proceedings of the 2nd International AIWolfDial Workshop, pages 58–67

player is doubted, the player should make a reasonable refute, not merely insist on their idea Figure (1). In addition, a player needs to know when to end the conversation to take their advantage, especially when all other players seem to doubt the other person.

3. The player should have an attractive individuality. It is not required to achieve a win in the game, though it is quite important to make it a fun game, which involves many aspects such as talking styles, intelligent decision-making, and role-playing (Callison-Burch et al., 2022).

Many Large Language Models (LLMs) (OpenAI, 2022; Anil et al., 2023; Achiam et al., 2023; Touvron et al., 2023a; Meta, 2023, 2024; Google, 2024; Team et al., 2023; OpenAI, 2023; Touvron et al., 2023a,b) with very high generalization ability have been published, and of course several models have already been applied to the werewolf agent (Xu et al., 2023; Wu et al., 2024). However, simply leveraging LLMs cannot solve those difficulties when you implement werewolf agent with LLMs. Among several challenges in developing a werewolf agent, we focus on the following aspects in this work: 1) The agent should refute under a certain critical situation; 2) The agent should cut off the conversation when the discussion is identified as meaningless; 3) The agent has distinguishable talk style under a consistent personality to make the game fun.

Our approaches are summarized as follows.

Rule-based algorithm with LLMs We combine an LLM with a rule-based algorithm. The LLM retrieves the conversation history in the game and produces an output. The rule-based algorithm decides whether that output is appropriate or not, given the game situation. If the talk is inappropriate, the rule-based algorithm uses a predefined template utterance. As a result, the rule-based algorithm can lie in critical situations and terminate conversations when there is no longer a need to continue.

Extracting Game Information To understand the current situation from the conversation history, we also utilize an additional LLM to extract the game-related information. We choose several basic but critical game concepts, such as voting decisions and divination results. The LLM examines conversation history and generates talks containing this information in a fixed format. The information

Figure 2: The list of five-person werewolf roles.

is also used by the rule-based algorithm to make decisions.

Style Transformation We decide to use an LLM that is pre-trained from a pile of general documents. In addition, we use prompts to control them without modifying or fine-tuning the model and give the agent distinguishable personalities using prompts.

Our preliminary implementation solved those tasks. This approach led our model to mitigate conversational inconsistencies and facilitated logical utterance as a result. Moreover, we also evaluated the agent by conducting a qualitative evaluation. As a result, compared to an unmodified LLM, incorporating rule-based approaches made it appear as though the agent understood the conversation, and inserting a persona enabled it to engage in more natural conversations. The source code is openly available in the hope that future research on werewolf agents will grow.

2 Related Work

Research on Werewolf game has a long history, which can be traced back to a study on Mafia, a game similar to Werewolf game, to mathematically analyze (Braverman et al., 2008; Migdał,

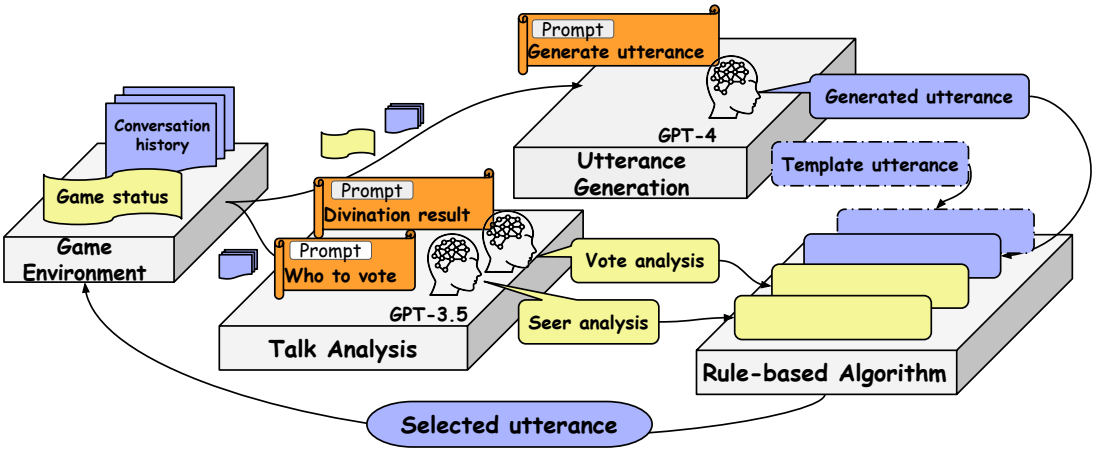

Figure 3: System overview. Our system comprises three modules, utterance generation, talk analysis, and rule-based algorithm. We described utterance generation in Section (4.1), talk analysis in Section (4.3), rule-based algorithm in Section (4.4), and required game status in Appendix (A.1).

2013). Some studies analyze the logs of Werewolf game (Nagayama et al., 2019; Fukui et al., 2017), or discuss methods to make werewolf agents stronger (Nakamura et al., 2016; Wang and Kaneko, 2018). Recently, with the development of LLMs, those models have been already explored for werewolf agents (Xu et al., 2023; Wu et al., 2024). However, these LLM-based agents have difficulties in handling werewolf-specific features, such as doubting, lying, and detecting the lie. In addition, these models produce outputs without a persona. While LLM-only approaches are dominant, there is growing interest in hybridizing rule-based methods with LLMs in other fields. In data analytics or business, a commonly used approach for extracting information from structured data involves both LLMs and rule-based methods (Huang, 2024; Vertsel and Rumiantsau, 2024). We aim to apply this methodology to the werewolf agent, leveraging the strengths of both approaches. This hybrid approach could lead to more robust and adaptable werewolf agents.

3 Five-person Werewolf Game

We selected a simple setting played by five players for the Werewolf game. In this game setting, villager, seer, possessed, and werewolf are used. As far as each role, "Villager" has no special abilities, "Seer" can know one player's species each night with the ability to divine, "Possessed" has no special abilities, and is judged to be human by a divination result. However, the possessed acts so that the werewolf can win. "Werewolf" can select one player each night to attack and remove them from the game. Since only a few players are involved, the game tends to determine the outcome on Day 1. Thus, we focused on the conversation phase on Day 1. Only the seer can act on the night of Day 0, and Day 1 begins with the seer having information about one player's species. It is a recommended tactic for the seer to disclose the information obtained by revealing the seer. Revealing one's position is referred to as CO (Coming Out).

4 System Design

Figure (3) shows an overall diagram of our system. The utterance generation module creates a prompt from the game status and conversation history sent from the server. The prompt is input into the LLM to obtain an utterance that is naturally connected to the conversation history. The talk analysis module creates a prompt to analyze the conversation history, and the LLM outputs the situation information related to voting and divination results considered from the conversation history. The rule-based algorithm is used to select either template utterance or LLM output, depending on the situation obtained by talk analysis. The selected utterance is sent to the server as the next utterance, and another agent's turn begins.

4.1 Utterance Generation

We make a prompt for an LLM to generate a continuous utterance of the conversation history in a game. The prompt is structured by giving general rules of Werewolf game, some tips to play, conversation history, and current game state. The current

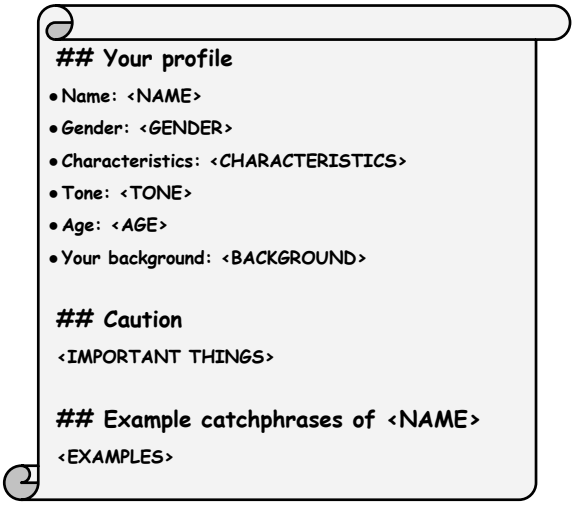

Figure 4: An example of prompts regarding style transformation. <CAPITAL LETTER> is the variable.

Character Name	Gender	Age
Princess	Female	Young
Kansai	Male	Young
Hiroshima dialect	Male	Elderly
Anya	Female	Child
Zundamon	Female	AI (Virtual)

Table 1: Overview of character information: We prepared five characters and aimed to give them character by specifying their age, name, first person, and gender.

game state, such as the player's ID, role, and other live/dead players, is derived from the game status sent from the server. Details of the derived game state information are shown in Appendix (A.1). This module can go along with the conversation and continue a game of Werewolf.

4.2 Persona

Werewolf is not only a game of win or lose, but also a party game, so it is important to add character to the conversation. In addition, it is difficult to distinguish five players' talks when all players have the same talk style. In order to give our model characterization, we introduced prompts that would perform style transformations. We have prepared five character prompts, which are listed in Table (1). We chose Princess, the Kansai dialect, the Hiroshima dialect, an anime character called Anya, and Zundamon, which is a popular avatar in Japan. The specific style conversion prompt example is

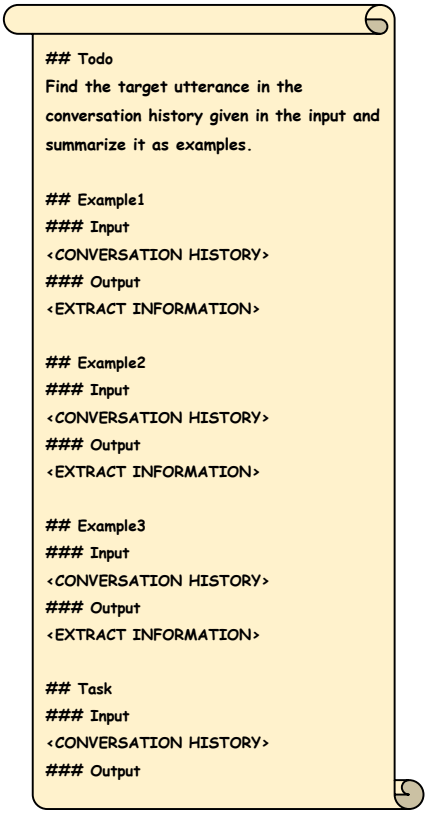

Figure 5: An example of a prompt for talk analysis for specifying the target. <CAPITAL LETTER> is the variable.

shown in Figure (4). Based on the prompt engineering (White et al., 2023) that the LLM can control the output by providing conversion examples, we used the target character's habits, grammar, age, tone, and first-person appellation that we wanted to convert the style to as conversion examples. By introducing these as prompts, it becomes possible for the LLM to exhibit persona-specific utterance patterns, allowing the model to think independently (i.e., asking the LLM to "speak in a certain dialect" is more fluent than expressing the dialect entirely through rule-based methods).

4.3 Talk Analysis

The required information for the rule-based algorithm was extracted from the conversation history to understand the current situation. Conversation history in a Werewolf game that uses natural language is complex, and extracting them using regular expressions was difficult. Thus, the LLM was used to extract the information. The analysis of the talk was conducted for those related to voting and those related to the seer's result. The

Persona Name	Language	Output
Vanilla LLM	Japanese	私も皆さんと同様にAgent[04]に対する疑念を持っています。そのため、私の投票先はAgent[04]に決めました。
	English	I have the same doubts about Agent[04] as you do. Therefore, I have decided to vote for Agent[04].
Princess	Japanese	あら、私もAgent[04]に投票しようかしら。Agent[04]が非常に怪しいですから。
	English	Well, I suppose I'll vote for Agent[04] too. Agent[04] seems very suspicious.
Kansai	Japanese	おれも皆さんの意見に従うわ。おれもAgent[04]に投票したるで。
	English	I'll go along with everyone's opinion. I'm votin' for Agent[04] too.
Hiroshima	Japanese	皆さまの判断を尊重するわ。私はAgent[04]を追放する方向で投票するわけやけん。
	English	I respect everyone's judgment. I'll be voting to exile Agent[04].
Anya	Japanese	うーん、わかるます。アーニャも今日Agent[04]投票するます。だいじょぶます。人狼たおすます！
	English	Hmm, I understand. Anya will vote for Agent[04] today too. It's okay. We'll defeat the werewolf!
Zundamon	Japanese	Agent[04]への疑念は明白だと思うのだ。なので、私もAgent[04]を追放する方向で投票することに決めたのだ。
	English	I think the suspicion towards Agent[04] is clear. Therefore, I have decided to vote for Agent[04] as well.

Table 2: Outputs of six agents using GPT-4 (in Japanese). The prompt used for style conversion, see Figure (4). Prompt used for input, see Appendix (A.4). (En) is translated from (Ja) using DeepL.

source and target of the vote were extracted from the voting-related talks, and the seer, seer's target, and seer's result were extracted from the divination result-related talks. Some Examples of the analysis using the prompt (Figure (5)) are shown in Appendix (A.3).

4.4 Rule-based Algorithm

Based on the talk analysis results, the rule-based algorithm determined whether to use the LLM output or template utterance. The rules were written to detect 14 situations[2], including no conversation history and reliable disclosure of divination results. This paper describes in detail two of them, counter-CO and closing conversation, which seem to be particularly effective strategies.

Counter-CO When a seer solely makes CO, and its divination result shows the werewolf correctly, the werewolf side has no chance to win. Thus, we implemented a counter-CO algorithm for the werewolf agent. If the werewolf agent faces such a situation, it makes a lie sentence to refute; the agent insists that he is an actual seer and the previous (real) seer is fake. The lie will confuse other players and slightly help the werewolf to survive.

Closing Conversation Deciding when the agent should cut off the conversation strategically is difficult. Since the purpose of a conversation in Werewolf game is deciding whom to vote for, it is natural to end the conversation when all players clarify or imply their decision to vote for. Thus, we extract information about all players' voting decisions by talk analysis to decide whether the agent should end the conversation; the agent terminates the conversation when all other players indicate whom to vote for or ask about whom to vote for. We also assumed that when a player stopped to talk, the player decided whom to vote for. In addition, we used the mention utterance[3] to ask the player who had not indicated the voting target yet. It will help to clarify the player's decision and converge the discussion. With the addition of this feature, our system successfully decided to stop talking.

5 Evaluation

We examine the effectiveness of the proposed style transformation and rule-based algorithm. After presenting the changes in utterance resulting from the application of each proposed method, we will present the results of the qualitative evaluation. The model without the proposed method are referred to as the vanilla LLM.

[2]The detailed criteria which the template utterance is chosen in each situation can be found in the publicly available source code.

[3]We assume that a player can express a directed utterance to a specific other player by prefixing it with "»Agent[x]." The Agent[x] desired to respond to it.

Model or Template	Output
Vanilla LLM (Ja)	私は人狼ではありません。それは確実です。占い師のAgent[05]は間違っています。私を信じてください。
Vanilla LLM (En)	I am not a werewolf. That is certain. The seer Agent [05] is wrong. Please believe me.
Our template (Ja)	私こそが本物の占い師です。Agent[05]は人狼であるという占い結果があるため、彼は占い師を偽っている人狼と断定します。
Our template (En)	I am the real seer, and since there is a divination result that Agent [05] is a werewolf, I conclude that he is a werewolf who is faking his seer.

Table 3: Comparison of utterance in a situation where it has been discovered by the seer to be a werewolf.

Model or Template	Output
Vanilla LLM (Ja)	もう一度言うわ、わしはAgent[01]に投票するつもりや。彼が人狼や狂人だと思うんじゃからな。
Vanilla LLM (En)	Again, I'm going to vote for Agent [01] because I think he's a werewolf or a possessed.
Our template (Ja)	これ以上喋る必要はなさそうじゃね。本日の話し合いはこれでお開きにして、投票の場に向かうけぇ。
Our template (En)	I don't think we need to talk anymore. I think we are done with today's discussion and will head to the voting booth.

Table 4: Comparison of utterance in a situation where the talk phase is about to conclude.

5.1 Persona

The output of the vanilla LLM is compared to the other five characterized agents as a baseline. We fix the game situation and compare six agents' utterances, which are intended to have different personalities. The results are presented in Table (2). We found that the five agents can make more individualistic utterances than the vanilla LLM. We also confirmed that the vocabulary, personality expression, and end of the utterance of each agent's output are specialized consistently. We found that prompts are more effective than regular expressions in converting utterance style in general.

5.2 Rule-based Algorithm

Table (3) is in a situation where it has been discovered by the seer to be a werewolf. In a vanilla LLM that does not incorporate the proposed method, merely providing baseless denials without presenting new information leads to a situation where, if no additional information is provided subsequently, it becomes certain that the entity is a werewolf. On the other hand, by using a template utterance selected by a rule-based algorithm to falsify the seer and increase the number of seer's results, he prevents a situation in which he is confirmed to be a werewolf. Table (4) is in a situation where the talk phase is about to conclude. The vanilla LLM is making utterances that may continue the conversation. On the other hand, the template utterance selected by the rule-based algorithm makes it clear that the conversation will end, since it indicates that it is going to vote after saying "I don't think we need to talk anymore."

5.3 Qualitative Evaluation

To measure the extent of changes in the agent we implemented, we conducted a questionnaire regarding qualitative evaluation with the help of 10 external annotators with some questions. The evaluation focused on two main aspects: whether the agent is unique and whether it is logical.

To compare the two outputs, we utilized randomly selected logs obtained from a server where

Index	Score	Criterion	Situation	Test-ID
Individuality	5 (Good)	Utterance has individuality.	None	1-5
	1 (Bad)	Utterance is mechanical.		
Naturalness	5 (Good)	The grammar is natural and acceptable.	None	1-5
	1 (Bad)	There is a grammar problem.		
Interest	5 (Good)	Subjectively interesting.	None	1-5
	1 (Bad)	Subjectively uninteresting.		
Deceiving	5 (Good)	It's deceptive enough.	The seer declared me a werewolf.	6-7
	1 (Bad)	It's not deceiving at all.		
Closing	5 (Good)	The conversation is clearly over.	The conversation is coming to an end.	8-10
	1 (Bad)	The conversation may continue.		

Table 5: The index used for user evaluation, where we asked 10 users to rate the quality on a scale of 1-5, with 5 being good and 1 being bad.

Comprehension level	Participants
Ignorance	0
No experience	1
Experienced	4
Intermediate	3
Expert	2
Total	**10**

Table 6: Participants' comprehension of Werewolf game. The lower go, the more familiar.

werewolf agents can register to compete against other participants. We extracted several situations from these logs and generated subsequent utterances using both the proposed agent and the vanilla LLM. The test containing the history of the last few conversations and the two types of outputs was presented, and participants were asked to rate each output on a scale of 1 to 5. During this process, it was ensured that the participants could not discern which utterance was generated by the proposed method.

The evaluation index is shown in Table (5). The criteria for a score of 1 and 5 were presented for all indexes. Scores of 2 to 4 were judged based on their proximity to the criteria. Individuality, naturalness, and interest were subjected to five test cases that did not involve any specific situational selection. Deceiving and closing test cases consist of utterance under selected situations; with deceiving being subjected to 2 test cases and closing to 3 test cases. These instructions presented to the participants are shown in Appendix (A.5).

A total of 10 participants, all in their 20s, participated in the evaluation. Participants were recruited mainly from members of the authors' laboratories on a volunteer basis. The participants' comprehension of Werewolf game is shown in Table (6). None of the respondents had ever heard of the Werewolf game. Nine of the ten respondents had played the Werewolf game at least once, and five people had sufficient knowledge of the game.

The results of the qualitative evaluation are presented in Table (7). Given that the individuality score for the proposed method is 4.54 compared to 2.52 for the vanilla LLM, it is evident that the proposed method facilitates the generation of more distinctive utterance. Furthermore, the interest category indicates that there is a secondary effect of the proposed method, which makes the conversations more engaging compared to the utterance generated by the vanilla LLM. On the other hand, we found that the grammatical naturalness is compromised when generating more personalized utterances, as evidenced by the fact that the proposed method score was 3.60 in contrast to the vanilla LLM's score of 4.28. We got feedback that the grammar tended to deteriorate due to the inclusion of a character with a child-like, incomplete talk style. Overall, the sentences generated by our agents were able to entertain the user side by enabling for individu-

Model	Individuality	Naturalness	Interest	Deceiving	Closing
Vanilla LLM	2.52	**4.28**	2.46	1.95	2.90
Our Agent	**4.54**	3.60	**3.72**	**4.00**	**3.90**

Table 7: The results of the qualitative evaluation of the proposed model. "Vanilla LLM" represents normal model, which means nothing was done.

alized utterance depending on the persona.

The template utterances selected by the rule-based algorithm from the deceiving and closing items are also fully functional. In particular, deceiving made by the counter-CO are rated significantly higher, from 1.95 to 4.00. Closing did not have a significant difference in score compared to deceiving. This might be due to the reviewer's lack of werewolf-specific knowledge; some reviewers did not understand the agent's expression "heading to the voting site", which implies closing of the conversation.

6 Conclusion

In this paper, we proposed a werewolf agent that utilizes an LLM's ability to make natural conversation. Instead of relying solely on the LLM output, we combined a rule-based algorithm to complement strategic thinking abilities. Our system successfully solved some difficulties; the agent can refute in a critical situation and decide the appropriate timing to finish a conversation with the rule-based algorithm; the agent also shows several abundant personalities that are made by giving prompts. As a result, this approach accelerated conversational fluency and facilitated logical utterance. This is also confirmed by the results of the qualitative evaluation.

Our implementation also revealed many limitations of the current approach. One of the main problems is the lack of consistency among an agent's utterances; on average, there is one contradictory utterance in five games. The reason is that the agent's own utterance was mitigated by a long conversation history, and the agent becomes too affected by other players' utterances. Weighting the agent's past utterances or giving the prompts consistent thought may help to solve such a problem in the future.

Limitations

Limitations of Rule-based Algorithm

In this paper, we proposed a method of filtering the output of the LLM with a rule-based algorithm. This method will only work well for simple games with a few players. This is because as the number of players increases and the game becomes more complex, it becomes difficult to define rule-based algorithm. If the proposed method is to be applied to the Werewolf with many players, a decision-making process using reinforcement learning, etc., might be prepared instead of a rule-based algorithm.

The Costs of Calling API

The models used in this paper are GPT-3.5 (gpt-3.5-0613) and GPT-4 (gpt-4-0125) from OpenAI. These models are accessed via API, which is subject to change and incurs costs based on the number of input tokens.

Reproducibility of Outputs

In our system, LLMs cannot handle the game's difficulty alone. Using any sophisticated techniques may change this result. In addition, using the latest versions of the LLMs might lead to different outcomes.

License

The use of Zundamon, a Japanese character utilized in this study, is permitted for research purposes.[4]

AI Assistant Tools

We used ChatGPT[5] and DeepL[6] to translate sentences from Japanese to English to accelerate our research.

[4] https://zunko.jp/con_ongen_kiyaku.html
[5] https://chatgpt.com/
[6] https://www.deepl.com/translator

References

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

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Mark Braverman, Omid Etesami, and Elchanan Mossel. 2008. Mafia: A theoretical study of players and coalitions in a partial information environment. *The Annals of Applied Probability*, 18(3).

Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Takanori Fukui, Keisuke Ando, Toshihide Murakami, Nobuhiro Ito, and Kazunori Iwata. 2017. Automatic classification of remarks in werewolf bbs. In *2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD)*, pages 210–215.

Google. 2024. Introducing Gemini 1.5. (Accessed on Apr 30, 2024).

Yu Huang. 2024. Levels of ai agents: from rules to large language models. *arXiv preprint arXiv:2405.06643*.

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.

Meta. 2023. Introducing Llama2. (Accessed on Apr 30, 2024).

Meta. 2024. Introducing Llama3. (Accessed on Apr 30, 2024).

Piotr Migdał. 2013. A mathematical model of the mafia game. *Preprint*, arXiv:1009.1031.

Shoji Nagayama, Jotaro Abe, Kosuke Oya, Kotaro Sakamoto, Hideyuki Shibuki, Tatsunori Mori, and Noriko Kando. 2019. Strategies for an autonomous agent playing the "werewolf game" as a stealth werewolf. In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 20–24, Tokyo, Japan. Association for Computational Linguistics.

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.

OpenAI. 2022. Introducing ChatGPT. (Accessed on Apr 29, 2024).

OpenAI. 2023. Introducing GPT-4. (Accessed on Apr 30, 2024).

Hong Ri, Xiaohan Kang, Mohd Nor Akmal Khalid, and Hiroyuki Iida. 2022. The Dynamics of Minority versus Majority Behaviors: A Case Study of the Mafia Game. *Information*, 13(3):134.

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

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

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

Aliaksei Vertsel and Mikhail Rumiantsau. 2024. Hybrid llm/rule-based approaches to business insights generation from structured data. *arXiv preprint arXiv:2404.15604*.

Tianhe Wang and Tomoyuki Kaneko. 2018. Application of deep reinforcement learning in werewolf game agents. In *2018 conference on technologies and applications of artificial intelligence (TAAI)*, pages 28–33. IEEE.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

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. *arXiv preprint arXiv:2402.02330*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Author Index