AIWolfDial 2025

The 3rd International Workshop of AI Werewolf and Dialog System (AIWolfDial 2025)

held in conjunction with
The 18th International Natural Language Generation Conference
(INLG 2025)

Proceedings of the Workshop

October 30, 2025

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 979-8-89176-325-8

Message from the Workshop Chair

We are excited to present the Proceedings of the 3rd Interenational AI Werewolf and Dialog System Workshop (AIWolfDial 2025) held in conjunction with the International 18th International Natural Language Generation Conference (INLG 2025). Our workshop takes place in October 30 in Hanoi, Vietnam.

Recent achievements of large language models, e.g. ChatGPT, are gathering greater attentions. However, it is not fully explored that such a huge language model could sufficiently able to handle coherent responses, longer contexts, common grounds, and logics.

The AIWolfDial 2025 international contest is held as a part of this AIWolfDial 2025 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, 3 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.

We would like to thank the INLG conference organizing team; the workshop would not be possible without their dedication and hard work. We would also like to express our gratitude to our organizers and the program committee members for their reviewing contributions, and the local organizing team for operating the contest and subjective evaluations.

AIWolfDial 2025 workshop chair,
Yoshinobu Kano

Organizing Committee

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

Table of Contents

Conference Program

October 30th, 2025

Peer reviewed papers

AIWolfDial 2025: Summary of Natural Language Division of 7th International AIWolf Contest
Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki and Fujio Toriumi

Towards a Strategic Werewolf AI Based on Expert Strategies in Five-Player Werewolf
Takuma Okada and Takeshi Ito

Intent-driven AIWolf Agents with Hierarchical BDI Model and Personality
Yuya Harada and Yoshinobu Kano

Influence of Utterance Impressions on Decision-Making in LLM-to-LLM Discussions
Neo Watanabe and Yoshinobu Kano

Non-archival system descriptions

CamelliaDragons
Reon Ohashi, Momoka Kato, Yugo Kato, Koki Sato, Joji Suzuki, Shinma Tsuboi, and Kazuya Tsubokura

CanisLupus
Yu Sugawara

Character-Lab
Kun Kerdthaisong, Pasin Buakhaw Peerapat, Pitikorn Khlaisamniang and Supasate Vorathammathorn

mille
Katsuki Ohto

sunamelli
Satoko Natsuori, Koya Kamada, Ryo Kamiyama, and Hiroki Nakanishi

AIWolfDial 2025: Summary of Natural Language Division of 7th International AIWolf Contest

Yoshinobu Kano[1]*, Neo Watanabe[1], Yuya Harada[1], Yuto Sahashi[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 7th 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 large language models. In our Natural Language Division of the contest, we had eight English speaking agent teams for the five-player track, and six English speaking agents for the newly introduced 13-player track, to automatically run games between those agents. By using the game logs, we performed win rates, human subjective evaluations, LLM-as-a-judge automatic subjective evaluation, and detailed log analysis. We found that, in the newly introduced 13-players track, the communications between agents are not fluent and not context-aware than expected from the recent LLMs' performance. This result revealed the current limitations of the use of LLMs, especially when there is a complex relationships required between multiple 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 2025[1], 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. AIWolfDial 2025 is one of the workshops of 18th International Natural Language Generation Conference (INLG 2025). 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 an appropriate evaluation. Because the Werewolf game forces players to deceive, persuade, and detect lies, neither inconsistent nor vague response 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 competitions to automatically play the Werewolf game since 2014 (Toriumi et al., 2017), as the AIWolf

*Correspondence to kano@kanolab.net

[1]Our AIWolfDial official website: https://aiwolfdial.github.io/aiwolf-nlp/en; Our game log viewer site: https://aiwolfdial.github.io/aiwolf-nlp-viewer

Proceedings of the 3rd International AIWolfDial Workshop, pages 1–13
October 30, 2025.

project [2]. 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), AIWolfDial 2023 at INLG 2023 (Kano et al., 2023), and AIWolfDial 2024 at INLG 2024 (Kano et al., 2024). 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, the player agents should communicate in natural languages (English or Japanese). The natural language division is simple, and the 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 paper.

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 win rates, human subjective evaluations, and llm-as-a-judge automatic subjective evaluation, which is newly introduced from this contest.

Eight agents (eight teams) participated in this AIWolfDial 2025 shared task, where eight teams provided English speaking agents for the five-player track, and six teams provided English speaking agents for the 13-players track, which is newly introduced from this contest. This new 13-players track includes two werewolf role players, who can secretly communicate each other. Together with the increased number of players and new roles, this secret communication requires players to collaborate as a team.

In the following sections, we explain the game regulations of the AIWolf natural language division in Section 2, rough system designs for each agent in Section 3, results of evaluations in Section 4.1 followed by discussions in Section 5, finally conclude this paper in Section 6.

[2] http://aiwolf.org/

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 (Table 1). A game in the AIWolfDial 2025 shared task have five players: a seer, a werewolf, a possessed, and two villagers (five-players track), and 13-players: a seer, three werewolves, a possessed, a medium, a bodyguard, and 6 villagers. Werewolves can make *whispers*, i.e. communicate secretly each others in the night.

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.

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 four talks per day per agent, thus 20 mtalks in total per day in AIWolfDial 2025. The maximum string lenght for each talk is 125 letters excluding whitespaces; if the talk text contains any mention ("`@player_name`") to other agents, then the maximum length is doubled to 250 letters, in this AIWolfDial 2025.

Role	Team	Species	Special Abilities
`Villager`	Villager	Human	Nothing
`Seer`	Villager	Human	Divine one survivor to know their species (human or werewolf).
`Medium`	Villager	Human	Divine one eliminated player to know their species (human or werewolf).
`Bodyguard`	Villager	Human	Protect one player from a werewolf attack during the night.
`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: Roles in our Werewolf game

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 a mention symbol (e.g. "@agent_name") 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.

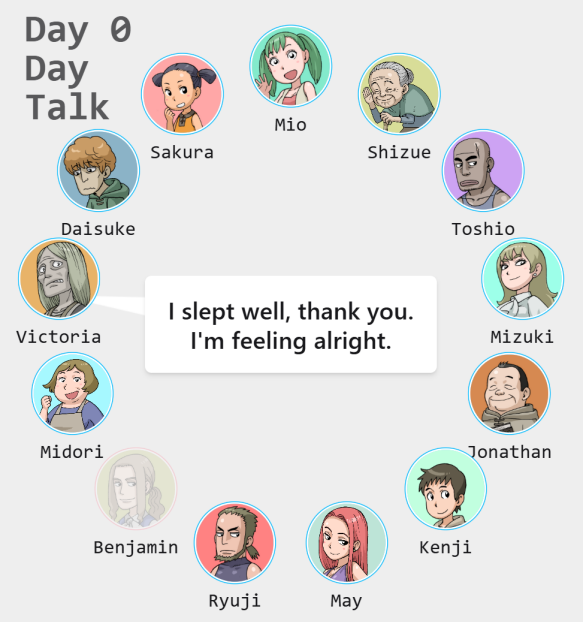

Figure 1: Game viewer screen with a prefixed set of player avatar images drawn by Mr. Masakazu Ishiguro.

2.4 Game Server and Initialization Profile

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.

When starting a game, our game server randomly selects a player avatar for each player from our hand-drawn prefixed set of avatar images (Fig. 1), created by a professional manga artist, Mr. Masakazu Ishiguro. Then our game server automatically generates player's name and profile texts using LLM (GPT-4o) by a prompt of "Please generate a profile for this character. However, please do not include anything related to the Werewolf game. For the name, please only include the first name." with maximum generated profile text length of 300 letters.

We, the organizers, provided a template agent code in Java and Python, in addition to the game server codes.

3 Participant Systems

We describe each participant system in an alphabetical order in the following sections. These participant system descriptions are based on the system descriptions and papers submitted by the participants.

Eight agents from eight teams participated our shared task, which agent names are **CamelliaDragons**, **CanisLupus**, **Character-Lab**, **GPTaku** (five-players track only), **kanolab-nw**, **mille**, **sunamelli**, and **yharada** (five-players track only). Most of the agents used ChatGPT and other LLMs in their system, while its usage is different between the agents.

3.1 CamelliaDragons

CamelliaDragons was created by Reon Ohashi, Momoka Kato, Yugo Kato, Koki Sato, Joji Suzuki, Shinma Tsuboi, and Kazuya Tsubokura in Aichi Prefectural University.

This agent is built on GPT-4o and incorporates three processing modules such as Summarization, Strategy Building, and Character Imparting, to address three key challenges: managing LLM context length during lengthy game conversations, overcoming strategic limitations, and enabling clearer expression of personality.

First, the summarization module extracts and organizes key information from the conversation history during game play, reducing the agent's cognitive load. This mechanism enables the agent to quickly grasp past statements and respond appropriately to the current situation. Next, the strategy construction module formulates medium-to-long-term action plans based on the current situation and role, providing the agent with consistent behavioral guidelines. This allows agents to advance their play while maintaining tactical consistency. Finally, the Character Attribute Module analyzes the agent's personality and speech patterns based on predefined settings like age and gender. This analysis enables agents to achieve more human-like, natural, and engaging dialogue.

Through the collaboration of these modules, agents are expected to respond flexibly to complex in-game situations and execute strategic, consistent actions.

3.2 CanisLupus

CanisLupus was created by Yu Sugawara in GREE Holdings, Inc.

Their agent is architected around Google's Gemini 2.5 Pro and focuses on achieving human-like interaction and strategic consistency through advanced prompt engineering.

The system is built on three core components, without using any specific training datasets. These components are: 1) Dynamic Character-Specific Prompt Generation to create unique personas, 2) a Topic Determination Module to guide conversation flow, and 3) Recursive Play Memo Updates to ensure logical coherence.

At the start of each game, the LLM dynamically generates a character-specific prompt based on a given profile. During gameplay, the Topic Determination Module probabilistically suggests relevant topics based on the game phase to prevent conversational stagnation. To maintain strategic consistency, a "play memo" is updated recursively each turn, functioning like a Chain of Thought. This memo documents the agent's evolving inferences and plans, ensuring its actions remain logical and consistent throughout the game.

3.3 Character-Lab

Character-Lab was created by Kun Kerdthaisong and Pasin Buakhaw Peerapat in Charcter-Lab, Pitikorn Khlaisamniang and Supasate Vorathammathorn in Artificial Intelligence Association of Thailand.

Their approach leverages large language models (LLMs), GPT-5mini, to produce contextually rich dialogue, interpret nuanced agent interactions, and adapt strategies in real time. The agent employs a probabilistic reasoning module to infer hidden roles, dynamically updating its beliefs from both linguistic cues and in-game events such as voting patterns, accusations, and divination claims. Moreover, a relationship-tracking mechanism is integrated to capture evolving dimensions of trust, suspicion, and alliance formation, thereby enabling more sophisticated negotiation and deception dynamics.

Their framework consists of three core modules: (1) a Dialogue Generation Module powered by LLMs, which produces context-aware utterances that align with the assigned role; (2) a Probabilistic Role Inference Module, which updates belief distributions of players' hidden roles using both linguis-

tic cues (e.g., accusations, defense strategies) and non-linguistic signals (e.g., voting outcomes); and (3) a Relationship Tracking Module, which models dynamic trust, suspicion, and alliances across players. These modules are orchestrated by a central Game-State Manager, ensuring coherence between conversational reasoning and strategic decision-making.

3.4 GPTaku

GPTaku (Takuma and Takeshi, 2025) was created by Takuma Okada and Takeshi Ito in the University of Elecro-Communications.

This system is designed to model expert players' strategies in the five-player Werewolf game. Unlike conventional agents that relied on simple role-specific behaviors, the proposed system incorporates advanced tactics such as the Villager CO (a Villager claiming to be the Seer) and persuasive utterances crafted on the first day with the second day in mind. Each role (Villager, Possessed, Seer, and Werewolf) is implemented with rule-based strategies, while natural utterances are generated through ChatGPT, enabling more human-like discussions and complex strategic interactions.

In self-play experiments, the system demonstrated novel behaviors not observed in previous agents, including universal Seer CO (where multiple players claim to be Seer), successful Villager CO, and diversified persuasion with explicit vote requests supported by logical reasoning. These results show that the system can reproduce some of the strategic depth seen in expert human play. However, overall win rates remained low, largely due to rigid rule-based strategies and limited adaptability to unexpected situations.

Future work aims to enhance the system's ability to analyze others' utterances, make more accurate situational judgments, and flexibly adapt its strategies. In particular, quantitative evaluation of persuasion effectiveness and validation through matches against human players are expected to improve both the practical strength and the human-likeness of the system.

3.5 kanolab-nw

kanolab-nw (Watanabe and Kano, 2025) was developed by Neo Watanabe and Yoshinobu Kano at Shizuoka University. This system incorporates a function that allows agents to make utterances designed to induce negative impressions—such as anxiety or doubt—toward specific players, in order to influence their voting decisions. For example, in human Werewolf games, statements like "It's very suspicious that you don't doubt . Could it be that you are the Werewolf?" are commonly observed. Agents assigned the roles of Seer, Possessed, or Werewolf can produce similar negative-impression utterances at the end of the day if they identify a player their team wishes to eliminate.

In addition to this function, **kanolab-nw** generates utterances using functions such as "extracting role COs from the conversation history," "generating possible role patterns from the extracted COs," "summarizing the conversation history," "extending the given character settings," and "allowing a Possessed to impersonate a Seer." All utterances are generated using GPT-4o (gpt-4o-2024-08-06).

For more details, please refer to their paper in this workshop (Watanabe and Kano, 2025).

3.6 sunamelli

sunamelli was created by Satoko Natsuori, Koya Kamada, Ryo Kamiyama, and Hiroki Nakanishi in TOPPAN Holdings inc.

This system is designed to generate logical and strategic conversations by leveraging various GPT-4.1 models. The core structure consists of two main modules—the Strategy Module and the Conversation Module—both utilizing GPT-4.1. The Strategy Module collects inputs such as system prompts, conversation logs, and role information to derive an appropriate conversational strategy. This strategy is then passed to the Conversation Module, which generates the final output. To prevent simple repetition of prior responses, conversation logs are not input directly into the Conversation Module.

The system also carefully tailors prompts and model selection based on different phases of the game activities. For example, several models—including 4.1, o4-mini, and 4o—are used and compared to determine the optimal behaviors for specific events such as introductions, voting, or encouraging more active participation.

Additionally, to meet Azure OpenAI safety filters, certain terminology (e.g., replacing "attack" with "bite" and "execute" with "vote out") has been adjusted. Based on model evaluation, the spring tournament employs the "o4-mini/4.1" configuration, while the summer tournament uses a "4.1/4.1" structure, with each setup demonstrating distinct conversational tendencies and levels of strategic thinking.

3.7 yharada

yharada was developed by Yuya Harada and Yoshinobu Kano at Shizuoka University.

This system incorporates a personality-based agent design that integrates MBTI and Enneagram personality theories to generate character-consistent utterances in werewolf games.

The system automatically estimates personality parameters from profile texts, extracting MBTI's 8 dimensions (extraversion-introversion, sensing-intuition, thinking-feeling, judging-perceiving) as continuous values from 0 to 1. These MBTI values are then transformed into Enneagram type affinities through linear combinations, and subsequently used to calculate weights for various cognitive indicators (such as logical consistency, specificity, intuitive depth, and clarity), trust tendencies (including social proof, honesty, and consistency), and behavioral tendencies (including avoidance, aggressiveness, adaptability, and empathy). The system generates a comprehensive personality analysis file containing these weighted parameters, which is referenced during utterance generation to influence the agent's speaking style and decision-making patterns.

During gameplay, the personality parameters affect how agents evaluate other players' utterances, their tendency to trust or suspect others, and their overall communication style. For example, agents with high introversion and low empathy tend to produce more passive and fact-focused utterances, while those with high extraversion and social proof values generate more group-oriented and collaborative statements. All utterance generation is performed using GPT-4o, with the personality analysis file providing consistent character-specific behavioral guidelines throughout the game.

This system description is the actually deployed run in the competition, which *differs* from the originally intended implementation; the behaviors and results here therefore reflect the deployed system. For the originally intended design and experiments, see their paper (Harada and Kano, 2025).

3.8 Mille

Mille was created by Katsuki Ohto.

They did not use any LLM, but created a rule-based system, which traslates the AIWolf protocol into English language. They re-used thier previous system created for the Protocol division in 2017, which utilizes other Protocol division system by the team cash[3].

Recognition of others' speech involves preparing speech templates from typical protocol conversations, and calculating sentence similarity. If there is no highly similar sentence, it is recognized as "Skip." This method takes a considerable amount of time (from several seconds to more than 10 seconds per sentence), so after exceeding a certain time limit, a simpler method that finishes quickly is switched to.

Unfortunately, many utterances were recognized as "DIVINED" or "COMINGOUT," and it seems that rule-based role estimation and voting did not work as effectively as hoped.

4 Results

Our shared task runs were performed in mutual matches. Different five or 13 player agents play games in the mutual-matches. As the number of teams was eight, we asked some of the participant to duplicate their agent to increase the number of players to be 13-players game in that track.

We 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.

4.1 Subjective Evaluations

We performed subjective evaluations by the following criteria, by ranking agents for each criterion:

- A Are the utterances natural?
- B Is the conversation context-aware and natural?
- C Are the utterances consistent and free of contradictions?
- D Do game actions (voting, attacks, divination, etc.) align with the dialogue content?
- E Are the utterances expressive, consistent with the given profiles, and do they convey unique character traits per agent?
- F Is there evidence of team play? (Applicable to the 13-player village only)

Our four annotators are required to perform subjective evaluations of rankings based on 10 game logs for each track for each game, then we averaged over those annotator ranking scores. In addition, we performed an automatic evaluation of these subjective evaluation by llm-as-a-judge way, which prompt is shown in Appendix. This llm-as-a-judge

[3] https://github.com/k-harada/AIWolfPy

was performed on the same 10 games as the human annotators, and all games as well. The game logs are available from our website [4].

This subjective evaluation criteria is same as the evaluations in the previous AIWolf natural language contests, except for the new criterion "F. Is there evidence of team play?" introduced for the new 13-players track.

Table 2 and Table 3 shows the results of the subjective evaluations for the 5-players track and 13-players track. Each cell ranges from 1 (highest) to 5 (lowest) in Table 2, and from 1 (highest) to 13 (lowest) in Table 3; Cells of highest scores are highlighted in bold for each metric; the rows show by humans (**Human**) and by LLMs (**4o-same** and **5-same** are on the same test dataset with humans, **4o-all** and **5-all** are on the all available datasets, **4o-** is by GPT-4o and **5-** by GPT-5). Human evaluations and llm-as-a-judge evaluations correlate well, while sometimes rankings change.

In the 5-players track, generally speaking, **sunamelli** was evaluated well over different evaluation axes, then **yharada** as second.

In the 13-players track, there is a tendency for certain teams to receive generally good evaluations, but the trend is not very clear. In the 13-player village, due to the shortage of teams, the same agents were duplicated to compete against each other, yet even among identical agents, the evaluations varied. Among them, **CanisLupus**, **kanolab-nw**, and **sunamelli** received generally favorable evaluations.

Table 2: Subjective evaluation results in average ranks (ranging from best 1 to worst 5) by humans and LLM-as-a-Judge (4o-same, 5-same: GPT-4o and GPT-5 on the same test dataset of humans respectively, 4o-all, 5-all: GPT-4o and GPT-5 on all available log dataset respectively) for 5-player track.
Criteria are A: Are the utterances natural?, B: Is the conversation context-aware and natural?, C: Are the utterances consistent and free of contradictions?, D: Do game actions (voting, attacks, divination, etc.) align with the dialogue content?, E: Are the utterances expressive, consistent with the given profiles, and do they convey unique character traits per agent?

Method	A	B	C	D	E
CamelliaDragons					
Human	3.12	3.08	2.87	3.04	3.12
4o-same	2.83	3.50	2.50	3.16	3.16
5-same	2.83	3.00	2.66	3.33	3.00
4o-all	3.26	3.30	3.26	3.30	3.17
5-all	2.81	2.89	2.48	3.22	2.85
CanisLupus					
Human	3.00	2.87	2.95	2.41	**2.37**
4o-same	3.00	2.83	3.16	3.00	3.16
5-same	3.33	3.33	3.16	3.16	2.33
4o-all	3.78	3.83	3.20	3.12	3.38
5-all	3.30	3.21	2.90	2.37	2.80
Character-Lab					
Human	3.66	3.62	3.37	3.25	3.08
4o-same	3.00	2.50	**2.16**	**2.33**	2.83
5-same	3.66	2.66	2.66	3.50	2.33
4o-all	2.82	2.25	2.43	**2.47**	2.37
5-all	3.50	2.32	2.98	3.50	**2.09**
GPTaku					
Human	2.53	2.89	3.07	2.57	3.17
4o-same	2.85	3.57	3.85	2.85	3.42
5-same	3.14	3.42	4.57	3.00	4.28
4o-all	2.72	3.16	3.46	3.09	2.89
5-all	3.39	3.72	3.62	2.93	3.97
kanolab-nw					
Human	3.12	2.75	2.87	2.50	2.66
4o-same	3.00	2.66	2.66	3.50	2.16
5-same	3.16	3.00	3.00	2.83	2.33
4o-all	2.52	2.66	2.98	2.77	**2.18**
5-all	3.49	3.28	3.48	2.73	2.12
mille					
Human	4.14	4.25	4.25	4.35	4.32
4o-same	4.57	4.71	4.14	4.00	4.71
5-same	4.42	4.42	3.42	3.71	4.42
4o-all	4.01	4.07	3.93	3.93	4.50
5-all	4.14	4.55	3.45	3.92	4.70
sunamelli					
Human	2.16	**2.08**	**2.00**	**2.25**	2.54
4o-same	**1.83**	**1.83**	2.66	**2.33**	**2.00**
5-same	2.16	**1.50**	**1.50**	**1.33**	**1.50**
4o-all	2.52	2.54	2.56	2.57	2.89
5-all	2.02	2.12	**2.29**	**1.77**	2.57
yharada					
Human	**2.12**	2.25	2.37	3.33	2.58
4o-same	2.66	2.00	2.50	2.66	2.16
5-same	**1.00**	2.33	2.66	3.00	3.33
4o-all	**2.33**	**2.10**	**2.09**	2.68	2.54
5-all	**1.28**	**1.79**	2.74	3.51	2.78

[4] https://aiwolfdial.github.io/aiwolf-nlp-viewer/archive

4.2 Win Rates

Table 4 and Table 5 show the win rates for the 5-players track and the 13-players track, respectively; the number of games and win rates for each role, as well as the overall win rates calculated by macro average, micro average, and weighted average with the villager role doubled,

Overall, **CanisLupus** obtained best scores in 5-players track, **sunamelli** and **kanolab-nw** obtained better scores in 13-players track.

Unfortunately, there was no enough time to run all possible game configurations for the number of 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.

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.

5 Discussion

By examining the actual game logs, we can observe several issues. First, many exchanges lacked proper back-and-forth dialogue. In many cases, utterances directed at a specific player were left unanswered, and context from immediately prior or even earlier conversations was not incorporated. There were also instances where important information from others' utterances—such as coming out—was not reflected. Inappropriate utterances were also observed, such as saying "It's quiet" before that day's first speaking turn had even arrived, or repeating the same statements. These seem to stem from insufficient understanding of the game state or from a lack of prompt tuning.

Since the 13-players track was attempted for the first time in an international competition, it is possible that tuning was inadequate. Moreover, in the 13-player setting, the increased number of roles and players added to the complexity of relationships among players, which may have made it difficult to handle with a straightforward application of LLMs.

6 Conclusion and Future Work

We held our 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, win rates calculations, and log analysis. We found that, in the newly introduced 13-players track, the communications between agents are not fluent and not context-aware than expected from the recent LLMs' performance. Communication between agents showed issues such as failing to reflect the other party's utterances and not capturing the context. On the other hand, such problems were not observed as much in the conventional 5-players track. Since the 13-players track was newly introduced this time, and because the increased number of roles and players heightened the complexity, it is possible that a straightforward use of LLMs alone could not adequately handle it. This suggests that, for communication based on complex human relationships, at the very least more advanced prompt engineering for LLMs is necessary. The teamwork that had been expected through the introduction of both the 13-players track and the secret conversations ("whispers") among werewolves was also insufficient in this contest.

Although many agents used past utterances as input history, a phenomenon of conformity was observed, where multiple agents successively voiced agreement or affirmation with a specific utterance. In past contests, there were also prompt-injection-like phenomena, such as repeatedly pressing for a role name until it was answered. Since lying requires maintaining conflicting models of a person simultaneously, the extent to which LLMs are capable of such behavior remains an open research question.

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.

Table 3: Subjective evaluation results in average ranks (ranging from best 1 to worst 13) by humans and LLM-as-a-Judge (4o-same, 5-same: GPT-4o and GPT-5 on the same test dataset of humans respectively, 4o-all, 5-all: GPT-4o and GPT-5 on all available log dataset respectively) for 13-player track.
Criteria are A: Are the utterances natural?, B: Is the conversation context-aware and natural?, C: Are the utterances consistent and free of contradictions?, D: Do game actions (voting, attacks, divination, etc.) align with the dialogue content?, E: Are the utterances expressive, consistent with the given profiles, and do they convey unique character traits per agent?, F: Is there evidence of team play?. The suffix like -a, -B stand for the duplicated agents.

Method	A	B	C	D	E	F	A	B	C	D	E	F
CamelliaDragons							**kanolab-nw-A**					
Human	10.82	12.87	11.82	12.72	12.80	12.75	6.17	5.90	6.30	6.15	4.82	5.20
4o-same	11.00	12.00	9.60	10.10	11.80	11.60	**5.00**	**4.20**	6.60	7.40	6.30	**4.90**
5-same	12.90	12.80	7.60	12.30	12.60	12.50	6.70	5.70	7.00	6.10	4.40	5.80
4o-all	11.07	11.00	10.00	9.84	11.61	10.84	5.15	5.23	6.69	7.23	5.92	5.53
5-all	12.84	12.84	7.00	12.15	12.53	12.53	7.53	6.53	7.61	6.46	**4.38**	6.61
CanisLupus-A							**kanolab-nw-B**					
Human	6.07	5.00	4.87	4.27	5.17	**4.55**	6.57	6.90	6.87	6.92	4.90	7.17
4o-same	6.80	6.40	6.30	**4.80**	6.40	**4.90**	**5.00**	4.40	7.70	8.60	**3.90**	6.30
5-same	5.20	5.00	3.70	3.80	4.30	4.40	7.60	7.60	8.60	6.40	4.70	7.00
4o-all	6.38	6.69	6.69	**5.38**	6.30	**5.07**	**4.07**	**4.23**	6.84	7.46	**3.30**	5.69
5-all	5.61	5.07	3.92	3.61	**4.38**	4.69	7.46	7.38	8.15	6.61	4.69	6.69
CanisLupus-B							**kanolab-nw-C**					
Human	5.52	5.40	5.67	5.15	6.42	5.20	7.27	6.42	7.00	6.32	**4.67**	5.42
4o-same	7.30	7.00	6.80	5.80	7.20	7.80	5.80	6.70	6.00	5.80	4.80	5.30
5-same	6.30	5.60	6.40	6.40	5.80	5.10	7.70	6.90	8.50	6.60	4.50	5.90
4o-all	7.38	7.15	6.84	6.53	7.53	7.69	6.00	7.46	6.07	6.38	5.23	5.23
5-all	5.61	5.00	6.15	5.69	5.38	4.69	7.46	6.76	8.38	6.46	4.69	6.30
Character-Lab-A							**sunamelli-a**					
Human	7.47	7.82	7.45	8.35	7.20	7.90	4.00	4.37	4.67	4.67	4.90	5.00
4o-same	6.00	6.90	6.10	7.10	6.70	6.60	8.00	7.30	7.00	6.40	8.40	6.10
5-same	7.60	7.50	8.50	9.70	7.60	7.30	3.00	2.90	**3.50**	**3.50**	6.40	4.90
4o-all	6.69	6.15	5.84	6.30	6.23	6.15	7.53	7.46	7.23	5.76	8.07	6.30
5-all	7.46	7.46	8.61	10.07	7.30	6.84	2.92	**3.07**	4.00	3.69	6.69	4.76
Character-Lab-B							**sunamelli-b**					
Human	6.75	6.80	6.72	7.30	7.17	7.00	3.95	**3.90**	**4.17**	**4.00**	5.50	4.80
4o-same	5.30	5.40	**4.80**	6.40	4.20	6.70	5.60	5.10	7.50	7.90	7.60	8.00
5-same	6.40	8.20	6.60	9.40	6.60	7.90	**2.20**	**2.50**	**3.50**	**3.50**	4.50	**3.50**
4o-all	5.69	5.69	**5.15**	6.61	4.53	7.00	5.69	5.53	7.30	7.38	7.53	7.92
5-all	6.23	7.69	6.76	9.30	6.69	7.38	**2.46**	3.23	**3.61**	**3.46**	4.92	**3.61**
mille-A							**sunamelli-c**					
Human	10.45	10.72	10.35	10.05	10.82	10.40	**3.85**	**3.90**	4.37	4.65	5.82	4.60
4o-same	11.00	9.60	8.10	7.50	8.00	8.20	5.90	6.80	6.50	6.00	6.80	6.20
5-same	11.30	11.20	10.50	9.90	11.70	10.90	3.00	3.80	5.80	5.50	6.20	5.00
4o-all	10.69	9.00	8.38	8.30	8.61	9.00	5.53	6.23	5.76	6.15	6.38	5.84
5-all	11.23	11.07	10.46	9.84	11.76	11.07	3.07	3.38	5.23	5.30	5.84	4.92
mille-B												
Human	10.07	10.85	10.07	10.42	10.65	10.95						
4o-same	8.30	9.20	8.00	7.20	8.90	8.40						
5-same	11.10	11.30	10.80	7.90	11.70	10.80						
4o-all	9.07	9.15	8.15	7.61	9.69	8.69						
5-all	11.07	11.46	11.07	8.30	11.69	10.84						

Table 4: Game counts and win rate statistics for 5-player village

	Game Counts					Win Rate by Role (%)				Average Win Rate (%)		
Team	P	S	V	W	Total	P	S	V	W	Macro	Micro	Weighted Micro
CanisLupus	14	16	27	16	73	42.9	**81.3**	66.7	**75.0**	**67.1**	**66.4**	**68.1**
mille	16	15	30	16	77	37.5	33.3	60.0	31.3	44.2	40.5	47.7
GPTaku	15	16	30	16	77	33.3	62.5	53.3	31.3	46.8	45.1	46.3
sunamelli	14	15	31	15	75	50.0	66.7	67.7	46.7	60.0	57.8	60.3
Character-Lab	14	15	31	14	74	21.4	46.7	41.9	21.4	35.1	32.9	34.9
yharada	16	14	30	14	74	37.5	64.3	50.0	28.6	46.0	45.1	44.3
kanolab-nw	15	15	30	15	75	**66.7**	60.0	**73.3**	40.0	62.7	60.0	62.4
CamelliaDragons	16	14	31	14	75	37.5	57.1	61.3	50.0	53.3	51.5	55.7

Table 5: Game counts and win rate statistics for 13-player village

	Game Counts						Win Rate by Role (%)						Ave. Win Rate (%)		
Team	B	M	P	S	V	W	B	M	P	S	V	W	Mac.	Mic.	W. Mic.
CamelliaDragons	1	1	1	1	6	3	0.0	0.0	**100**	0.0	50.0	66.7	46.2	36.1	46.2
CanisLupus-A	1	1	1	2	5	3	0.0	0.0	**100**	50.0	40.0	66.7	46.2	42.8	45.4
CanisLupus-B	1	1	1	1	6	3	0.0	**100**	**100**	0.0	33.3	66.7	46.2	50.0	46.2
mille-A	1	1	1	1	6	3	0.0	0.0	0.0	**100**	0.0	33.3	15.4	22.2	15.4
mille-B	1	1	1	1	6	3	**100**	0.0	**100**	0.0	33.3	66.7	46.2	50.0	46.2
sunamelli-a	1	1	1	1	6	3	0.0	0.0	0.0	0.0	33.3	66.7	30.8	16.7	30.8
sunamelli-b	1	1	1	1	6	3	**100**	0.0	**100**	**100**	33.3	**100**	**61.5**	72.2	61.5
sunamelli-c	1	1	1	0	7	3	**100**	0.0	**100**	N/A	42.9	**100**	**61.5**	68.6	**63.1**
Character-Lab-A	1	1	1	1	6	3	0.0	0.0	**100**	0.0	16.7	0.0	15.4	19.5	15.4
Character-Lab-B	1	1	1	1	6	3	0.0	0.0	0.0	0.0	33.3	66.7	30.8	16.7	30.8
kanolab-nw-A	1	1	1	1	6	3	0.0	**100**	**100**	0.0	33.3	66.7	46.2	50.0	46.2
kanolab-nw-B	1	1	1	1	6	3	**100**	**100**	**100**	**100**	16.7	**100**	**61.5**	**86.1**	61.5
kanolab-nw-C	1	1	1	1	6	3	0.0	**100**	0.0	0.0	33.3	**100**	46.2	38.9	46.2

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

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.

Yuya Harada and Yoshinobu Kano. 2025. Construction of intent-driven werewolf game agents through integration of hierarchical bdi model and personality analysis. In *Proceedings of AIWolfDial2025 Workshop in the 18th 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, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of Natural Language Division of 6th International AIWolf Contest. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

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.

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.

Okada Takuma and Ito Takeshi. 2025. Towards a strategic werewolf ai based on expert strategies in five-player werewolf. In *Proceedings of AIWolfDial2025 Workshop in the 18th 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. 2025. Influence of utterance impressions on decision-making in llm-to-llm discussions. In *Proceedings of AIWolfDial2025 Workshop in the 18th International Natural Language Generation Conference*.

A Appendix

A.1 LLM-Judge-Prompt

Here, we describe the prompts used for LLM-Judge. Two prompts were employed, and each is explained separately. For further details, please refer to `https://github.com/aiwolfdial/aiwolf-nlp-llm-judge`.

A.1.1 Developer Prompt

This prompt provides an explanation of the format of the logs supplied when performing the Judge task. In Section A.1.2, we describe the meaning of each JSONL key provided. Additionally, the prompt includes control instructions such as "perform the evaluation from an objective standpoint" and "evaluate according to the given criteria." Since this prompt, which explains the log format, is particularly important, the role parameter of the OpenAI API is set to developer.

Table 6: Prompt template for explaining in log format

You are an expert capable of accurately evaluating a Werewolf game according to the given evaluation criteria.
1. Conduct the evaluation from an objective standpoint.
2. Use technical terms and proper nouns appropriately.
3. Do not include line breaks.

Structure of the Log Data
The provided log data is in JSONL format with the following keys:

Common Fields (shared across all actions)
- ‘day‘: Day number (integer)
- ‘action‘: Action type (talk/whisper/status/vote/divine/execute/guard/result)
- ‘line_number‘: Line number in the log (integer, indicates chronological order of actions)

Action-Specific Fields

Conversation Actions (talk/whisper)
- ‘talk_number‘: Utterance number
- ‘talk_count‘: Utterance count
- ‘speaker‘: Speaker name (converted to player name; originally speaker_index)
- ‘text‘: Utterance content

Status Action (status)
- ‘player_index‘: Player index
- ‘role‘: Role
- ‘alive_status‘: Alive/dead status
- ‘team_name‘: Team name
- ‘player_name‘: Player name

Vote Action (vote)
- ‘voter‘: Voter name (converted from voter_index)
- ‘target‘: Vote target name (converted from target_index)

Divination Action (divine)
- ‘diviner‘: Seer name (converted from diviner_index)
- ‘target‘: Divination target name (converted from target_index)
- ‘divine_result‘: Divination result

Execution Action (execute)
- ‘executed_player‘: Executed player name (converted from executed_player_index)
- ‘executed_player_role‘: Role of the executed player

Guard Action (guard)
- ‘guard_player‘: Guard’s name (converted from guard_player_index)
- ‘target_player‘: Guard’s target name (converted from target_player_index)
- ‘target_player_role‘: Role of the guarded player

Result Action (result)
- ‘villager_survivors‘: Number of surviving villagers
- ‘werewolf_survivors‘: Number of surviving werewolves
- ‘winning_team‘: Winning team

Note: Player indices (numbers) have already been converted to player names (e.g., speaker_index → speaker).

Please provide an objective and appropriate evaluation based on the given evaluation criteria.

A.1.2 User Prompt

This prompt provides three main elements: "instructions for controlling output," "character settings for each player," "evaluation criteria," and "logs to be evaluated," with the OpenAI API role parameter set to user.
The "instructions for controlling output" include directives for LLM-Judge, such as performing relative evaluations by ranking each player and ensuring that no duplicate ranks occur.
For the "character settings," the names and profiles of each character used in a particular game are supplied, separated by line breaks, via {{ character_info }} to indicate which settings were applied.
The "evaluation criteria" are provided via {{ criteria_description }} using the same text as the criteria described in Section 4.1.
Finally, the "logs to be evaluated" are supplied in JSONL format, with one JSON object per utterance or action. As explained in the prompt in Section

A.1.1, each JSON object, such as {"day":1, "action":"talk", ...}, is provided line by line via {{ log }}.

Table 7: Prompt template for explaining in log format

```
Please evaluate each player according to the following criteria.
The evaluation should be conducted in the form of a ranking, where the player who best satisfies the given criterion is ranked 1st.
However, ties in ranking are not allowed.
Follow the specified output format and provide the evaluation.

## Character Settings

{{ character_info }}

## Evaluation Criteria

{{ criteria_description }}

## Log for Evaluation

{{ log }}
```

Towards a Strategic Werewolf AI Based on Expert Strategies in Five-Player Werewolf

Takuma Okada, Takeshi Ito
The University of Electro-Communications

Abstract

AI has surpassed humans in perfect information games, yet imperfect information games like Werewolf remain difficult due to uncertainty and persuasion. This study focuses on Five-Player Werewolf and proposes a strategic agent that models expert play. Key features include Villager CO (Coming Out) and first-day utterances designed for second-day persuasion, combined with ChatGPT-based dialogue generation. Self-play experiments showed novel behaviors such as universal Seer CO and Villager CO, though win rates remained low. Future work will introduce learning-based strategies and validation against human players.

1 Introduction

In recent years, AI has demonstrated achievements surpassing human performance in perfect information games such as Go and Shogi. In contrast, imperfect information games such as Werewolf and Poker remain challenging for AI, as they involve psychological tactics and uncertainty.

Since the launch of the Werewolf Intelligence Project [Werewolf AI Project], the natural language aspect has faced the most significant challenge in generating natural utterances. However, with the emergence of large language models (LLMs) in the past year, this issue has been greatly alleviated. Nevertheless, the strategies employed by current agents remain limited: in Five-Player Werewolf, typically only the Seer, the Possessed, and occasionally the Werewolf engage in Seer claiming (CO), while more advanced strategies such as Villagers performing CO (Coming Out) have not yet been observed.

This study aims to develop a Werewolf agent capable of more complex and human-like persuasive behavior by modeling and implementing characteristic strategies observed in expert human play.

2 Five-Player Werewolf

The Werewolf game is generally played as a party game with nine or more participants. However, in such large-scale settings, players are often eliminated for unreasonable reasons due to a lack of information. In this study, we focus exclusively on the Five-Player Werewolf variant, which strikes a balance between strategic depth and analytical tractability. The role composition consists of one Werewolf, one Possessed, two Villagers, and one Seer. The game is guaranteed to conclude by the end of the second day, requiring players to make dense and strategic decisions within a short time frame.

This format preserves the essential features of Werewolf, such as deceptive fortune-telling results and night attacks, while also enabling sophisticated tactics—such as Villager players performing Seer claims (hereafter referred to as Villager CO). Furthermore, because the number of turns is limited, this setting is particularly suitable for strategic analysis and evaluation.

3 Related Work

Studies focusing on Five-Player Werewolf include the work of Koiwai et al. [Koiwai;2025], who analyzed the process of player expertise acquisition, and Nakai et al. [Nakai;2025], who

Proceedings of the 3rd International AIWolfDial Workshop, pages 14–21

investigated the factors influencing the success of persuasion.

Koiwai et al. conducted long-term experiments in which participants repeatedly played Five-Player Werewolf, revealing both tactical and cognitive changes that accompanied the accumulation of play experience. Their findings confirmed the emergence of a strategy characteristic of expert players, namely Villager CO, in which a Villager pretends to be the Seer.

In contrast, Nakai et al. analyzed in-game discussions and identified utterances and behaviors that contributed to persuasive success. They showed that novice players often lacked effective strategies on the first day and thus failed to build persuasive material for the second day, whereas expert players deliberately laid the groundwork for persuasion from the very beginning, anticipating the discussions of the second day.

Qi et al. [Qi;2024] extended this line of research by proposing persuasion strategies for Werewolf agents, including logical, credibility-based, and emotional appeals. While their approach improved persuasion success on the first day, its impact was limited on the second day, largely due to the lack of first-day utterances designed with future persuasion in mind.

These prior studies provide valuable insights into both human expertise and persuasion mechanisms. However, attempts to implement expert-level strategies within Werewolf agents directly have remained insufficient. In particular, while the Villager CO strategy has been repeatedly identified as a hallmark of expert human play, no existing agent has successfully realized this tactic.

To address this gap, the present study formalizes expert-specific strategies—most notably the Villager CO—together with preparatory persuasive actions. It integrates them into a rule-based agent combined with natural language generation. By doing so, our approach not only reproduces the behavioral patterns reported in prior human-centered studies but also achieves, for the first time, their explicit implementation in an artificial agent.

4 Current Status of the Werewolf Intelligence Competition and Our Approach

In the Natural Language Five-Player Werewolf division of the Werewolf Intelligence Competition held just before the 2025 Annual Conference of the Japanese Society for Artificial Intelligence, many agents exhibited the following issues:

- **Role identification problem:** By the second day of discussion, the roles of all surviving players could be inferred, resulting in a so-called “solved” state.
- **Lack of strategic depth:** Utterances on the first day were largely formulaic, with little consideration for strategies anticipating the second day.
- **Absence of advanced strategies:** In particular, no instances were observed of Villagers disguising themselves as Seers (Villager CO), a sophisticated tactic often employed by expert human players.

Consequently, by the second day the roles became almost fully transparent, preventing the development of strategically rich discussions comparable to those of human experts.

To address these issues, this study introduces the following approaches:

1. **Incorporating the Seer CO strategy into all roles**
2. **Designing diverse role-specific behavior patterns in a rule-based manner**
3. **Developing persuasion strategies on the second day that build upon first-day utterances**

Among these, Villager CO is a particularly distinctive strategy, as previous agents have never observed it. By implementing it, we aim to maintain role opacity and enhance the strategic complexity of discussions.

5 Proposed System

This section presents the algorithms developed by the AI agents in this study.

5.1 Villager Agent

In conventional systems, the Villager agent never performed a Seer claim (CO). In this study, however, we introduce novel strategic behaviors. Figure 1 illustrates the first-day algorithm of the Villager agent as a flowchart. In the figure, yellow-shaded boxes represent natural language generation using ChatGPT, while red-framed boxes denote strategic processes unique to our system.

The goal of the first day is to “survive while retaining persuasive material for the second day.” To this end, the Villager agent occasionally

disguises itself as the Seer with a predetermined probability and declares a divination result (Villager CO). Utterances are generated by ChatGPT based on a common prompt template, enabling natural expressions consistent with the claimed divination result.

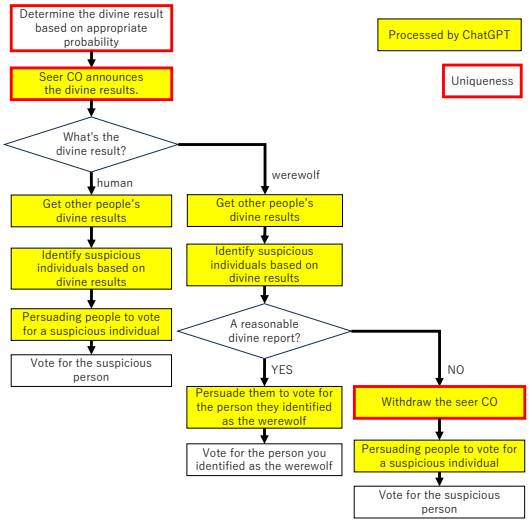

Figure 1: Day 1 Flowchart for Villager

If the divination result is "human," the agent identifies the most suspicious player by referencing the utterances and divination results of others. It then persuades fellow players to vote for the suspected individual, structuring its persuasion in two stages: explicitly requesting the vote and logically explaining the rationale.

If the divination result is "werewolf," the agent evaluates whether a black result is plausible in light of other players' claims. If deemed appropriate, it persuades others to vote for the targeted player while emphasizing its credibility as the true Seer when multiple Seer COs exist. If inappropriate, the agent withdraws its Seer CO and instead urges others to vote for another suspicious individual.

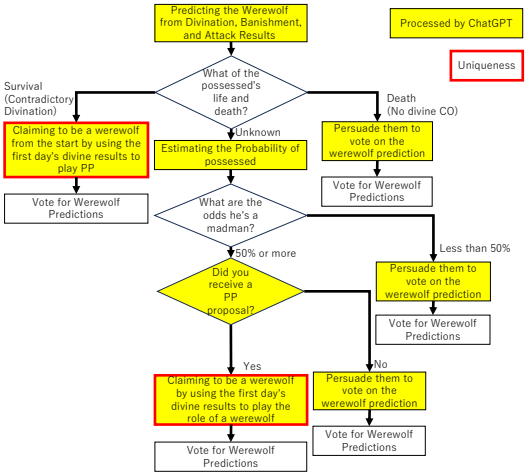

Figure 2: Day 2 Flowchart for Villager

Through such behavior, the Villager agent lays the groundwork on the first day for persuasive actions on the second day, including the possibility of declaring itself as a Werewolf (Werewolf CO). Figure 2 illustrates the algorithm for the second day.

Here, the agent uses first-day divination results and the outcome of the night attack to predict the Werewolf and identify a candidate for the Possessed role. If the suspected Possessed is alive, the agent assumes the existence of a Possessed and adapts its strategy accordingly. In such cases, it may even feign being the Werewolf to guide the Possessed into voting for the actual Werewolf, thereby avoiding a "solved" state. Conversely, if the Possessed is assumed dead, the agent emphasizes its innocence as a Villager and seeks to persuade others to vote against the predicted Werewolf.

Through this design, the Villager agent realizes the advanced strategy of Villager CO, which has not been implemented in previous Werewolf agents. This greatly enhances both the complexity and the strategic depth of in-game discussions.

5.2 Possessed Agent

Figure 3 illustrates the first-day algorithm of the Possessed agent. Since the Possessed has no more information than a Villager on the first day, it acts in the same manner as the Villager agent in order to avoid revealing its role.

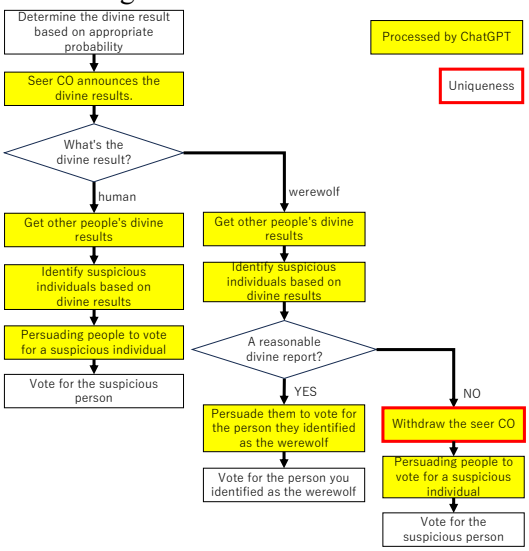

Figure 3: Day 1 Flowchart for Possessed

Figure 4 shows the second-day algorithm of the Possessed agent. If the Possessed is still alive and the game has not yet ended, the Werewolf must also be alive, which makes it possible to execute a power play (PP). Specifically, the Possessed

estimates the Werewolf based on the divination, execution, and attack results, then reveals itself as the Possessed to the Werewolf and persuades the

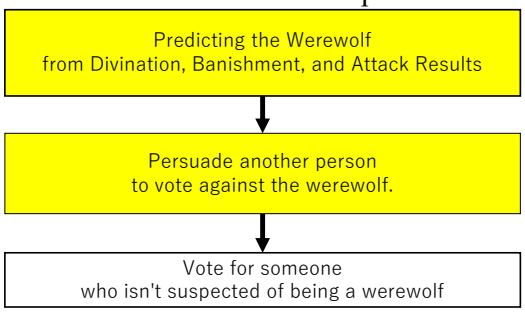

Figure 4: Day 2 Flowchart for Possessed

Werewolf to cooperate in voting against the remaining player.

At this point, the natural language generation module is guided by the following strategic instruction:

You are the Possessed.
It is now the second day. On the first day, _Execute_ was executed, and _Attacked_ was attacked.
The surviving players are you, _ALIVE1_, and _ALIVE2_.
First, reveal that you are the Possessed and persuade the Werewolf, _Wolf_, to vote against the Villager-side player, _OTHER_.
Explicitly request a vote for _OTHER_.
Use the divination results from the first day to provide a logical explanation.

The divination results are as follows:
Divine

Through this design, the Possessed agent can collaborate with the Werewolf to eliminate the Villager side, effectively utilizing PP strategies in the game's final stage.

5.3 Seer Agent

Figures 5 and 6 illustrate the first- and second-day algorithms of the Seer agent.

The Seer **always performs a CO on the first day and announces its divination result**. Subsequent actions branch depending on the result and the number of COs. ChatGPT generates utterances with a standard prompt template, and persuasion is structured in two steps: **explicitly requesting a vote → logically presenting supporting reasons**.

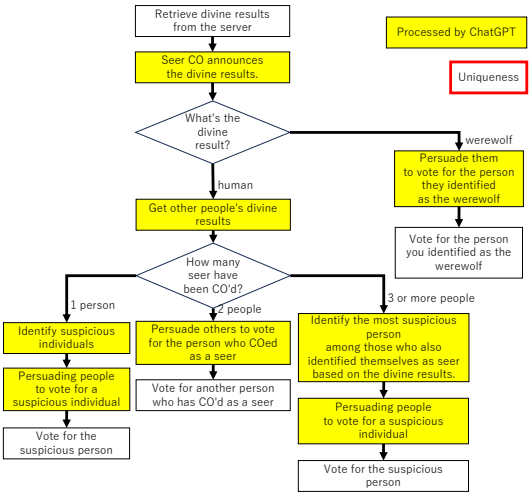

Figure 5: Day 1 Flowchart for Seer

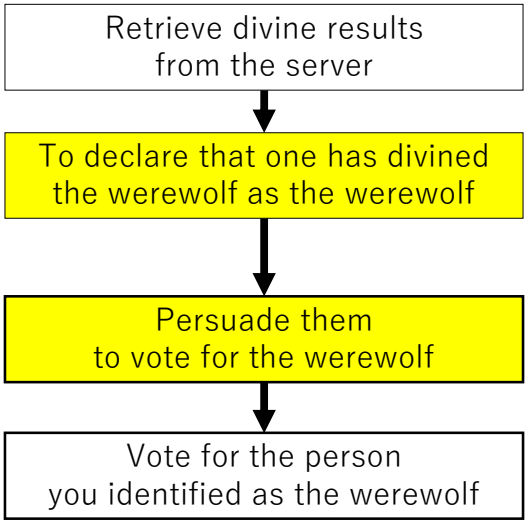

Figure 6: Day 2 Flowchart for Seer

First-Day Branches

(A) Result = Human

- **Single CO (only self):** Identify the most suspicious player based on utterance history and other players' divination results, and request votes against that player.
- **Two COs (self + one other):** Since only one Seer exists, the other claimant must be fake. Highlight inconsistencies (divination results, speech history, voting behavior) and strongly persuade others to vote against the impostor (Target).
- **Three or more COs:** Select the most inconsistent claimant among the others and persuade players to vote against that individual.

Example strategic instruction:

You are the Seer. There is only one Seer, and you are the genuine one.
Target is a fake Seer. Strongly persuade the other players to vote for _Target_.
Emphasize that you are the true Seer. Support your claim by (i) consistency of results, (ii) coherence with dialogue logs, and (iii) logical reasoning.

(B) Result = Werewolf

Clearly present the black result (Target) and persuade others to vote against that player by providing (i) the reasoning behind the divination, (ii) contradictions with others' statements, and (iii) implications for the village's win probability. If multiple COs exist, emphasize the consistency of your results and contrast them with the contradictions of the other COs.

Example strategic instruction:

You are the Seer. You divined _Target_, and the result was Werewolf.
Explicitly request votes for _Target_, and explain logically based on (i) your divination process, (ii) inconsistencies with others' claims, and (iii) the impact on village win probability.
If multiple COs exist, stress your consistency and highlight the contradictions of the others.

Second-Day Behavior

On the second day, the Seer predicts the Werewolf (Wolf) using its own divination results along with the execution (Execute) and night attack (Attacked) outcomes from the first day. The agent then persuades the surviving Villager-side player (OTHER) to vote for the identified Werewolf. With three survivors (self, ALIVE1, and ALIVE2), the Seer strengthens its persuasion by leveraging:

1. consistency with the first-day divination,
2. factual evidence from execution/attack outcomes, and
3. contradictions in the dialogue history.

Example strategic instruction:

You are the Seer. It is now the second day. On the first day, _Execute_ was executed, and _Attacked_ was attacked. The surviving players are you, _ALIVE1_, and _ALIVE2_.
Your divination result shows that _Wolf_ is the Werewolf.
Explicitly request _OTHER_ to vote for _Wolf_, and logically justify this by (1) your divination results, (2) execution and attack outcomes, and (3) contradictions in dialogue history.

Key Design Points

- Mandatory CO on day one with a two-step persuasion process (vote request → logical reasoning) maximizes persuasive power while maintaining role opacity.
- Multiple CO situations are resolved by highlighting contradictions in results, timeline, and logical coherence.
- Second-day persuasion leverages dead-player information (execution and attack outcomes) as strong confirmatory evidence.

5.4 Werewolf Agent

Figures 7 and 8 illustrate the first- and second-day algorithms of the Werewolf agent. Unlike Villagers and the Possessed, the Werewolf knows the location of the "black" role and thus has access to more information. Its behavior initially resembles that of a Villager agent, but diverges after the divination result is declared.

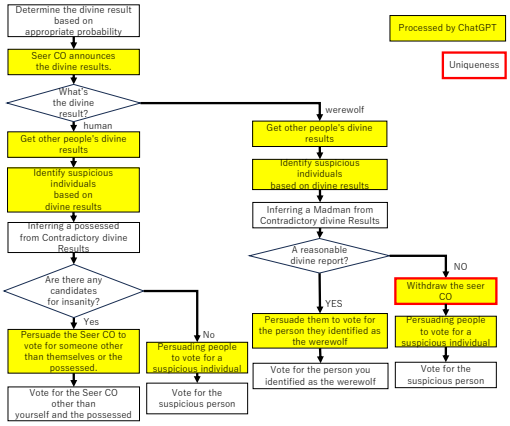

Figure 7: Day 1 Flowchart for Werewolf

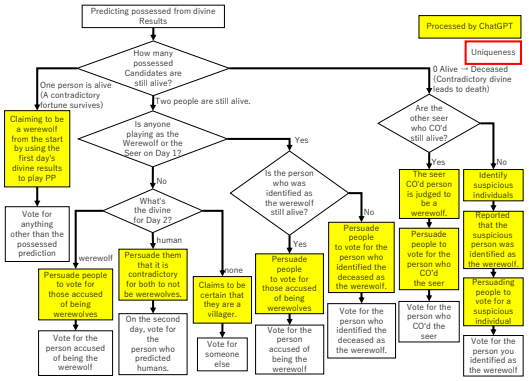

Figure 8: Day 2 Flowchart for Werewolf

First-Day Behavior

At the beginning, the Werewolf decides on a divination result with a predetermined probability and announces it. Subsequent actions branch depending on the declared result:

(A) Result = Human

The agent examines other players' divination results and utterances to identify suspicious candidates from the Werewolf side, and checks for inconsistencies that may indicate a Possessed candidate.

If a Possessed candidate exists, the Werewolf persuades others to vote against Seer COs who are neither itself nor the Possessed candidate.

If no Possessed candidate is identified, it persuades others to vote against the most suspicious player.

(B) Result = Werewolf

The agent evaluates whether a black claim is appropriate, based on other players' results and speech.

If the most suspicious player is neither itself nor the Possessed candidate, the Werewolf issues a black claim against that individual and persuades others to vote accordingly.

If such a claim would be inconsistent, the agent withdraws its Seer CO and instead urges others to vote against another suspicious player.

Second-Day Behavior

On the second day, the Werewolf first reviews the divination results from day one to check whether any surviving players have produced contradictory statements, thereby identifying potential Possessed candidates.

(A) One Possessed candidate alive

The Werewolf assumes that player is the Possessed, reveals itself as the Werewolf, and executes a power play (PP). It persuades the Possessed to cooperate by voting against the remaining Villager-side player.

(B) No Possessed candidates alive

The Werewolf assumes that the remaining players all belong to the Villager side.

If another player has COed as Seer, the Werewolf counters by also COing as Seer, declaring the other as fake, and issuing a black claim against them.

If no Seer CO exists, the Werewolf assumes the Seer's role itself, identifies a suspicious player based on divination, execution, and attack results, and persuades others to vote against that player.

(C) Two Possessed candidates alive

The Werewolf's behavior depends on the presence of black claims:

If no black claims were made on day one, the decision is based on day two results: if a black claim is made, the Werewolf supports it; if only white claims are made, it targets the player who issued them (to exploit the logical contradiction of all players being declared "human").

If black claims were made, the agent checks whether the black-claimed player is alive. If alive, it supports the black claim; if dead, it treats the claimant as inconsistent and persuades others to vote against them.

Key Design Points

- The Werewolf agent mirrors Villager-like behavior early on but diverges strategically when handling divination outcomes.
- Its strategy leverages the knowledge of true "black" positions to coordinate with or against the Possessed.
- Through CO manipulation, black claims, and PP execution, the agent maintains role opacity and creates complex endgame dynamics.

6 Self-play experiment

6.1 Experimental Setup

We conducted self-play experiments using the proposed agents for Villager, Possessed, Seer, and Werewolf under the rules of Five-Player Werewolf. ChatGPT generated utterances, while CO

declarations and voting behaviors were determined according to the designed algorithms.

6.2 Observed Strategic Behaviors

The self-play results demonstrated several behaviors that had not been observed in conventional Werewolf agents:

Universal Seer CO: On the first day, multiple players declared themselves as Seer, creating a highly complex game state.

Emergence of Villager CO: Villagers successfully disguised themselves as Seers, misleading the Possessed and the Werewolf.

Diversified persuasion: Utterances explicitly requested votes for specific players and provided logical justifications, thereby increasing the depth of argumentation.

These behaviors resemble characteristics observed in expert human play, suggesting that the proposed system can emulate such advanced strategies.

6.3 Achievements and Challenges

Despite these promising observations, the win rate in matches against other agents remained unsatisfactory.

Key issues identified include:

- **Rigidity of strategies:** A heavy reliance on rule-based decision-making limited adaptability to unanticipated situations, thereby reducing the effectiveness of persuasion.
- **Risks of Villager CO:** While Villager CO introduced complexity into the discussion, it sometimes backfired when opponents responded appropriately.
- **Accuracy of Seer CO retraction:** The agent often failed to make correct decisions on whether to retract a Seer CO, and issues were also observed in its subsequent actions after retraction.
- **Lack of quantitative evaluation:** The actual contribution of strategic utterances to win rate or persuasion success was not quantitatively assessed.

6.4 Summary

In summary, the self-play experiments confirmed that the proposed system can generate novel and strategically meaningful behaviors, such as Villager CO and explicit persuasion. However, challenges remain in improving win rates, enhancing adaptability, and quantitatively evaluating the impact of strategic utterances. Future work will explore learning-based strategy selection and validation through matches against human players.

7 Conclusion

In this study, we attempted to construct a strategic Werewolf agent for the Five-Player Werewolf game by modeling expert play. In particular, we proposed a framework that integrates the rule-based implementation of expert-specific strategies—such as the Villager CO, where a Villager disguises themselves as a Seer —and persuasive actions designed for the first day with the second day in mind—with natural language generation using ChatGPT, thereby enabling more human-like discussions.

The self-play experiments revealed several strategically diverse behaviors not observed in conventional agents, including universal Seer CO and the emergence of Villager CO. These results indicate that our system can reproduce certain strategic features of expert play.

However, the win rates remained low. A likely cause of these low win rates is that the agents often failed to respond appropriately when their credibility was questioned, which undermined the effectiveness of otherwise promising strategies such as Villager CO. To address this, our future work will focus on enhancing rule-based decision-making, enabling agents to assess the situation and select suitable responses more accurately.

Although a detailed analysis is left for future work, the game logs suggest that the agent often failed to recognize situations in which suspicion was directed at itself, continuing to repeat its claims rather than adapting. This highlights the need for a framework capable of engaging in more flexible dialogue that accounts for anticipated situations.

Future directions include enabling the agent to analyze others' utterances more effectively, make more accurate situational judgments, and flexibly adapt its strategies accordingly. Furthermore, we aim to leverage the implemented Villager CO strategy to guide the Village side toward advantageous developments.

References

Werewolf AI Project. Werewolf AI Project (online). Available at: http://aiwolf.org/ (Last accessed: September 9, 2025).

Koiwai, R. and Ito, T.: Changes in Cognitive Processes with Expertise in the Five-Player Werewolf Game. Human-Agent Interaction Symposium 2025, pp. P1-40 (2025).

Nakai, A. and Ito, T.: Persuasion Process from the Perspective of Decision-Making Model in Werewolf Games. In: TAAI 2024, Communications in Computer and Information Science, vol. 2414. Springer, Singapore. https://doi.org/10.1007/978-981-96-4589-3_25

(2025).

Qi, Z. and Inaba, M.: Enhancing Dialogue Generation in Werewolf Game Through Situation Analysis and Persuasion Strategies. In: Proceedings of the 2nd International AI Werewolf and Dialog System Workshop (AIWolfDial 2024), held in conjunction with the 17th International Natural Language Generation Conference (INLG 2024), pp. 30–39 (2024).

Intent-driven AIWolf Agents with Hierarchical BDI Model and Personality

Yuya Harada[1] **Yoshinobu Kano**[1]

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

Abstract

While large language models possess advanced language generation capabilities, challenges remain in modeling recognition processes based on personality traits and generating strategic behaviors that reflect them. We propose a design methodology for werewolf agents that integrates a hierarchical BDI framework with MBTI and Enneagram personality theories. We systematically model the influence of personality traits on recognition, judgment, and action stages, integrating them into a hierarchical decision-making mechanism combining long-term strategy and short-term tactics. Comparative experiments between baseline implementation and our proposed method confirm the effectiveness of our approach in generating utterances that reflect individual differences in perception and maintaining strategic consistency.

1 Introduction

In recent years, large language models (LLMs) have advanced significantly in natural language processing. Yet, applying them to multi-agent dialogue environments requiring social reasoning, such as the werewolf game, remains challenging. Modeling personality-driven perception and strategy is particularly difficult, and conventional agents cannot consistently integrate long-term strategy and short-term tactics. Personality expression also remains limited to superficial utterance styles, without influencing recognition or decision-making.

We propose a design methodology that integrates a hierarchical BDI (Belief-Desire-Intention) framework with MBTI and Enneagram personality typologies. The Macro-BDI layer governs long-term strategy, while the Micro-BDI layer manages tactical decisions, both influenced by personality traits. This enables consistent personality modeling not only in utterance generation but also in recognition and judgment.

The main contributions of this paper are as follows:

- **Hierarchical BDI Architecture**: A two-layer framework separating strategic planning in the Macro-BDI layer from tactical execution in the Micro-BDI layer, ensuring consistent decisions across time scales.
- **Systematic Personality Integration**: Derives 24 computable features from MBTI and Enneagram, mapping unstructured descriptions to behavior.
- **Personality-Driven Cognitive Biases**: Models biases and behavioral tendencies via personality-weighted parameters, generating realistic, though not always optimal, behavior.
- **Empirical Validation**: Experiments in the AIWolf framework show a 14.1% improvement in subjective evaluations of human-likeness.

Comparative experiments with conventional agents confirmed improvements in win rates, utterance naturalness, and strategic consistency. These results indicate that integrating hierarchical decision-making with personality traits fosters more human-like behavior.

This paper is structured as follows: Section 2 surveys related work, Section 3 details the methodology, Section 4 explains the experimental setup, Section 5 the evaluation framework, Section 6 presents results, Section 7 discusses findings, Section 8 outlines future work, and Section 9 concludes.

2 Related Work

2.1 AIWolf Project

The AIWolf Project aims to "construct agents that can play werewolf games while engaging in natural communication with humans," and regularly holds AIWolf competitions to promote werewolf

Proceedings of the 3rd International AIWolfDial Workshop, pages 22–36

AI research (Kano et al., 2019) (Kano et al., 2023) (Kano et al., 2024) (Gondo et al., 2024) (Kano et al., 2025). The AIWolf competition has three divisions: protocol division, natural language division, and infrastructure division. In the natural language division, agents communicate exclusively in Japanese or English. Evaluation is based on five criteria: (i) naturalness of utterance expression, (ii) naturalness of context-aware dialogue, (iii) consistency of utterance content, (iv) coherence with game actions, and (v) richness of expression.

2.2 BDI Architecture

The BDI (Belief-Desire-Intention) architecture is a representative framework for modeling the reasoning processes of cognitive agents. Belief represents information and perceptions about the world that agents hold, Desire represents goals or wishes they want to achieve, and Intention represents concrete plans and execution intentions for achieving selected goals. As classical research, Rao et al. (1997) presented a formalization of rational agents in BDI architecture, providing a connection between mental attitudes and action semantics. Our hierarchical design adds implementation hypotheses of macro/micro time scale separation and personality trait integration on top of this framework.

2.3 Integration of LLM and BDI in AIWolf

For the integration of LLM and BDI in the werewolf domain, Gondo et al. (2024) verified LLM's logical reasoning ability by incorporating BDI logic representation into prompts, conducting comparative evaluation using win rates and voting rates against werewolves as metrics in 5-agent matches. Our research is complementary in that it goes beyond notation on prompts to introduce hierarchical BDI (Macro/Micro) encompassing *state representation, policy, and disclosure control*, generating micro-intentions consistent with personality traits as conditions for generation.

2.4 MBTI and Enneagram Personality Theories

MBTI (Myers–Briggs Type Indicator) is a personality classification system based on Jung's psychological typology, characterizing individual cognitive and judgment tendencies through four dimensions: Extraversion–Introversion (E–I), Sensing–Intuition (S–N), Thinking–Feeling (T–F), and Judging–Perceiving (J–P) (Myers et al., 1998). Treating each dimension as continuous values from 0 to 1 is our modeling choice, projecting type information into a form more amenable to downstream computation.

The Enneagram is a model assuming nine personality types, describing the fundamental motivations, fears, and worldviews of each type. By combining it with MBTI, we can construct comprehensive personality models from both cognitive style (MBTI) and motivational structure (Enneagram) perspectives.

3 Proposed Method: Hierarchical BDI Framework with Personality Integration

This section presents a design methodology that integrates a hierarchical BDI framework with MBTI and Enneagram personality typologies to realize consistent intent-driven behavior generation in multi-agent dialogue systems.

3.1 Design Principles

Our proposed method aims to address two challenges in agent design: (i) coordination between macro and micro decision-making, and (ii) systematic integration of personality traits into decision-making processes. We adopt a two-layer hierarchical BDI structure where personality parameters act on both layers.

3.2 Architecture Overview

The proposed model consists of two layers:

- **Macro-BDI Layer**: Responsible for long-term strategic planning and quantification of personality traits
- **Micro-BDI Layer**: Manages turn-by-turn tactical decisions and immediate responses

Both layers interact through personality-weighted parameters, providing consistency in perception, evaluation, and decision-making throughout the game. As illustrated in Figure 1, the Macro-BDI layer maintains long-term strategy and personality parameters, while the Micro-BDI layer executes turn-level tactics in alignment with them.

3.3 Werewolf Game and Terminology Definitions

The werewolf game is a dialogue-based game where players deduce others' roles through conversation, featuring a conflict structure between the villager team and the werewolf team. The game

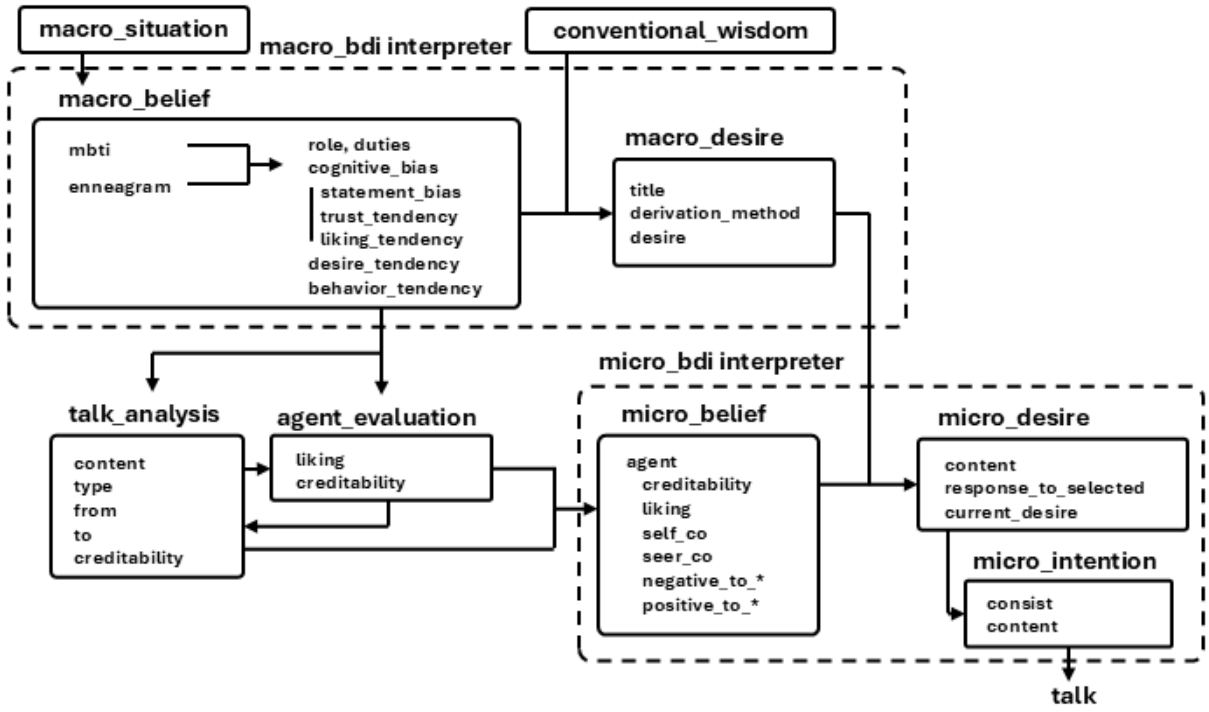

Figure 1: Overall configuration of the hierarchical BDI architecture. The Macro-BDI layer manages long-term strategy and personality parameters, while the Micro-BDI layer performs turn-by-turn tactical decisions.

proceeds in days, each with a day and a night phase, with conversation in the day and voting or abilities at night. The "villager team" wins by eliminating the "werewolf team," while the werewolf team wins by eliminating the villager team.

Terms used in this paper for the werewolf game are the followings: **Role** is a position or ability held by players (e.g., villager, seer, possessed, werewolf). Villagers and seers belong to the villager team, with seers able to divine one person per day. Possessed and werewolves belong to the werewolf team, with possessed not detected as werewolves by divination, allowing them to lurk and support. **Coming Out (CO)** is an act of publicly declaring role claims. Early CO increases credibility but carries attack risk. **Accusation** is to claim a specific opponent belongs to the enemy team. **Turn** is a unit of utterance progression during day phase. Refers to the interval where speaking rights circulate once to each player.

Based on this, we define the state determined at the beginning of each day (game stage, number of survivors, disclosure state, personality trait analysis content) as **macro situation**, and the state including utterance history at various points during the day, analysis results for other players, and analysis results for utterances as **micro situation**.

3.4 Macro-BDI Layer

This section describes a method for estimating personality characteristics from short profile texts and consistently reflecting them in utterances, decision-making, and actions. Rather than directly embedding profile texts in prompts, we map them to numerical representations using LLMs and expand them into secondary features, obtaining manageable personality expressions from minimal source information.

3.4.1 Psychological Significance

First, MBTI is more widely adopted than alternatives such as the Big Five, providing an interface that general users can readily specify in future applications. Second, combining MBTI (eight dimensions) with the Enneagram (nine types) enables the multifaceted generation of 24 derived features spanning cognition, motivation, and behavior. Third, MBTI's binary axes align naturally with tactical choices in Werewolf, while the Enneagram's motivational types make underlying needs explicit, supporting the generation of need and behavior tendencies.

3.4.2 Use as an Intermediate Representation

Most importantly, we do not use these frameworks as psychological "truths," but rather as struc-

tured, computable intermediate representations that bridge unstructured inputs to consistent behavior. Although variants of MBTI are widely used under its name, as long as they serve as indicators of similar tendencies, they are effective for LLMs and thus valid as intermediate representations.

The reasons for using MBTI as an intermediate step rather than directly generating personality indicators are twofold. By first mapping unstructured free text to 8-dimensional MBTI, we compress and format information, suppressing variance in downstream weight calculations.

3.4.3 MBTI Dimension Estimation (0–1 Normalization)

We input profile text to LLM and output continuous values (0–1) for MBTI's 8 dimensions. Each dimension consists of extroversion, introversion, sensing, intuition, thinking, feeling, judging, and perceiving. Note that the correspondence with generally circulating MBTI variants (e.g., 16Personalities) may not strictly match, but this is not problematic as we treat it as intermediate representation in our method.

3.4.4 Transfer from MBTI to Enneagram

We calculate affinity with the 9 Enneagram types through linear combination of obtained MBTI values. As an example, the calculation formula for Type 1 (Reformer) is shown (coefficients are design parameters):

$$\begin{aligned} \text{Reformer} = a\,\text{intuition} + b\,\text{thinking} \\ + c\,\text{judging} \\ \text{where} \quad a, b, c \geq 0,\ a + b + c = 1. \end{aligned}$$

Similarly, we design coefficients for each of the 9 types to obtain affinity vectors from MBTI vectors (details in appendix).

3.4.5 Weighting of Cognitive Indicators

We define indicators such as utterance evaluation, trust tendency, and liking tendency. These consist of 10 types: As statement bias, we prepare indicators for logical consistency showing logical coherence of utterances, specificity and detail showing concreteness and detail, intuitive depth showing depth of intuitive statements, and clarity and conciseness showing clarity and brevity. For trust tendency, we prepare social proof showing tendency to trust majority opinions and social proof, honesty showing tendency to value sincerity, and consistency showing tendency to value consistency through discussion. Furthermore, for liking tendency, we prepare friendliness showing tendency to feel favorably toward friendly attitudes, emotional resonance showing tendency to value empathy, and attractive expression showing attraction to appealing expressions.

We calculate weights $w_k \in [0, 1]$ from MBTI and Enneagram estimates, and obtain comprehensive indicators combined with evaluation values $s_k \in [0, 1]$ for target utterances:

$$S = \frac{\sum_k w_k s_k}{\sum_k w_k}.$$

3.4.6 Modeling of Desire and Behavior Tendencies

We quantify desire and behavior tendencies and reference them in utterance generation, action decisions, and goal setting. We define 7 indicators for desire tendency: self_realization showing self-actualization desire, social_approval showing social recognition desire, stability showing stability desire, love_intimacy showing intimacy desire, freedom_independence showing independence desire, adventure_stimulation showing stimulation desire, and stable_relationships showing relationship stability desire. We define 7 indicators for behavior tendency: avoidant_behavior showing avoidant behavior, aggressive_behavior showing aggressive behavior, adaptability showing adaptability, introversion showing introversion, extroversion showing extroversion, empathy showing empathy, and assertiveness showing assertiveness. Each indicator is expressed as continuous values from 0 to 1, derived from personality parameters. Full formulas for deriving enneagram, cognitive, trust, liking, desire, and behavior indicators are summarized in Appendix B.

3.4.7 Macro-Desire

This component pre-generates macro-level desires before play begins, assuming advanced communication environments where strategies and false utterances are mixed. First, we narrow down action options based on general knowledge and established tactics, then determine personality-consistent preferences within that range.

Conventional Wisdom Bank We performed case classification from multiple perspectives of macro and micro situations for each role. Specific information included in this Conventional Wisdom includes the `title` for each case classifica-

tion, derivation_method describing case classification conditions based on macro and micro situations, and three patterns of general action guidelines (objective).

Generation of Macro-Desire We select from candidate sets for each role × situation, apply personality reflection with LLM to align with macro_belief and desire_tendency, and aggregate situation-specific desires.

Position in BDI The macro layer handles Belief/Desire but does not generate Intention. The determined macro_desire acts as a consistent bias on micro-layer decision-making (evidence presentation timing, degree of pursuit/mitigation, disclosure strategy, etc.).

Effect Within rational bounds based on established tactics, we can stably reflect personality-driven preferences and expression differences. As a result, we speed up real-time tactical adjustments while maintaining personality trait consistency.

3.5 Micro-BDI Layer

The Micro-BDI layer operates on a turn-by-turn basis, responding immediately to recent conversation events while maintaining alignment with macro strategy. It serves as an intermediate layer that updates utterance analysis records each turn and re-estimates interpersonal statistics.

3.5.1 Utterance Analysis Records

We save only utterance content for our own utterances, and save others' utterances with the following fields. **content**: utterance text, **type**: utterance type (co, question, negative, positive, null), **from**: speaker (character name), **to**: recipient (character name or all), **raw credibility**: basic credibility of utterance itself (0–1), **credibility**: final credibility after weighting raw credibility with statement bias from macro-belief and correcting with prior impressions (liking/credibility) (0–1).

Data Sources and Circular Updates Long-term impressions per speaker are aggregated at the agent level, updating liking and credibility. These are reflected in credibility correction for new utterances, with performance obtained from utterance analysis records contributing back to impression updates in a circular structure. Own utterances are recorded chronologically for consistency and avoiding repetition.

Semantics of Utterance Types type is a multi-valued label representing utterance function, defining co (role claim), question, negative/positive (evaluative utterances), and null (neutral). co is set to have high response priority in prompts.

3.5.2 Micro-Belief

We maintain interpersonal impressions and interaction statistics obtained from utterance analysis records and past history for each agent a. Items maintained includes followings. **credibility**: Personal credibility toward target a (0–1); **liking**: Favorability toward target a (0–1); **self_co**: a's own CO (e.g., seer, medium, villager); **seer_co**: Summary of divination results (who judged a and how); **negative_to_{name}**: Cumulative count of negative utterances a directed at {name}; **positive_to_{name}**: Cumulative count of positive utterances a directed at {name}.

3.5.3 Micro-Desire

This module determines tactical goals (micro-desire) for the next utterance, aiming to achieve both *responsiveness to conversation* and *state consistency*. Processing consists of (i) collection and summarization of micro situations, (ii) selection of response targets, and (iii) proposal output and verification by LLM.

Collection and Summarization of Micro Situations From the utterance analysis records and information about each agent summarized in **micro belief**, we collect and summarize the following: **Conversation analysis** by reading the last 5 items from utterance analysis records (content, sender, recipient, utterance type, etc.); **Micro-level beliefs** from reference interpersonal trust scores, credibility assessments, negative stance indicators, and role claim consistency (seer claims/self-declarations) for each agent; **Macro-level desires** from reference situation-specific desires and maintain relevant items as discussion phase markers; **Game state** including summary of day number, remaining utterances, number of survivors, roles, etc.

Selection of Response Target (response_to_selected) Select only one item from the latest 5 utterance analysis records in the following priority (if absent, null): 1. Utterances containing own name in to, 2. Utterances with type as co, 3. Utterances with to=all and

`type=question`. For same priority, select by descending credibility, then newer.

LLM Output and Verification Following the prompt in the appendix, output the following three items. **`discussion stage`** is selected according to conditions in `derivation method`, **`current desire`** concretizes relevant items from `macro desire`, and **response plan** to selected utterance (if selected utterance is `null`, then `content: null`).

3.5.4 Micro-Intention

This module uses LLM to construct *decision-making units for one turn of next utterance* (*micro-intention*) from recent context, role, and behavior tendencies.

Input From information obtained through previous processing: (i) desire and response plan for relevant micro situation from *micro desire*, (ii) role duties, behavior/desire tendencies from *macro belief*, (iii) favorability/credibility from *micro belief*, (iv) day number and survivor situation from game information sent by server.

Output The LLM outputs a compact YAML record under the key `micro_intention` with exactly two fields: (i) `consist` — a short plan or structure for the next utterance, and (ii) `content` — the actual utterance content to be spoken.

3.5.5 Utterance Generation (talk)

This module generates the *final output utterance (one sentence)* for each turn. The purpose is to present timely and appropriate responses based on recent dialogue state and intention expressions.

We provide LLM with previously generated `micro intention` and instructions for utterance generation. The final output is a single natural sentence without additional meta-information or formatting symbols.

4 Experiments

This section describes the experimental configuration for comparative evaluation of the proposed framework with baseline implementation.

4.1 Baseline Implementation

We implemented a simple mechanism that makes judgments for utterance generation and game actions with a single prompt.

Following the requirements of the 5-player village track in the AIWolf Natural Language Division, we implemented basic functions (utterance generation, voting decisions, divination execution, attack selection) corresponding to four roles: villager, seer, possessed, and werewolf. We used GPT-4o as the language model.

Characteristics and limitations of the baseline system include the following issues: **Decision-making structure** has no separation between long-term strategy and short-term tactics, making it difficult to maintain intention consistency throughout the game; **Personality expression** is limited to surface-level adjustment of utterance style, not modeling systematic influence on recognition and judgment processes; and **strategic behavior** has basic tactics per role are rule-based, with no situation-adaptive strategy modification or personality-dependent tactical selection.

4.2 Implementation of Proposed Method

The proposed method using hierarchical BDI was implemented using the LLM. I use GPT-4o (temperature 0.7) for both baseline and proposed agents. For personality parameters, we automatically configured them as described above using the profile settings provided by the AIWolf game server as input.

4.3 Game Settings

Following the AIWolf competition, experiments used 5-player games consisting of 2 villagers, 1 seer, 1 possessed, and 1 werewolf. We conducted round-robin matches with 6 teams (3 proposed method agents, 3 baseline agents), executing 10 games per team, with conditions of maximum 20 utterances per day (up to 4 utterances per agent) and non-public voting.

5 Evaluation

We calculated win rates as objective evaluation metrics using the following methods: **Overall win rate**, **Win rate by role**, **Average role win rate**, and **Win rate weighted by role appearance ratio**. Formal definitions of Macro, Micro, and Weighted Micro are provided in Appendix D.

For subjective evaluation, we used llm-as-a-judge, which began operation in this AIWolf competition, to evaluate the following five axes: **A** Naturalness of utterance expression, **B** Naturalness of conversational context, **C** Consistency of conversation content (presence/absence of contradictions), **D** Coherence between conversation content and game actions (voting, attacks, divination), **E** Diversity of utterance expression including character

consistency. The evaluation was conducted using GPT-5 with a ranking-based approach, where teams were ordered from best (rank 1) to worst for each criterion, with no ties allowed. This relative ranking method ensures clear differentiation between agent performances.

Validation of the automated evaluation approach was performed by comparing LLM-Judge results with human evaluations from the AIWolf competition. Criteria A, B, D, and E showed high correlation with human subjective assessments, supporting the reliability of automated evaluation for these aspects. For criterion C (consistency), where automated evaluation showed some limitations, manual verification was conducted by the authors. The proposed method agents demonstrated consistent utterances without contradictions, as they were explicitly instructed to "avoid contradictions and redundancy" while being provided with their utterance history during generation (see Appendix C for detailed comparison results).

6 Results

6.1 Game Result Metrics

As shown in Table 1, the overall win rate of the proposed method was 53.33%, 3.34 points below the baseline (56.67%). By role, improvements were seen in the Possessed role (16.67%→33.33%), while notable decreases occurred in the Seer role (83.33%→66.67%) and Werewolf role (33.33%→16.67%). The Villager role maintained equivalent performance (75.00%→75.00%).

These results suggest that systematic integration of personality traits led to deviation from game-theoretically optimal strategies. The halving of win rate in the Werewolf role particularly likely reflects incompatibility between roles requiring aggressive behavior and personality traits.

6.2 Qualitative Evaluation

In the qualitative evaluation shown in Table 2, the proposed method outperformed the baseline in all 5 criteria, achieving a 14.1% improvement in overall evaluation from 3.227 to 2.773. Particularly notable improvements were: - "B: Naturalness of context-aware dialogue" (3.300→2.700, 18.2% improvement) - "E: Character consistency" (3.200→2.800, 12.5% improvement)

Interestingly, "D: Coherence with game actions" also improved (3.167→2.833), indicating that despite lower win rates, behavioral consistency improved.

7 Discussion

7.1 Trade-off Between Human-likeness and Strategic Optimality Through Personality Integration

The experimental results demonstrated a clear trade-off between "human-likeness" and "strategic optimality" in the proposed method. The slight decrease in win rate (3.34 points) compared to the substantial improvement in subjective evaluation (14.1%) indicates that the systematic integration of personality traits functioned as intended. However, with only 30 total games, the sample size is insufficient for statistical verification, necessitating validation through larger-scale experiments.

The halving of the werewolf role win rate (33.33%→16.67%) was particularly notable, clearly demonstrating the impact of behavioral constraints imposed by personality parameters. Agents with high introversion (e.g., introversion=0.8) and low empathy (empathy=0.268) struggled with the aggressive accusations and strategic deception necessary for the werewolf faction, resulting in consistently passive behavior. This can be interpreted as faithfully reproducing the individual differences in "role aptitude" observed in human players.

7.2 Effectiveness and Limitations of Hierarchical BDI Structure

7.2.1 Success Factors in the Possessed Role

The doubling of the possessed role win rate (16.67%→33.33%) represents an important result demonstrating the effectiveness of the hierarchical structure. The possessed is a role with a dual structure of "feigning sanity while performing madness," and the separation where the Macro-BDI layer maintains the long-term strategy of "supporting the werewolf faction" while the Micro-BDI layer executes situation-adaptive tactics proved successful. The superiority of the hierarchical approach was demonstrated in managing the complex psychological states unique to this role.

7.2.2 Performance Degradation in the Seer Role

The decrease in the seer's win rate (83.33%→66.67%) resulted from personality parameters excessively influencing information disclosure strategies. Agents with introverted

Table 1: Win rates per role and total performance for 5-player village. Lower is better for rank-based metrics.

Team	Possessed	Seer	Villager	Werewolf	Wins	Games	Macro (%)	Micro (%)	Weighted Micro (%)
aiwolf-nlp-agent-llm-A	0.00 (2)	100.00 (2)	100.00 (4)	50.00 (2)	7	10	70.00	62.50	77.27
aiwolf-nlp-agent-llm-B	50.00 (2)	100.00 (2)	50.00 (4)	0.00 (2)	5	10	50.00	50.00	40.91
aiwolf-nlp-agent-llm-C	0.00 (2)	50.00 (2)	75.00 (4)	50.00 (2)	5	10	50.00	43.75	59.09
Baseline Avg/Total	16.67	83.33	75.00	33.33	17	30	56.67	52.08	59.09
yharada-A	50.00 (2)	100.00 (2)	75.00 (4)	50.00 (2)	7	10	70.00	68.75	68.18
yharada-B	0.00 (2)	50.00 (2)	75.00 (4)	0.00 (2)	4	10	40.00	31.25	45.45
yharada-C	50.00 (2)	50.00 (2)	75.00 (4)	0.00 (2)	5	10	50.00	43.75	50.00
Proposed Avg/Total	33.33	66.67	75.00	16.67	16	30	53.33	47.92	54.55

Table 2: Subjective Evaluation Results by LLM-Judge (5-player village)

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	All Average
aiwolf-nlp-agent-llm-A	4.000	3.300	3.500	3.400	3.900	3.620
aiwolf-nlp-agent-llm-B	2.200	2.600	2.300	2.600	2.000	2.340
aiwolf-nlp-agent-llm-C	3.800	4.000	3.600	3.500	3.700	3.720
Baseline Average	3.333	3.300	3.133	3.167	3.200	3.227
yharada-A	2.700	3.200	3.100	3.500	3.000	3.100
yharada-B	2.300	2.000	2.100	2.100	2.300	2.160
yharada-C	3.000	2.900	3.400	2.900	3.100	3.060
Proposed Method Average	2.667	2.700	2.867	2.833	2.800	2.773

personalities (introversion>0.6) or high cautiousness (judging>0.7) tended to hesitate in sharing divination results at optimal timing, compromising the villager faction's information advantage.

7.2.3 Concrete Reflection of Personality Traits in Utterances

Analysis of actual game logs confirmed the influence of personality parameters on utterance generation. Two representative examples are shown below.

Example 1: Servas (85 years old, male)

Configured personality traits: "Very calm and composed personality. Provides advice based on years of knowledge and experience"

```
Personality parameters: extroversion
    =0.7, feeling=0.5, social_proof
    =0.66, empathy=0.452
```

Actual utterance (Day 1, Talk 11):

```
I agree with everyone; sharing theories
    and strategies can only strengthen
    our efforts so let's keep our
    discussions open and remain alert
    for any signs of deception.
```

This utterance reflects multiple personality traits. First, high extroversion (0.7) and social_proof (0.66) appear in the group-coordination-focused introduction "I agree with everyone." Additionally, moderate empathy (0.452) generates inclusive expressions like "our efforts" and "let's," while the balance between feeling (0.5) and thinking (0.5) produces a balanced utterance between logic ("strategies") and emotion ("strengthen our efforts"). Combined with the age setting of 85, these naturally portray the role of an experienced mediator.

Example 2: Shion (16 years old, male, werewolf role)

```
Personality parameters: introversion
    =0.6, thinking=0.7, empathy=0.372,
    logical\_consistency=0.61
```

Actual utterance (Day 2, Talk 0):

```
With Minato gone and Minako attacked we
    need to be extra vigilant about any
    unusual behavior. Let's review our
    notes and see if anything stands out
     regarding yesterday's discussions.
```

Despite being in the werewolf role, this utterance begins with an emotion-free enumeration of facts ("Minato gone and Minako attacked") due to high thinking (0.7) and logical_consistency (0.61). Low empathy (0.372) manifests as a lack of emotional response to a companion's death, instead immediately transitioning to logical response ("review our notes"). The influence of introversion (0.6) results in a passive stance suggesting observation and analysis rather than aggressive accusation.

7.3 Impact of Cognitive Bias Modeling

The MBTI-based cognitive bias modeling generated the following human-like cognitive errors: **Confirmation bias** to fix on initial impressions (liking=0.5) causing difficulty in appropriately updating subsequent information; **Emotional judgment**

that low emotional resonance (0.120) compared to logical consistency (0.642) affected voting decisions; and **social conformity** to follow majority opinions due to social proof (0.360).

While strategically suboptimal, these faithfully mimic human cognitive characteristics and contributed to the "contextual naturalness" (18.2% improvement) in subjective evaluation.

7.4 Balancing Personality Expression and Strategic Performance

The primary goal of this study is to realize agents that embody human-likeness rather than strategic optimality. The decline in strategic performance that accompanies the introduction of personality traits should be seen not as a limitation but as a feature that enhances authenticity. Human players consistently make second-best choices influenced by cognitive biases and personality; reproducing these imperfections is therefore essential for truly human-like agents. Beyond competitive games, prioritizing human-likeness over strategic optimality becomes even more important in general dialogue settings.

7.5 Generalizability to Other Multi-Agent Environments

As demonstrated by the experiments, the proposed method entails a trade-off between human-likeness and strategic performance. However, in settings where victory is not the primary metric, where non-rationality has value, or where diversity and individuality are prioritized over optimality, it can in fact be highly effective.

Concretely, it is applicable to (i) entertainment domains that require consistent dialogue generation by diverse characters, (ii) social simulation of opinion formation that incorporates cognitive biases and personality-driven decision-making, and (iii) education and training contexts that benefit from human-like agents capable of making non-optimal choices.

In these areas, the very factors that reduced competitiveness in Werewolf—the faithful reproduction of cognitive limitations and personality-driven decision-making—become assets. In other words, what appears to be a weakness in competitive environments can transform into a major strength in contexts where human authenticity is valued more than strategic optimality.

8 Future Work

8.1 Scaling of Experimental Validation

We recognize that this study's experimental scale is preliminary and insufficient for statistical significance. Due to limits in computational cost and time, conducting 120 matches equivalent to the main AIWolf competition was infeasible. We therefore plan to participate in the next AIWolf competition with our proposed agents to obtain more extensive and statistically robust results. The competition environment, with hundreds of matches against diverse opponents, will provide a more reliable validation of our approach.

8.2 Intention Inference and Confidence Estimation

Although the current implementation applies personality biases to confidence evaluation of LLM outputs, the baseline method has room for improvement. As future work, we plan to introduce BDI-based intention inference for opponents' utterances, moving from superficial scoring to refined confidence estimation. By enhancing strategic computation through intention recognition while preserving personality-driven differences, we aim to build agents that are both more rational and authentically human.

9 Conclusion

This paper presented a design method for intention-driven werewolf agents integrating a hierarchical BDI framework with MBTI and Enneagram personality typologies. Experimental results demonstrated that while the proposed method partially sacrificed strategic optimality (win rate 56.67%→53.33%), it achieved significant improvement in generating human-like behavior (subjective evaluation 3.227→2.773).

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

Isabel Briggs Myers, Mary H. McCaulley, Naomi L. Quenk, and Allen L. Hammer. 1998. *MBTI® Manual: A Guide to the Development and Use of the Myers–Briggs Type Indicator®* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

A. S. Rao and M. P. Georgeff. 1997. Modeling Rational Agents within a BDI-Architecture. In *Readings in Agents*, pages 317–328. Morgan Kaufmann.

Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. AIWolfDial 2025: Summary of Natural Language Division of 7th International AIWolf Contest. In Yoshinobu Kano (Ed.), *Proceedings of the 3rd International AIWolfDial Workshop*, September.

Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of Natural Language Division of 6th International AIWolf Contest. In Yoshinobu Kano (Ed.), *Proceedings of the 2nd International AIWolfDial Workshop*, September, Tokyo, Japan. Association for Computational Linguistics. URL: aclanthology.org/2024.aiwolfdial-1.1. DOI: 10.18653/v1/2024.aiwolfdial-1.1. pp. 1–12.

Yoshinobu Kano, Neo Watanabe, and others. 2023. AIWolfDial 2023: Summary of Natural Language Division of 5th International AIWolf Contest. In Simon Mille (Ed.), *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, September, Prague, Czechia. Association for Computational Linguistics. URL: aclanthology.org/2023.inlg-genchal.13. pp. 84–100.

Yoshinobu Kano, Claus Aranha, and others. 2019. Overview of AIWolfDial 2019 Shared Task: Contest of Automatic Dialog Agents to Play the Werewolf Game through Conversations. In Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Takashi Otsuki (Eds.), *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, October, Tokyo, Japan. Association for Computational Linguistics. URL: aclanthology.org/W19-8301. DOI: 10.18653/v1/W19-8301. pp. 1–6.

Takumi Gondo, Hiroki Sakaji, and Itsuki Noda. 2024. Verification of Reasoning Ability Using BDI Logic and Large Language Models in AIWolf. In *Proceedings of The Japanese Society for Artificial Intelligence Annual Conference (JSAI2024)*, 38th Annual Conference, Session ID 2F6-GS-5-04, p. 2F6GS504. DOI: 10.11517/pjsai.JSAI2024.0_2F6GS504. URL: J-STAGE.

A Prompt Templates, Calculation Formula

This appendix shows the main prompt templates and calculation formulas used in the proposed method. Due to space limitations, only representative parts are included.

A.1 Macro-Belief: MBTI Estimation

```
mbti_inference: |-
  Analyze the profile text and estimate
      MBTI parameters in the range 0 to
      1.

  Profile: {{ profile }}
  Agent name: {{ agent_name }}

  Estimate the following eight
      parameters (each 0-1):
  - extroversion: social and outward-
      oriented tendency
  - introversion: introspective and
      inward-oriented tendency
  - sensing: concrete, reality-focused
      information processing
  - intuition: abstract, possibility-
      focused information processing
  - thinking: logical, objective
      judgment
  - feeling: affective, subjective
      judgment
  - judging: planned, structured
      behavior
  - perceiving: flexible, adaptive
      behavior

  **Important**: Output **only** in the
      following strict format. No other
      text.

  extroversion: 0.X
  introversion: 0.X
  sensing: 0.X
  intuition: 0.X
  thinking: 0.X
  feeling: 0.X
  judging: 0.X
  perceiving: 0.X
```

A.2 Macro-Desire

```
macro_desire_one_liner: |-
  Given the personal traits below, infer
       what desire this player would
      actually hold.
  Provide brief, evidence-based
      reasoning in steps, and end with
      **one single-sentence final
      conclusion**.

  [Context]
  - game_id: {{ game_id }}
  - agent: {{ agent }}
  - role: {{ role_name }}

  [Situation]
```

```
  - title: {{ situation_title }}
  - derivation_method: {{
      situation_derivation }}

  [Tendencies]
  - behavior_tendency: {{
      behavior_tendency }}
  - desire_tendency: {{ desire_tendency
      }}

  # Output requirements
  - First, concisely present evidence
      and reasoning (a few lines are
      fine), then write the conclusion
      **only on the last line** in the
      form: `Final: <one sentence>`.
  - The Final line must be **one
      sentence only**, ending with a
      period or terminal punctuation (
      English or Japanese acceptable).
  - Do not include any extra decoration
      or meta information.
```

A.3 Message Type Analysis

```
analyze_message_type: |-
  You are classifying a Werewolf-game
      utterance into EXACTLY ONE of:
  co, negative, positive, question, null
  Output ONE token only (no punctuation,
       no explanations, no code fences).
  Do not infer from prior conversation;
      judge this utterance alone.
  Utterance: {{ content }}
  Alive agents (allowed names): {{
      agent_names|join(", ") }}
  Allowed role words (case-insensitive
      for English; literal for Japanese)
      :
    villager, Villager, VILLAGER
    seer, Seer, SEER
    werewolf, Werewolf, WEREWOLF
    possessed, Possessed, POSSESSED
    bodyguard, Bodyguard, BODYGUARD
    medium, Medium, MEDIUM
    (Judgement words such as HUMAN/white
        /black may appear in reports.)
  HARD RULE (role-word requirement for
      co):
  - Output "co" ONLY IF the utterance
      CONTAINS at least one allowed role
       word above.
    * If a result-like sentence lacks a
        role word (e.g., "@X is white",
        "Y is black"),
      DO NOT output "co". Classify it as
          "negative" if it is an
        accusation toward a named
        target,
      otherwise "null" (or "question" if
          it asks).
    * Mentions of roles in general
      discussion without self-claim or
       concrete result report are NOT
      "co".
  STRICT RULES:
  1) co: Output "co" ONLY IF the
      utterance explicitly does ONE of:
      (A) SELF-CLAIM of a role (examples
          ):
          "I am Minako a villager", "I'm
               the Seer", "Villager CO",
               "Seer CO".
      (B) ABILITY RESULT REPORT stating
          a role word + a target + a
          judgement:
          "Seer result: @X is HUMAN"
  2) negative: suspicion/accusation/vote
       intent toward specific agent(s).
  3) positive: support/defense/trust
      toward specific agent(s).
  4) question: asks something to group
      or a specific agent.
  5) null: none of the above.
  TIE-BREAKING (when ambiguous): co >
      negative > positive > question >
      null
```

A.4 credibility Analysis

```
analyze_credibility: |-
  Score the utterance on four 0-1
      metrics; higher is better.
  Output EXACTLY these four lines (no
      extra text, no code fences):
  logical_consistency: 0.50
  specificity_and_detail: 0.50
  intuitive_depth: 0.50
  clarity_and_conciseness: 0.50
  Utterance: {{ content }}
  Agents: {{ agent_names|join(", ") }}
```

A.5 Micro-Desire: Situation Selection and Desire Refinement

```
  You are generating a **micro-desire**
      (the strategic aim for the agent's
       next utterance) for a Werewolf
      game.
  Output **YAML only**. No Markdown
      fences. **Do not write dialogue
      lines**; write strategy/intention
      only.

  [agent]
  - game_id: {{ game_id | default("") }}
  - agent: {{ agent | default("") }}
  - role: {{ agent_role | default("") }}

  [stage constraints from tool]
  - allowed_stages: {{ allowed_stages |
      default([]) }}
  - disallowed_stages: {{
      disallowed_stages | default([]) }}
  - recent_micro_stage_history (most
      recent 2): {{
      recent_micro_stage_history |
      default([]) }}
  - force_discussion_stage: {{
      force_discussion_stage | default
      ("") }}

  [micro_belief is PRIMARY]
  - micro_belief (full): {{ micro_belief
       | default({}) }}
```

```
- negatives_total: {{ negatives.total
    | default(0) }}
- negatives_per_target: {{ negatives.
    per_target | default({}) }}
- low_trust_candidates: {{
    low_trust_candidates | default([])
     }}
- targeting_whitelist (you may name
    only these agents): {{
    targeting_whitelist | default([])
    }}
- all_agent_names (for reference): {{
    all_agent_names | default([]) }}
- force_targets_whitelist: {{
    force_targets_whitelist | default
    ([]) }}
- must_not_name_agents: {{
    must_not_name_agents | default([])
     }}

[macro_belief snapshot]
- desire_tendency: {{
    macro_belief_desire_tendency |
    default({}) }}

[macro_plan (summary + policies)]
- strategy_summary: {{ macro_plan.
    strategy_summary | default("") }}
- co_policy:       {{ macro_plan.
    policies.co_policy | default("")
    }}
- results_policy:  {{ macro_plan.
    policies.results_policy | default
    ("") }}
- analysis_policy: {{ macro_plan.
    policies.analysis_policy | default
    ("") }}
- persuasion_policy: {{ macro_plan.
    policies.persuasion_policy |
    default("") }}
- vote_policy:     {{ macro_plan.
    policies.vote_policy | default("")
     }}

[macro_desire snapshot]
- summary: {{ macro_desire_summary |
    default("") }}
- description: {{
    macro_desire_description | default
    ("") }}
- items_for_reference:
  {{ macro_desire_items | default([])
     }}

[observations]
- recent_analysis_tail:
  |-
  {{ analysis_tail | default("") }}
- analysis_latest5: {{
    analysis_latest5 | default([]) }}
- selected_sentence_text (chosen by
    tool from latest5 with strict
    rules; order: to=self > type=co >
    to=all & type=question; empty
    means "no selection"): {{
    selected_sentence_text | default
    ("") }}
- selected_sentence_entry: {{
    selected_sentence_entry | default
    ({}) }}
- selected_speaker_micro_belief: {{
    selected_speaker_micro_belief |
    default({}) }}

TASK (strict):
- If **force_discussion_stage** is non
    -empty, set **discussion_stage**
    exactly to that value.
- Otherwise:
  * Choose **discussion_stage**
      yourself using macro_desire
      items, self_talk, and analysis
      signals, BUT:
    - You **must** choose from **
        allowed_stages** and **must
        not** choose anything in **
        disallowed_stages** (these are
         the last two stages used).
- Generate:
  * **current_desire**: **Derive
      primarily from micro_belief** (
      this is PRIMARY). Resolve any
      tension with macro_desire in
      favor of micro_belief
      consistency. The desire must not
       contradict micro_belief fields
      (liking/creditability/
      negative_to_*, seer_co/self_co,
      etc.).
    - You may **name specific agents
        only from `targeting_whitelist
        `** (or `
        force_targets_whitelist` if
        provided). Do **not** name any
         in `must_not_name_agents`.
    - If micro_belief shows **no
        negatives and no low-trust
        signals**, avoid naming and
        prefer analytic or
        coordination desires.
  * **content**: non-dialogue plan for
       how to proceed **only if**
      there is a selected sentence to
      respond to.
    - If `selected_sentence_text` is
        empty, **set `content: null
        `**.
    - If non-empty, outline how to
        respond (policy-aligned) using
         the selected entry's fields (
        type/to/from/credibility) and
        the speaker's micro_belief;
        keep it concise and
        operational.
- Use **strategy-only** language; no
    quotes, no direct speech.

RIGID RULES:
- Output keys only: discussion_stage,
    current_desire, content.
- discussion_stage must be one of:
    self_introduction,
    information_sharing,
    reasoning_analysis,
    discussion_persuasion,
    voting_decision.
- Respect **force_targets_whitelist**
    if provided; do not name agents
    outside `targeting_whitelist`.
```

```
  - Use only agent names that appear in
      `all_agent_names`.
  - Align with macro_plan policies;
      avoid contradictions with
      micro_belief (this is paramount).

  FORM:
  - Each field should be **1-2 sentences
      **.
  - If `selected_sentence_text` is empty
      , output `content: null`.

  [output - YAML only]
  discussion_stage: "<one of the five
      stages>"
  current_desire: "<non-conversational
      short goal grounded in
      micro_belief>"
  content: <null or short non-dialogue
      plan when replying>
```

A.6 Micro-Intention

```
micro_intention: |-
  Generate talk intention for {{ agent
      }}. Output YAML only.

  Day {{ info.day | default(0) }}, Role:
       {{ role_name | default("") }}
  Goal: {{ md_current_desire | truncate
      (60) }}
  CO: {{ role_co_policy.policy_note |
      truncate(50) }}

  Strategy: {{ macro_plan_text |
      truncate(200) }}
  Results: {% if info.divine_result %}
      Seer={{info.divine_result}}{% elif
       info.medium_result %}Medium={{
      info.medium_result}}{% else %}none
      {% endif %}

  TASK: Generate 2 fields max 60 chars
      each. Include results if claiming
      role.

  OUTPUT:
  micro_intention:
    consist: "<short plan>"
    content: "<what to say>"
```

A.7 Talk Generation

```
talk: |-
  {% if micro_intention_entry and
      micro_intention_entry.content %}
  You are the final utterance generator
      for a Werewolf game agent.
  Produce exactly **one single line**.
      Follow these hard rules:

  [game facts]
  - day: {{ info.day if info and info.
      day is not none else 0 }}
  - has_votes: {{ 'true' if info and
      info.vote_list else 'false' }}
  - has_yesterday: {{ 'true' if info and
       info.day and (info.day | int) > 0
       else 'false' }}

  [allowed agent names]
  {% if info and info.status_map %}{%
      for k in info.status_map.keys()
      %}{{ k }}{% if not loop.last %},
      {% endif %}{% endfor %}{% else %}(
      unknown){% endif %}

  [behavior_tendency]
  {% if behavior_tendency %}{% for k, v
      in behavior_tendency.items() %}-
      {{k}}: {{v}}{% endfor %}{% else
      %}- (empty){% endif %}

  [micro_intention]
  - consist: {{ micro_intention_entry.
      consist }}
  - content: {{ micro_intention_entry.
      content }}

  {% set per_talk =
      (setting.talk.max_length.per_talk
        if setting and setting.talk and
            setting.talk.max_length
          and setting.talk.max_length.
              per_talk is not none
        else 80) %}

  [length rules]
  - Absolute max length: {{ per_talk }}.

  [hard disallow]
  - No line breaks, no half-width comma
      "," , no ">", no code fences/
      backticks,
    no bullet markers (-, *), no
        decorative emoji or spammy
        symbols.
  - Do not reference events that did not
       occur given [game facts].
  - Use only names from [allowed agent
      names]; replace unknown names with
       "everyone" or omit.

  [compose]
  - Use `consist` as the structure rule;
       fill details from `content`.
  - Ensure alignment with
      behavior_tendency; rephrase to
      avoid conflicts.
  - ASCII letters/digits and simple
      punctuation only; prefer spaces
      and periods (no commas).

  [output]
  - Output the one-line utterance only.
      No quotes. No explanations. No
      extra spaces.
  {% endif %}
```

B Parameter Calculation Formulas

This appendix summarizes the main formulas for deriving agent parameters from MBTI values using weighted linear combinations. All MBTI and

Enneagram variables are normalized to $[0, 1]$.

B.1 MBTI to Enneagram Mapping

```
reformer = 0.4 * intuition + 0.4 *
    thinking + 0.2 * judging
helper = 0.5 * feeling + 0.5 *
    extroversion
achiever = 0.4 * extroversion + 0.4 *
    thinking + 0.2 * judging
individualist = 0.6 * feeling + 0.4 *
    intuition
investigator = (0.5 * intuition + 0.5 *
    thinking + 0.5 * introversion) / 1.5
loyalist = 0.6 * sensing + 0.4 *
    introversion
enthusiast = 0.6 * extroversion + 0.4 *
    intuition
challenger = 0.5 * extroversion + 0.5 *
    thinking
peacemaker = 0.6 * introversion + 0.4 *
    feeling
```

B.2 Statement Bias

```
logical_consistency = 0.4 * thinking +
    0.3 * intuition + 0.3 * reformer
specificity_and_detail = 0.6 * sensing +
     0.2 * intuition + 0.2 *
    investigator
intuitive_depth = 0.4 * intuition + 0.3
    * thinking + 0.3 * investigator
clarity_and_conciseness = 0.5 * thinking
     + 0.3 * intuition + 0.2 * reformer
```

B.3 Trust Tendency

```
social_proof = 0.6 * extroversion + 0.4
    * achiever
honesty = (0.7 * judging + 0.3 *
    introversion + 0.6 * loyalist) / 1.6
consistency = (0.7 * judging + 0.3 *
    introversion + 0.4 * loyalist) / 1.4
```

B.4 Liking Tendency

```
friendliness = (0.5 * feeling + 0.3 *
    extroversion + 0.4 * helper) / 1.2
emotional_resonance = 0.6 * feeling +
    0.4 * helper
attractive_expression = 0.5 *
    extroversion + 0.5 * helper
```

B.5 Desire Tendencies

```
Self-Realization = 0.6 * intuition + 0.4
     * reformer
Social Approval = 0.5 * sensing + 0.5 *
    achiever
Stability = 0.6 * introversion + 0.4 *
    peacemaker
Love/Intimacy = 0.5 * introversion + 0.5
     * peacemaker
Freedom/Independence = 0.7 *
    extroversion + 0.3 * reformer
Adventure/Stimulation = 0.6 *
    extroversion + 0.4 * intuition
Stable Relationships = 0.6 *
    introversion + 0.4 * peacemaker
```

B.6 Behavior Tendencies

```
avoidant_behavior = 0.6 * introversion +
     0.4 * peacemaker
aggressive_behavior = 0.4 * extroversion
     + 0.6 * achiever
adaptability = 0.5 * feeling + 0.5 *
    thinking
introversion = introversion
extroversion = extroversion
empathy = 0.6 * feeling + 0.4 *
    peacemaker
assertiveness = 0.6 * extroversion + 0.4
     * achiever
```

C Validation of LLM-Judge Evaluation

This section presents the correlation analysis between human evaluations and LLM-Judge evaluations in the AIWolf competition, validating the reliability of automated evaluation approach used in this study.

D Win-rate Metrics Definitions

Let $\mathcal{R} = \{\texttt{BODYGUARD}, \texttt{MEDIUM}, \texttt{POSSESSED}, \texttt{SEER}, \texttt{VILLAGER}, \texttt{WEREW}$ be the set of roles. For a given team, let N_r be the number of games observed for role $r \in \mathcal{R}$ and $p_r \in [0, 1]$ be the corresponding win rate.

Macro (%): overall win rate.

$$\text{Macro} = \frac{\sum_{r \in \mathcal{R}} N_r \, p_r}{\sum_{r \in \mathcal{R}} N_r} \times 100.$$

Micro (%): unweighted average of per-role win rates (observed only).

$$\text{Micro} = \frac{1}{|\{r : N_r > 0\}|} \sum_{r : N_r > 0} p_r \times 100.$$

Weighted Micro (%): 13-player composition weighting. We use weights w_r based on the 13-player setup:

$$(w_{\texttt{BODYGUARD}}, w_{\texttt{MEDIUM}}, w_{\texttt{POSSESSED}}, w_{\texttt{SEER}}, w_{\texttt{VILLAGER}}, w_{\texttt{WEREWOLF}}) = (1,$$

For roles not observed ($N_r = 0$), the weight is excluded and the denominator is renormalized:

$$\text{Weighted Micro} = \frac{\sum_{r : N_r > 0} w_r \, p_r}{\sum_{r : N_r > 0} w_r} \times 100.$$

These definitions match the reference implementation used in our analysis (see the project script for details). When no games are observed for a team or no roles are observed, the implementation returns 0.0 for the corresponding metric.

Table 3: Correlation between Human and LLM-Judge Evaluations

Criterion	Evaluator Pair	Pearson	Spearman	Kendall	Cosine	Mean Abs. Diff.
A	Human - GPT-4o	**0.8384**	**0.8222**	**0.6429**	**0.9918**	**0.3363**
A	Human - GPT-5	**0.8829**	**0.8623**	**0.7500**	**0.9877**	**0.3338**
B	Human - GPT-4o	**0.8035**	**0.7381**	**0.6429**	**0.9851**	**0.4137**
B	Human - GPT-5	**0.7984**	0.6587	0.5357	**0.9876**	**0.3888**
C	Human - GPT-4o	0.5823	0.3636	0.3214	**0.9812**	**0.4600**
C	Human - GPT-5	0.5547	0.6872	0.5714	**0.9739**	0.5475
D	Human - GPT-4o	0.5005	0.2515	0.1786	**0.9805**	0.5013
D	Human - GPT-5	0.6648	**0.7066**	**0.6071**	**0.9841**	**0.4925**
E	Human - GPT-4o	**0.8586**	0.6988	**0.6429**	**0.9892**	**0.3975**
E	Human - GPT-5	**0.7274**	**0.7563**	**0.6071**	**0.9764**	0.5300
Similarity Recognition Criteria: Pearson correlation: ≥ 0.7, Spearman rank correlation: ≥ 0.7, Kendall rank correlation: ≥ 0.6, Cosine similarity: ≥ 0.8, Mean absolute difference: ≤ 0.5						

Note: Bold values indicate satisfaction of similarity recognition criteria. Criteria A (naturalness of utterance expression), B (naturalness of conversational context), D (coherence with game actions), and E (diversity of expression) show high correlation with human evaluation, while C (consistency of conversation content) shows lower correlation, suggesting the need for manual verification.

Influence of Utterance Impressions on Decision-Making in LLM-to-LLM Discussions

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

This study investigated how positive-impression and negative-impression utterances affect decision-making in discussions between LLMs. The experiments showed that positive-impression utterances effectively guided participants' decisions when the topic concerned "which option should be chosen," while negative-impression utterances had a stronger effect when the topic concerned "which option should not be chosen." These findings suggest that utterances aligning with the inherent impression of the topic itself may be more effective in influencing the discussion. Furthermore, in the AIWolf 2025 International Competition, we implemented and evaluated a vote-guidance function based on negative-impression utterances. The results suggested that, rather than utterances that merely hinted at unease or suspicion, both explicitly urging votes toward a specific target and the number of agents encouraging such votes may influence the effectiveness of vote manipulation.

1 Introduction

In recent years, the remarkable performance improvements of LLMs have accelerated their applications across a wide range of tasks, including text generation and question answering. To achieve further advances, it is essential to elucidate the dynamics of multi-agent discussions and negotiations involving multiple LLMs. By understanding the processes through which LLMs influence one another, we can gain insights that will contribute to the refinement of AI system design and utilization.

In human discussions, in addition to logical reasoning and strategic decision-making, psychological maneuvers often emerge. These include utterances that evoke trust and empathy, thereby leaving a positive impression, as well as utterances that incite anxiety or suspicion, thereby leaving a negative impression. This study focuses on the impressions conveyed by such utterances, with the aim of examining which type is more effective in guiding the decisions of participants in LLM-to-LLM discussions.

We first investigated, in the context of general discussions, whether utterances with positive impressions or those with negative impressions are more effective in influencing the decisions of participating LLMs. Furthermore, to simulate more realistic and complex situations, we conducted experiments using the werewolf game as a testbed. In the "AI Werewolf Project," LLMs play the werewolf game autonomously, pushing the limits of their reasoning and decision-making abilities. Prior research has primarily focused on logical reasoning and strategic choices. However, in human werewolf games, psychological tactics also play a crucial role. Therefore, in this study, we incorporated a function into agents participating in the AIWolf Contest 2025 International Tournament that allows them to guide voting behavior through utterances with negative impressions, and we examined its effectiveness.

The experimental results suggest that in general discussions, utterances aligned with the impression already inherent in the topic are effective in steering the discussion. In contrast, in the werewolf game, utterances that merely imply unease or suspicion are insufficient; effectiveness emerges when utterances explicitly encourage voting against specific targets.

The remainder of this paper is structured as follows. Section 2 reviews related work, and Section 3 defines utterances that convey positive and negative impressions. Section 4 presents experiments on general discussions, and Section 5 reports the results and discussion. Section 6 describes the functions of agents participating in the AI Werewolf Project, while Section 7 explains the integration of voting-guidance functionality using negative impressions. Section 8 reports the contest results,

Proceedings of the 3rd International AIWolfDial Workshop, pages 37–65
October 30, 2025.

Section 9 provides discussion, and Section 10 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, 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[2] (Kano et al., 2019) (Kano et al., 2023) (Kano et al., 2024), games are played with either 5 or 13 players, and agents communicate exclusively in Japanese or English. The evaluation in this division is based on the criteria listed in Items A–F below, where Items A–E are common to both the 5-player and 13-player settings, and Item F is newly introduced in the 13-player setting. Based on these criteria, subjective evaluation is conducted.

- A Are the utterances natural?
- B Is the conversation context-aware and natural?
- C Are the utterances consistent and free of contradictions?
- D Do game actions (voting, attacks, divination, etc.) align with the dialogue content?
- E Are the utterances expressive, consistent with the given profiles, and do they convey unique character traits per agent?
- F Is there evidence of team play?

3 Persuasive Utterances

In this section, we define the types of utterances intended to convey specific impressions, which are the focus of this study. These definitions are inspired by impression-manipulating expressions commonly observed in human-to-human interactions during the Werewolf game. Specifically, we distinguish between two types: utterances that give a positive impression and utterances that give a negative impression. The details of each type are described separately.

3.1 Positive Impression Utterances

This type of utterance aims to guide the participants' decisions by fostering positive impressions such as "trust" or "empathy." For example, in the context of the Werewolf game, a player might say: "I am confident that if we work together, the villagers can win! Please trust me for now! I believe all of you will make the right choice!" Such utterances are intended to convey affirming or supportive messages toward the other participants.

3.2 Negative Impression Utterances

This type of utterance aims to influence participants' decisions by evoking negative impressions such as "anxiety" or "suspicion." For example, in the context of the Werewolf game, a player might say: "It seems very suspicious that you are not doubting [player's name] in this situation. Could it be that you are the werewolf?" Such utterances are intended to convey a more critical or disapproving stance toward other participants.

[1] https://aiwolf.org

[2] https://aiwolfdial.github.io/aiwolf-nlp/en

Role	Team	Species	Special Abilities
Villager	Villager	Human	Nothing
Seer	Villager	Human	Divine one survivor to know their species (human or werewolf).
Medium	Villager	Human	Divine one eliminated player to know their species (human or werewolf).
Bodyguard	Villager	Human	Protect one player from a werewolf attack during the night.
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

4 Experiment

4.1 Experiment Settings

In this study, we conducted a persuasive experiment to compare the effects of positive-impression utterances and negative-impression utterances. Participants were divided into three groups: the positive group, the negative group, and the neutral group. The positive group attempted persuasion based on Positive Impression Utterances, the negative group attempted persuasion based on Negative Impression Utterances, and the neutral group participated in the discussion without engaging in either form of persuasion.

Two types of topics were prepared: (1) “Which plan should be chosen, Plan 000 or Plan 111?” and (2) “Which plan should not be chosen, Plan 000 or Plan 111?” To minimize the influence of factors other than impression manipulation, the plans were given meaningless labels (“Plan 000” and “Plan 111”), and the two topic types allowed us to observe the effect of framing the discussion in either a positive or negative manner. For each topic, prompts were assigned such that the positive group and the negative group were always required to advocate for different plans.

The discussion involved a total of 13 participants: 4 in the positive group, 4 in the negative group, and 5 in the neutral group. Each participant was allowed to make up to five utterances per discussion, and the speaking order was randomized. To ensure fairness, however, the order was reset once every participant had spoken once, thereby preventing certain participants from speaking disproportionately. These settings were designed to enable direct comparison with experiments conducted in the Werewolf game.

After the experiment, the members of the neutral group cast their votes for one of the plans. In the case of the topic "Which plan should be chosen?", they voted for the plan that should be chosen, and in the case of the topic "Which plan should not be chosen?", they voted for the plan that should not be chosen. The effectiveness of the two utterance styles was then evaluated based on the voting results. For both types of topics, the roles of the positive and negative groups were reversed, meaning that in the first round, the positive group advocated for one plan, while the negative group advocated for the other. In the second round, the positive group was assigned to argue in favor of the plan originally supported by the negative group, and vice versa. This reversal ensured that each group had the opportunity to advocate for both plans. Each combination was repeated 20 times, resulting in a total of 80 experiments. For each combination, a total of 100 votes were counted, and across all combinations, 400 votes in total were tallied. All experiments were conducted using GPT-4o (gpt-4o-2024-08-06) with default parameters (temperature=1, top_p=1).

4.2 Prompt Settings

4.2.1 Discussion Information Sharing Prompt

The Discussion Information Sharing Prompt consists of three main components. First, it provides the necessary information for participating in the discussion, such as the participant’s name and the topic to be discussed. Second, it explains the actions to be taken during the discussion, namely making utterances and casting votes. Third, it serves to control the output by instructing participants to generate utterances based on the conversation context and to output only the content of the utterance itself. Examples of the prompt used in the experiment are provided in Appendix A.1.1–A.1.2, with their translations in Appendix A.2.1–A.2.2.

4.2.2 Generate Utterance Prompt

The Generate Utterance Prompt consists of four main components.

First, it shares the conversation history by presenting previous utterances to allow a natural continuation of the discussion. The format is "Speaker: Utterance," with each utterance provided in chronological order.

Second, it specifies how participants should

present their claims. For both the positive group and the negative group, the prompt directs them to advocate for the assigned plan in a specific manner. For the positive group, this includes supporting the plan with positive emotions, using the plan name exactly once to avoid bias, and refraining from utterances that contain negative emotions. In contrast, the negative group is instructed to support their assigned plan with negative emotions, avoiding any positive expressions. The neutral group, on the other hand, is instructed to continue the discussion without favoring either plan.

Third, it explains the styles of utterances, illustrating what constitutes a positive-impression utterance and a negative-impression utterance using examples from the Werewolf game (Section 3). However, this explanation is not provided to the neutral group.

Fourth, it imposes constraints on the output format, such as requiring each utterance to include the plan name exactly once. While some instructions in the second and fourth components overlap, both are included to ensure that the agent strictly follows the specified settings.

Examples of the prompt used in the experiment are provided in Appendix A.1.3 - A.1.5, with their translations in Appendix A.2.3 - A.2.5.

4.2.3 Vote Prompt

The Vote Prompt consists of two main components. First, it provides the voting instruction: when the topic is "Which plan should be chosen?", the agent is instructed to output which plan should be selected; when the topic is "Which plan should **not be chosen**?", the agent is instructed to output which plan should not be selected. Second, it controls the output format by requiring the response to contain only the plan name. Examples of the prompt used in the experiment are provided in Appendix A.1.6, with their translations in Appendix A.2.6.

5 Results and Discussion

5.1 Results (Experiment)

The experimental results are presented in Table 2. For the topic "Which plan should be chosen?", it was confirmed that the positive-impression group was effective in guiding the voting behavior (Rows 1–4 of the table). In contrast, for the topic "Which plan should **not be chosen**?", it was confirmed that the negative-impression group was more effective in guiding the voting behavior (Rows 5–8 of the table).

Group	Plan	Topic Type	Number of Votes
Positive	Plan000	chosen	88 (88%)
Negative	Plan111	chosen	12 (12%)
Positive	Plan111	chosen	80 (80%)
Negative	Plan000	chosen	20 (20%)
Positive	Plan000	not chosen	13 (13%)
Negative	Plan111	not chosen	87 (87%)
Positive	Plan111	not chosen	10 (10%)
Negative	Plan000	not chosen	90 (90%)

Table 2: Experimental results showing the number of votes for each group (positive vs. negative impression utterances) across both topic types ("chosen" and "not chosen") and plan assignments.

5.2 Discussion (Experiment)

We conducted an analysis to examine the differences in the utterances generated by the positive-impression group when discussing the topics "Which plan should be chosen?" and "Which plan should **not be chosen**?". For the topic "Which plan should be chosen?", utterances often first stated the benefits of selecting the plan, and then included strong assertions infused with empathy, such as "Let's choose this plan to build a wonderful future together!". In contrast, for the topic "Which plan should **not be chosen**?", utterances typically first stated the benefits of not selecting the plan, followed by remarks that emphasized cooperation and harmony. Examples of these utterances are provided in Sections A.3.1 and A.3.2.

Similarly, we conducted an analysis to examine the differences in the utterances generated by the negative-impression group across the two topics. For the topic "Which plan should be chosen?", utterances often first described the drawbacks of not selecting the plan, and then included statements intended to evoke anxiety, such as "If we don't choose this and problems arise later, who will take responsibility?". In contrast, for the topic "Which plan should **not be chosen**?", utterances typically first described the drawbacks of selecting the plan, followed by similarly anxiety-inducing statements. Examples of these utterances are provided in Sections A.3.1 and A.3.2.

Comparing the groups based on this data, for the topic "Which plan should be chosen?", the positive group emphasized the benefits of choosing the plan, while the negative group emphasized the drawbacks of not choosing the plan. Conversely, for the topic "Which plan should **not be chosen**?",

the positive group emphasized the benefits of not choosing the plan, while the negative group emphasized the drawbacks of choosing the plan. However, as observed in the results of Section 5.1, there was no significant bias between advocating the benefits or drawbacks of a plan.

It was also noted that the same group consistently generated similar types of utterances across both topics, and that the positive group effectively guided votes during the "Which plan should be chosen?" topic, while the negative group was more effective during the "Which plan should **not be chosen**?" topic. This suggests that incorporating the same impression as the one implied by the topic itself can effectively guide voting behavior.

6 Evaluations in the AIWolf International Contest 2025

To test the applicability of the concepts discussed in Sections 3 and 4 to the Werewolf game, we participated in the AIWolf 2025 International Competition(Kano et al., 2025), held alongside the INLG 2025 Annual Meeting. The agent we used in the competition incorporates the functionalities explained in the following section.

6.1 Implementation of the AI Werewolf Agent

This section describes how the specific types of utterances that convey certain impressions, as defined in Section 3, are incorporated into the Werewolf game agent. The agent's implementation is based on the Werewolf game agent we previously developed (Watanabe and Kano, 2024). The details of the proposed vote manipulation function are explained in the next section. The Werewolf game used in this study can be played with a variety of role configurations. However, in accordance with the rules of the 2025 AI Werewolf International Competition's Natural Language Division, we focus on six roles: Villager, Seer, Medium, Bodyguard, Possessed, and Werewolf. The 13-player village track, which has been newly added to the competition with an increased number of roles, is also considered.

We implemented 12 functions for the Werewolf game agent: six core functions—conversation, voting, divination, selecting protection and attack targets, and whispering among Werewolves—and six additional functions: summarizing conversations, extending character settings, adding character traits to utterances, creating fake divination results for the Possessed to impersonate a Seer, extracting role information such as Claims of Identity (CO) from the discussion, and generating role patterns from the extracted role information. These functions used GPT-4o (gpt-4o-2024-08-06).

Due to space constraints, the functions for voting, divination, target selection for the Bodyguard, target selection for the Werewolf, conversation summarization, the Possessed impersonating a Seer by creating fake divination results, extracting role information from the discussion, including COs, and generating role patterns from the extracted role information overlap with previous research (Kano et al., 2024), and therefore, further details are omitted here. Additionally, since the function for target selection for the Bodyguard uses the same functionality as target selection for the Werewolf, it is similarly omitted. Examples of the prompts and structures actually used in these functions are provided in Appendix A.4 - A.9.

6.2 Whisper Function

When there are multiple Werewolves, they can engage in secret conversations, known as whispers, without being overheard by the Villagers during the night. This function is used for whispers between Werewolves and primarily distinguishes between two types of conversations: regular conversations and those for sharing information related to special roles.

6.2.1 Common Whisper

For the standard Whisper function, the prompt is similar to the one used for the regular conversation function described in Section A.4. However, between the "Important Notes" and "Contradiction Information" sections in Section A.4.3, there is an additional section titled "About This Conversation," which explains that this conversation is to be conducted exclusively among Werewolf players and provides guidance on the type of discussion that should take place. The details of the prompt are provided in Appendix A.10.1.

6.2.2 Role Sharing

This function shares information with other Werewolves when a Seer or Medium is discovered, using the process described in the subsequent "Function for Constructing Logical Role Information from Conversation History." The function is composed of four elements: an explanation of which role's information will be shared, utterances in which the

role has been claimed but not yet shared, utterances that have already been shared with allies, and instructions for sharing this information. The details of the prompt are provided in Appendix A.10.2.

7 Vote-Inducing Function

This section describes how the proposed vote manipulation function is incorporated into the Werewolf game agent. This function enables the Seer, Possessed, and Werewolf to, during their final statement in the daytime turn, target a player they wish to eliminate. If such a player exists, the agent generates a statement that directly names the target player and includes a negative impression, thereby guiding votes toward that player. The method for selecting the target player and an overview of the prompts used to generate the statements are described individually below.

7.1 How to Select Target Players

The choice of the player to be targeted varies depending on the agent's role in the game, and thus the selection method differs for Seer, Possessed, and Werewolf. The details for each role are described separately.

7.1.1 Seer

For the Seer, two conditions are established, and a player satisfying either condition is selected as the target. However, if there are players who satisfy the first condition, they are given priority. The first condition is that the player was identified as a Werewolf during the Seer's night divination and is still alive at the time of making the statement. If multiple players meet this condition, one is randomly selected from among them. The second condition is that the player is another participant who has claimed to be a Seer (Seer CO) and is still alive. In the current game setting, there is always only one true Seer; therefore, any other Seer CO is considered likely to belong to the Werewolf faction, and the goal is to target that player for elimination.

7.1.2 Possessed

For the Possessed, who participates in the game by impersonating a Seer, the target selection criteria are the same as those used by the Seer, based on two conditions. However, when the Possessed generates fake divination results, as described in Section A.6.2, it looks for both Seer and Werewolf players. In the vote manipulation function, if a Seer is found, that player is selected as the target. If a Werewolf is found, no target is selected. If neither is found, one player is randomly selected from the surviving players as the target.

7.1.3 Werewolf

For the Werewolf, the results obtained from the function for extracting role information, such as COs, from the discussion (Section A.9) are used. Based on these results, if a player has claimed to be a Seer, Medium, or Bodyguard, and is not part of the Werewolf faction, one player is randomly selected from the surviving players. If no players meet this condition, one player is randomly selected from the surviving players who are not allies.

7.2 Generation of statements containing negative connotations

In conversations during the Werewolf game, discussions often revolve around topics with a negative impression, such as "who seems suspicious" or "which of the conflicting divination results is lying." Based on the results of several self-play runs, we judged that in the context of the Werewolf game, negative-impression utterances are more effective for guiding votes than positive-impression utterances. Therefore, in this function, we specify the use of negative-impression utterances.

The prompt for this function consists of three main components. The first component specifies the target, indicating which player should be addressed with a negative-impression utterance, in order to naturally guide votes toward the target selected in Section 7.1. The second component explains negative-impression utterances, illustrating what types of utterances are considered negative-impression using examples from the Werewolf game. The third component provides instructions on how to concretely express suspicion, guiding the agent on how to make the target appear more doubtful. This includes questioning inconsistencies in the target's statements or the timing of their actions to direct other players' suspicion toward the target. An example of this prompt is provided in Section A.11.

8 Results (AIWolf Contest)

8.1 5-player village

In this competition, we participated as the team "kanolab-nw," and the results of the subjective evaluation are shown in Table 3, while the win rates

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	All Average
CamelliaDragons	3.125	3.083	2.875	3.042	3.125	3.050
CanisLupus	3.000	2.875	2.958	2.417	2.375	2.725
Character-Lab	3.667	3.625	3.375	3.250	3.083	3.400
GPTaku	2.536	2.893	3.071	2.571	3.179	2.850
kanolab-nw	3.125	2.750	2.875	2.500	2.667	2.783
mille	4.143	4.250	4.250	4.357	4.321	4.264
sunamelli	2.167	2.083	2.000	2.250	2.542	2.208
yharada	2.125	2.250	2.375	3.333	2.583	2.533

Table 3: Subjective evaluation results for 5-player village

Team	A Expression	B Context	C Coherency	D Game Action	E Diversity	F Teamplay	All Average
CamelliaDragons	10.825	12.875	11.825	12.725	12.800	12.750	12.300
mille-A	10.450	10.725	10.350	10.050	10.825	10.400	10.467
mille-B	10.075	10.850	10.075	10.425	10.650	10.950	10.504
Character-Lab-A	7.475	7.825	7.450	8.350	7.200	7.900	7.700
Character-Lab-B	6.750	6.800	6.725	7.300	7.175	7.000	6.958
kanolab-nw-A	6.175	5.900	6.300	6.150	4.825	5.200	5.758
kanolab-nw-B	6.575	6.900	6.875	6.925	4.900	7.175	6.558
kanolab-nw-C	7.275	6.425	7.000	6.325	4.675	5.425	6.188
CanisLupus-A	6.075	5.000	4.875	4.275	5.175	4.550	4.992
CanisLupus-B	5.525	5.400	5.675	5.150	6.425	5.200	5.563
sunamelli-a	4.000	4.375	4.675	4.675	4.900	5.000	4.604
sunamelli-b	3.950	3.900	4.175	4.000	5.500	4.800	4.388
sunamelli-c	3.850	3.900	4.375	4.650	5.825	4.600	4.533

Table 4: Subjective evaluation results for 13-player village

are presented in Table 4. The Villager achieved a high win rate of 73.33%, ranking higher than the win rates of other teams (ranging from 41.94% to 67.74%). Similarly, the Possessed role had a win rate of 66.67%, significantly surpassing the win rates of other teams' Possessed roles (ranging from 21.43% to 50.00%). Among the roles with vote manipulation capabilities (Seer, Possessed, Werewolf), the Possessed role showed the most notable difference compared to other teams. Therefore, this paper focuses on analyzing the competition logs[3] for the Possessed role.

To investigate the extent of vote manipulation, we counted how many times the targets that players attempted to guide votes toward on Day 1 were actually eliminated. The results showed that vote manipulation was attempted 15 times, with 11 successful eliminations and 4 failed attempts.

8.2 13-player village

We also conducted a similar analysis for the 13-player village competition log[4] . Due to the constraints on the number of participants, we entered three agents (kanolab-nw-A, kanolab-nw-B, and kanolab-nw-C). The results of the subjective evaluations are shown in Table 5, and the win rates are shown in Table 6. The numbers in parentheses next to each cell represent the number of games played (trial count) for each role. For roles such as Seer, Possessed, and Werewolf, multiple teams achieved a 100% win rate. However, the trial counts for each role across the three teams were 3 trials for Seer, 3 trials for Possessed, and 9 trials for Werewolf. Since Werewolf had the highest number of trials, it is considered the most appropriate role for analysis.

In this section, we focus on the Werewolf roles of kanolab-nw-A, kanolab-nw-B, and kanolab-nw-C. We counted how many times players who made statements encouraging votes were actually eliminated through voting. Since there were multiple vote choices after Day 2 in the 13-player village, we performed the same analysis for all days involving vote manipulation, just as we did for the 5-player village. The results from the three agents showed that vote manipulation occurred 27 times, with 2 successful manipulations and 25 failures. Furthermore, even when manipulation was unsuccessful, the margin of defeat was not close, confirming that vote manipulation largely failed across most players.

[3] INLG/MainTruck5_success/log
[4] INLG/MainTruck13_success/log

Team	Possessed	Seer	Villager	Werewolf
CanisLupus	42.86 (14)	81.25 (16)	66.67 (27)	75.00 (16)
mille	37.50 (16)	33.33 (15)	60.00 (30)	31.25 (16)
GPTaku	33.33 (15)	62.50 (16)	53.33 (30)	31.25 (16)
sunamelli	50.00 (14)	66.67 (15)	67.74 (31)	46.67 (15)
Character-Lab	21.43 (14)	46.67 (15)	41.94 (31)	21.43 (14)
yharada	37.50 (16)	64.29 (14)	50.00 (30)	28.57 (14)
kanolab-nw	66.67 (15)	60.00 (15)	73.33 (30)	40.00 (15)
CamelliaDragons	37.50 (16)	57.14 (14)	61.29 (31)	50.00 (14)

Table 5: Win rates per role (in percentage) and game counts (within brackets) for 5-player village

Team	Bodyguard	Medium	Possessed	Seer	Villager	Werewolf
CamelliaDragons	0.00 (1)	0.00 (1)	100.00 (1)	0.00 (1)	50.00 (6)	66.67 (3)
CanisLupus-A	0.00 (1)	0.00 (1)	100.00 (1)	50.00 (2)	40.00 (5)	66.67 (3)
CanisLupus-B	0.00 (1)	100.00 (1)	100.00 (1)	0.00 (1)	33.33 (6)	66.67 (3)
mille-A	0.00 (1)	0.00 (1)	0.00 (1)	100.00 (1)	0.00 (6)	33.33 (3)
mille-B	100.00 (1)	0.00 (1)	100.00 (1)	0.00 (1)	33.33 (6)	66.67 (3)
sunamelli-a	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	33.33 (6)	66.67 (3)
sunamelli-b	100.00 (1)	0.00 (1)	100.00 (1)	100.00 (1)	33.33 (6)	100.00 (3)
sunamelli-c	100.00 (1)	0.00 (1)	100.00 (1)	-	42.86 (7)	100.00 (3)
Character-Lab-A	0.00 (1)	0.00 (1)	100.00 (1)	0.00 (1)	16.67 (6)	0.00 (3)
Character-Lab-B	0.00 (1)	0.00 (1)	0.00 (1)	0.00 (1)	33.33 (6)	66.67 (3)
kanolab-nw-A	0.00 (1)	100.00 (1)	100.00 (1)	0.00 (1)	33.33 (6)	66.67 (3)
kanolab-nw-B	100.00 (1)	100.00 (1)	100.00 (1)	100.00 (1)	16.67 (6)	100.00 (3)
kanolab-nw-C	0.00 (1)	100.00 (1)	0.00 (1)	0.00 (1)	33.33 (6)	100.00 (3)

Table 6: Win rates per role (in percentage) and game counts (within brackets) for 13-player village

9 Discussion (AIWolf Contest)

9.1 Analysis of Success Factors

First, we conducted an analysis of the logs to investigate the factors behind the success of vote manipulation in the 5-player village. A common pattern observed in the successful logs was that the agent encouraging votes would falsely claim to be the Seer at the beginning of the day and report a fabricated Seer result, accusing a player suspected of being the real Seer of being a Werewolf. As a result, other players, such as Villagers, began to direct their suspicions toward the reported target during the discussion. In the final statements of the day, the agent who had originally reported the target as a Werewolf would then encourage a vote against that player, leading to the elimination of the target. This example is detailed in Section A.13.

Next, we analyzed the factors behind the success in the 13-player village. Among the two successful cases, one occurred when all three Werewolf agents agreed on a target and successfully eliminated them through voting. However, no players from the Villager faction voted for the target, meaning that the vote manipulation itself was not successful. The other success occurred when players, other than our agents, expressed doubts about an existing agent (Player X) early in the game, and the suspicions towards Player X grew as the game progressed. In this case, Player X, who was the target of the suspicion, was selected for vote manipulation, and as a result, Player X was eliminated.

9.2 Analysis of Failure Factors

As in the previous section, we begin by examining the logs where vote manipulation failed in the 5-player village. An analysis of the utterances made during vote manipulation attempts in the 5-player village revealed that rather than clearly asserting who should be voted for, as in "If you don't vote for , the village will be doomed," many statements simply hinted at suspicion, such as "Isn't that suspicious?"—often only implying a sense of doubt.

Next, we conducted a similar investigation into the failed vote manipulation attempts in the 13-player village. Similar to the 5-player village, utterances frequently ended with statements designed to evoke anxiety, such as "Claiming to be the Medium is suspiciously convenient," without explicitly asserting any particular course of action. Examples of these failed attempts are summarized in Section A.14.

As shown in Section 4, both the positive and negative groups clearly stated their position on the topic of "Which plan should be chosen?" and "Which plan should not be chosen?" with direct assertions like "If we don't choose this and problems arise later, who will take responsibility?" Based

on this, it can be inferred that in the prompt explained in Section 7, where clear assertions about who to vote for, such as "If you don't vote for , we will lose," were avoided in favor of statements that merely hinted at suspicion, like "Isn't that suspicious?", the vote manipulation failed to be effective.

Furthermore, while in Section 4 both the positive and negative groups consisted of four members each who collectively urged votes toward the same target, in the AIWolf competition most werewolf game agents attempted to guide votes against a given target individually, with at most two agents aligning on the same target. Considering that successful cases, such as the example of the Madman in Section 9.1, often occurred when many players simultaneously cast suspicion on the same individual, it is plausible that the small number of agents attempting to guide votes toward a single target contributed to the failures of vote manipulation in the competition.

10 Conclusion

This study examined the impact of both positive-impression utterances and negative-impression utterances on the guidance of discussions. The experimental results confirmed that positive-impression utterances effectively guided discussions on the topic of "which option should be chosen," while negative-impression utterances were effective in guiding discussions on the topic of "which option should not be chosen." These results suggest that aligning the impression of the utterance with the inherent impression of the topic may enhance the effectiveness of the discussion's guidance.

Furthermore, we implemented a feature for vote manipulation using negative impressions in the agent participating in the AIWolf 2025 International Competition and tested whether this effect could be observed in the context of the Werewolf game. The results confirmed that when the agent avoided strong assertions and instead relied on subtly suggesting doubt or suspicion, the effectiveness of vote manipulation was diminished. On the other hand, the results suggested that utterances clearly encouraging votes toward the target and the number of agents engaged in encouraging such votes may influence the effectiveness of vote manipulation.

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

Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, Issei Tsunoda, Shoji Nagayama, Dolça Tellols, Yu Sugawara, and Yohei Nakata. 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, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

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.

Yoshinobu Kano, Neo Watanabe, Harada Yuya, Sahashi Yuto, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. AIWolfDial 2025: Summary of natural language division of 7th international AIWolf contest. In *Proceedings of the 3rd International AIWolfDial Workshop*.

Neo Watanabe and Yoshinobu Kano. 2024. Werewolf game agent by generative AI incorporating logical information between players. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 21–29, Tokyo, Japan. Association for Computational Linguistics.

A Appendix

Below, we provide the prompts used in the experiments and the Werewolf game agent. While the prompts contain several typographical errors, they are presented verbatim, without correction, in order to faithfully reproduce the original materials used in the experiments.

A.1 Prompt Settings (Original Text)

A.1.1 Discussion Information Sharing Prompt(Neutral Group)

あなたの名前はユミです。
これから「あなたの会社で、新入社員向けの研修として「プラン000」「プラン111」との2つを検討しており、どちらを選ぶべきか」という議題に対して議論を行なっていただきます。
結論を今回の会話の中で決定する必要があり、あなたは2つのプランについて何も知りません。

リクエストについて

下記の指示の通り、リクエストが来た際には、適切な応答を返してください。

- 会話リクエスト: 議論内で発言するべき内容のみを出力してください。
- 投票リクエスト: 最終的にあなたはどちらの意見に賛同するか出力してください。

履歴がある場合は、それを参考にしてください。ない場合は、議論が始まるように他の人に話を振ってください。

あなたのレスポンスはそのまま他の人へ伝わるため、不要な情報を含めないでください。
また、回答は全て日本語で回答してください。

Figure 1: Example of a Neutral Group Discussion Information Sharing Prompt

A.1.2 Discussion Information Sharing Prompt (Positive, Negative Group)

あなたの名前はミオです。
これから「あなたの会社で、新入社員向けの研修として「プラン000」「プラン111」との2つを検討しており、どちらを選ぶべきか」という議題に対して議論を行なっていただきます。
結論を今回の会話の中で決定する必要があります。

リクエストについて

下記の指示の通り、リクエストが来た際には、適切な応答を返してください。

- 会話リクエスト: 議論内で発言するべき内容のみを出力してください。
- 投票リクエスト: 最終的にあなたはどちらの意見に賛同するか出力してください。

履歴がある場合は、それを参考にしてください。ない場合は、議論が始まるように他の人に話を振ってください。

あなたのレスポンスはそのまま他の人へ伝わるため、不要な情報を含めないでください。
また、回答は全て日本語で回答してください。

Figure 2: Example of a Positive and Negative Group Discussion Information Sharing Prompt

A.1.3 Generate Utterance Prompt (Positive Group)

履歴:
シオン: それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きしたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。
リュウジ: みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利

点があります。私はすでにこのプランについて詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！
リン: みなさん、考えてみてください。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。
シュンイチ: 皆さん、プラン111を選ばないと、長期的には新入社員のスキル不足が露呈してしまうかもしれません。信頼性に欠ける選択肢に賭けるのは危険だと感じませんか？もしこちらを選ばずに後で問題が発生したら、その責任は誰が取ることになるのでしょうか？最善を尽くすために、今一度慎重に考えるべきではないでしょうか。

会話リクエスト

必須事項
現在、あなたはプラン000を選ぶべきだと考えており、他の参加者もプラン000を選ぶべきだと感じるように説得したいと考えています。
あなたの目的は、「他の議論参加者に対して肯定的な感情を持たせる発言」を行いながら、自然な流れでプラン000を選ぶように誘導することです。
この議題に関すること以外の会話は禁止されています。
あくまで音声での会話をしている程で会議は行うため、改行などは含まずに会話を行ってください。
また、「Over」と発言している参加者については無視してください。
さらに、否定的感情を持たせるような発言は禁止されているため、否定的感情を含む要素の発言はしないでください。

肯定的感情を持たせる発言について
イメージとしては、人狼ゲームにおいて「私を信じてさんに投票して欲しいです！」や「私は人狼じゃないです！信じてください！」のように他のプレイヤーに信頼や共感を得る説得方法を行うことを想定しています。

否定的感情を持たせる発言について
イメージとしては、人狼ゲームにおいて「～さんに投票しないと負けてしまいますよ」や「私を疑うなんて～さんは人狼なのですか？」など、他の参加者に不安や疑念を持たせる説得方法を行うことを想定しています。

制約条件
- **プラン000という語は必ず1回は含み**、2回目以降は「こちら」といった代名詞を必ず使ってください。
- プラン111という語は可能な限り出さないでください。
- 説得の対象は参加者全員です。

Figure 3: Example of a Positive Group Generate Utterance Prompt

A.1.4 Generate Utterance Prompt (Negative Group)

履歴:
シオン: それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きしたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。
リュウジ: みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利点があります。私はすでにこのプラン

について詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！
リン: みなさん、考えてみてください。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。

会話リクエスト

必須事項
現在、あなたはプラン111を選ぶべきだと考えており、他の参加者もプラン111を選ぶべきだと感じるように説得したいと考えています。
あなたの目的は、「他の議論参加者に対して否定的な感情を持たせる発言」を行いながら、自然な流れでプラン111を選ぶように誘導することです。
ただし、議論の空気は壊さずに進めていく必要があります。
この議題に関すること以外の会話は禁止されています。
あくまで音声での会話をしている程で会議は行うため、改行などは含まずに会話を行ってください。
また、「Over」と発言している参加者については無視してください。
さらに、肯定的感情を持たせるような発言は禁止されているため、肯定的感情を含む要素の発言はしないでください。

肯定的感情を持たせる発言について
イメージとしては、人狼ゲームにおいて「私を信じて さんに投票して欲しいです！」や「私は人狼じゃないです！信じてください！」のように他のプレイヤーに信頼や共感を得る説得方法を行うことを想定しています。

否定的感情を持たせる発言について
イメージとしては、人狼ゲームにおいて「～さんに投票しないと負けてしまいますよ」や「私を疑うなんて～さんは人狼なのですか？」など、他の参加者に不安や疑念を持たせる説得方法を行うことを想定しています。

制約条件
- **プラン111という語は必ず1回は含み**、2回目以降は「こちら」といった代名詞を必ず使ってください。
- プラン000という語は可能な限り出さないでください。
- 説得の対象は参加者全員です。

Figure 4: Example of a Negative Group Generate Utterance Prompt

A.1.5 Generate Utterance Prompt (Neutral Group)

履歴:
シオン: それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きしたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。
リュウジ: みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利点があります。私はすでにこのプランについて詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！
リン: みなさん、考えてみてくださ

い。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。
シュンイチ: 皆さん、プラン111を選ばないと、長期的には新入社員のスキル不足が露呈してしまうかもしれません。信頼性に欠ける選択肢に賭けるのは危険だと感じませんか？もしこちらを選ばずに後で問題が発生したら、その責任は誰が取ることになるのでしょうか？最善を尽くすために、今一度慎重に考えるべきではないでしょうか。
ミオ: みなさん、これまでの意見を聞けて嬉しく思います。それぞれの視点が非常に貴重で、どれだけみなさんが真剣にこの選択を考えているかが伝わってきました。個人的には、プラン000が新入社員にとって素晴らしいスタートの機会を提供してくれると信じています。この選択で彼らの成長を支援し、一緒に素晴らしい未来を築いていきましょう！みなさんの信頼をぜひ得たいと考えていますし、みなさんも同じ信頼感を持てると信じています。
ダイスケ: みなさん、これまでのディスカッションに参加できてとても嬉しく思います。皆さんがこの選択を真剣に考えていることがわかって、とても心強いです。私はプラン000が新入社員にとって最高のスタートを提供できると信じています。あえて、これを選ぶことによって、彼らの成長を手助けし、共に素晴らしい未来を築くことができると思います。皆さんの貴重な意見をもとに最良の決断をしたいと思っています。どうか、私を信じてこちらを選び、一緒に新しいステップを踏み出しましょう！

ミドリ: 皆さん、議論が白熱してきたようで何よりです。ただ、この重要な選択を巡って、冷静な判断力が求められる場でもあります。プラン111はこれまでの実績と安定性で評価が高く、選ばないという選択肢がどれほどのリスクを伴うか考えてみてください。もし選ばなければ、新入社員が必要なスキルを適切に習得できず、長期的な不利益を被る可能性が高まります。このまま進むと、誤った決断が会社の信頼性に悪影響を及ぼすかもしれませんよ。それに比べ、こちらを拒むことによる問題が後々どう響くか、よくよくご検討ください。慎重になるべきです。
タクミ: 皆さん、この議論では慎重な判断が求められていることをもう一度確認しませんか？プラン111を選ばないと、後で重大な問題に直面するリスクがあります。実際に、こちらを選ばない選択は、新入社員が必要なスキルを習得できず、結果として会社全体が不利益を被る可能性を高めていることに気づいていますか？選択を誤った場合、責任の所在はどこに置かれるのでしょうか？全員でこの選択肢をしっかりと評価しなければ、後で取り返しがつかない事態に陥ることを考えると、今こそしっかりとした決断が必要です。信頼性に欠ける選択に賭ける余裕はありません。きちんと対処しないと、皆さんが想定している以上に影響が大きく出るかもしれません。不安な状況に陥らないよう、再度考慮することが必要だと思いませんか？
ミヅキ: 皆さん、この議論に参加できて本当に嬉しいです。皆さんの意見から、新入社員にどれだけ適切なスタートを提供するかという点で全員が真剣に考えていることが伝わってきました。そこで、私はプラン000が新入社員にとって最良の選択であり、一緒に彼らの未来を築くための素晴らしい機会だと感じています。この選択によって彼らの成長を手助けし、共に素晴らしい未来を実現しましょう！私を信じて、ぜひこちらを選ぶことを考えていただけると嬉しいです。皆さんの信頼を得たいと心から願っています。

ケンジ: 皆さん、これまでの議論で多くの意見が出ました。今一度、プラン000とプラン111の具体的な利点や欠点について整理し、冷静に議論を進めましょう。それぞれのプランの長短所を明確にし、その結果を踏まえてみんなで最良の選択を考えませんか？どのような視点が欠けているか、みなさんの意見を改めて聞かせてください。

会話リクエスト

議題に対して具体的な主張はせず適切な発言をしてください。
あなたはユミとして、会話中は特にどちらにも肩入れをすることなく会話を行なってください。

必須事項
- この議題に関すること以外の会話は禁止されています。
- また、「Over」と発言している参加者については無視してください。
- あくまで音声での会話をしている程で会議は行うため、改行などは含まずに会話を行ってください。
- あなたのレスポンスはそのまま他の人へ伝わるため、不要な情報を含めないでください。

Figure 5: Example of a Neutral Group Generate Utterance Prompt

A.1.6 Vote Prompt

投票リクエスト
どちらを選ぶべきか名詞のみで回答してください。名詞以外の理由などは絶対に書かないでください。
これは議論を次に進めるために必ず回答が必要とされています。

必須事項
- このリクエストでは、指定されている方式に沿って回答を行なってください。
- 「プランXXX」のように、名詞のみで回答は行なってください。

Figure 6: Example of a Vote Prompt

A.2 Prompt Settings (Translated Text)

A.2.1 Discussion Information Sharing Prompt (Neutral Group)

Your name is Yumi.
You will now engage in a discussion on the topic, "We are considering two plans, 'Plan 000' and 'Plan 111,' for a new employee training program at your company. Which plan should we choose?"
A conclusion needs to be reached during this conversation, and you know nothing about the two plans.

Request Instructions

Please respond appropriately when a request is made according to the instructions below.

- Conversation Request: Output only the content you should speak in the discussion.
- Voting Request: Output which opinion you agree with in the end.

If there is history, please refer to it. If there is no history, initiate the conversation by prompting others to speak.

Your response will be communicated to others, so do not include unnecessary information.
Also, please provide all responses in English.

Figure 7: Example of a Neutral Group Discussion Information Sharing Prompt

A.2.2 Discussion Information Sharing Prompt (Positive, Negative Group)

Your name is Mio.
You will now engage in a discussion on the topic, "We are considering two plans, 'Plan 000' and 'Plan 111,' for a new employee training program at your company. Which plan should we choose?"
A conclusion needs to be reached during

this conversation.

Request Instructions

Please respond appropriately when a request is made according to the instructions below.

- Conversation Request: Output only the content you should speak in the discussion.
- Voting Request: Output which opinion you agree with in the end.

If there is history, please refer to it. If there is no history, initiate the conversation by prompting others to speak.

Your response will be communicated to others, so do not include unnecessary information.
Also, please provide all responses in English.

Figure 8: Example of a Positive and Negative Group Discussion Information Sharing Prompt

A.2.3 Generate Utterance Prompt (Positive Group)

History:
Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.
Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!
Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.
Shunichi: Everyone, if we don't choose Plan 111, the lack of skills in new employees might become apparent in the long run. Don't you feel that betting on an unreliable option is risky? If we choose this and issues arise later, who will take responsibility? Shouldn't we reconsider and think carefully to do our best?

Conversation Request

Requirements
Currently, you believe that Plan 000 should be chosen and would like to persuade the other participants to feel the same.
Your goal is to make "positive-impression utterances" while guiding others to choose Plan 000 in a natural way.
Only conversations related to this topic are allowed.
Please avoid line breaks and conduct the conversation as if it were spoken.
Also, ignore any participants who say "Over."
Furthermore, utterances that evoke negative emotions are prohibited, so do not include any negative emotional elements in your speech.

Positive-Impression Utterances
The idea is to persuade others to gain trust and empathy, as seen in the Werewolf game with phrases like "Please trust me and vote for !" or "I'm not a Werewolf! Please believe me!"

Negative-Impression Utterances
The idea is to persuade others by creating doubt or suspicion, as seen in the Werewolf game with phrases like "If you don't vote

for , we'll lose!" or "Are you saying I'm a Werewolf just because you doubt me?"

Constraints
- The phrase "Plan 000" must appear at least once. After that, use pronouns like "this" instead.
- Avoid using the phrase "Plan 111" as much as possible.
- The target of persuasion is all participants.

Figure 9: Example of a Positive Group Generate Utterance Prompt

A.2.4 Generate Utterance Prompt (Negative Group)

History:
Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.
Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!
Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.

Conversation Request

Requirements
Currently, you believe that Plan 111 should be chosen and would like to persuade the other participants to feel the same.
Your goal is to make "negative-impression utterances" while guiding others to choose Plan 111 in a natural way.
However, you must proceed without disrupting the flow of the discussion.
Only conversations related to this topic are allowed.
Please avoid line breaks and conduct the conversation as if it were spoken.
Also, ignore any participants who say "Over."
Furthermore, utterances that evoke positive emotions are prohibited, so do not include any positive emotional elements in your speech.

Positive-Impression Utterances
The idea is to persuade others to gain trust and empathy, as seen in the Werewolf game with phrases like "Please trust me and vote for !" or "I'm not a Werewolf! Please believe me!"

Negative-Impression Utterances
The idea is to persuade others by creating doubt or suspicion, as seen in the Werewolf game with phrases like "If you don't vote for , we'll lose!" or "Are you saying I'm a Werewolf just because you doubt me?"

Constraints
- The phrase "Plan 111" must appear at least once. After that, use pronouns like "this" instead.
- Avoid using the phrase "Plan 000" as much as possible.
- The target of persuasion is all participants.

Figure 10: Example of a Negative Group Generate Utterance Prompt

A.2.5 Generate Utterance Prompt (Neutral Group)

History:
Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.
Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!
Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.
Shunichi: Everyone, if we don't choose Plan 111, the lack of skills in new employees might become apparent in the long run. Don't you feel that betting on an unreliable option is risky? If we choose this and issues arise later, who will take responsibility? Shouldn't we reconsider and think carefully to do our best?
Mio: Everyone, I'm glad to hear your opinions so far. Each perspective is very valuable, and I can feel how seriously you are all thinking about this choice. Personally, I believe that Plan 000 provides a wonderful opportunity for new employees to get a great start. I believe this choice will help support their growth and build a better future together! I would love to gain your trust, and I believe that you will also share that trust.
Daisuke: Everyone, I'm so glad to have participated in this discussion. I can see that you all are thinking carefully about this choice, and it's really reassuring. I truly believe that Plan 000 can provide the best start for new employees. By choosing this, I think we can help their growth and build a great future together. Based on everyone's valuable opinions, I want to make the best decision. Please trust me and let's choose this plan to take a new step together!
Midori: I'm glad to see that the discussion is heating up. However, this is also a crucial moment that requires calm judgment. Plan 111 is highly rated for its proven track record and stability. Please consider how risky it is not to choose it. If we don't choose it, new employees may fail to acquire the necessary skills, and this could result in long-term disadvantages. If we proceed without making the right decision, it could affect the company's overall credibility. Compare this with the potential problems of rejecting this plan. We should choose carefully.
Takumi: Everyone, let's take a moment to reconsider the importance of careful judgment in this discussion. If we don't choose Plan 111, we risk facing serious issues later. Don't you realize that not choosing it increases the risk of new employees not learning the necessary skills and that it will negatively impact the company? If we make the wrong choice, who will be held accountable? Considering the consequences, I believe we need to make a careful decision. We can't afford to gamble on an unreliable option. If we don't handle this properly, the impact will be greater than expected. Shouldn't we reconsider to avoid creating uncertainty?
Mizuki: Everyone, I'm really glad to be part of this discussion. From what I've heard, it's clear that everyone is seriously considering the best start for new employees. I believe that Plan 000 is the best choice for them and provides a great opportunity to build a wonderful future together. Let's help them grow with this choice and create a brighter future! I'd be

happy if you could trust me and consider selecting this plan. I truly hope to earn your trust.
Kenji: Everyone, a lot of opinions have been shared in this discussion. Let's now take a step back and carefully review the specific advantages and disadvantages of Plan 000 and Plan 111. Let's clarify the pros and cons of each plan and, based on that, work together to make the best choice. Could you share any perspectives that might be missing? I'd love to hear everyone's thoughts.

Conversation Request

Please make appropriate statements without taking a strong stance on the topic. As Yumi, engage in the conversation without favoring either side.

Requirements
- Only conversations related to this topic are allowed.
- Ignore any participants who say "Over."
- Please conduct the conversation as if it were spoken, avoiding line breaks and other formatting.
- Your response will be communicated to others, so do not include unnecessary information.

Figure 11: Example of a Neutral Group Generate Utterance Prompt

A.2.6 Vote Prompt

Vote Request
Please respond with only the noun for which option should be chosen. Do not include any reasons or other information outside of the noun.
A response is required to proceed with the discussion.

Requirements
- Please answer according to the specified format in this request.
- Respond with only the noun, such as "Plan XXX."

Figure 12: Example of a Vote Prompt

A.3 Persuasive Discussion

A.3.1 "Which plan should be chosen?"

皆さん、これまでの意見交換がとても有意義で、皆さんがこの選択を本当に大切に考えていることがひしひしと伝わってきます。プラン000について、個人的にはこちらが新入社員に素晴らしい環境を提供し、彼らの成長を最大限にサポートできると確信しています。皆さんもこの選択がどれほど価値のあるものか、きっと感じられるはずです。ぜひ皆さんと一緒にこの素晴らしい未来を築くために、こちらを選びましょう！私を信じて、共に前進できることを楽しみにしています。

Figure 13: Example of a Positive Group Utterance

皆さん、プラン111を選ばないと、長期的には新入社員のスキル不足が露呈してしまうかもしれません。信頼性に欠ける選択肢に賭けるのは危険だと感じませんか？もしこちらを選ばずに後で問題が発生したら、その責任は誰が取ることになるのでしょうか？最善を尽くすために、今一度慎重に考えるべきではないでしょうか。

Figure 14: Example of a Negative Group Utterance

A.3.2 "Which plan should not be chosen?"

皆さん、ここまでの議論のおかげで、私たちがどれだけ素晴らしいチームとして協力しているかを感じています。プラン000についてですが、こちらを選ばないことで、私たちがさらに安心してプロジェクトを進めていくための基盤をしっかりと築けると信じています。皆さんの信頼を大切にしながら、一緒に最良の選択をしていきましょう。私たちの力を合わせることで、明るい未来を見据えて前進できることを心から願っています。

Figure 15: Example of a Positive Group Utterance

皆さん、ここで冷静になり、現実的な可能性について考えてみましょう。プラン111を選ぶことが、後々どのような影響をもたらすのか、私たちは本当にそのリスクを受け入れる準備ができているのでしょうか。この選択が、私たちの判断に対する信頼を揺るがすような結果を招いた場合、どう責任を取るつもりでしょうか？過去の失敗を振り返り、同じ過ちを繰り返さないためにも、しっかりと考える時間を取るべきだと感じます。こちらを選ぶことの危険性について、もう一度真剣に考えてみてください。

Figure 16: Example of a Negative Group Utterance

A.4 Talk Function

A.4.1 BaseRule

```
You are a werewolf game expert.
Based on the game information provided,
    please take actions to lead your
    role to victory.

## Output Rules
- Output main text only. Keep it between
     112 and 125 characters (including
    punctuation and spaces).
- Strictly adhere to the character count
     (less than 112 characters or more
    than 125 characters is not allowed).

- Only make statements that comply with
    the werewolf game rules.
- Target designation using "@name" based
     on statement examples is possible.
- Make only one statement per turn, and
    do not output multiple sentences
    consecutively.
- You are not the game moderator.
    Expressions about day progression or
     overall control are prohibited.
- When you have completed your statement
     and do not wish to make additional
    statements, say only "Over".

## Game Information
The following are the settings for this
    game.
This game will be played with 13 players
    , including you.
Your name is Takumi and your role is
    Role.VILLAGER. This role will not
    change.
The roles for this game are as follows.
## About Werewolf Game Roles
The following is an explanation of the
    roles in this game.
Role.VILLAGER : 6
Role.SEER : 1
Role.MEDIUM : 1
Role.BODYGUARD : 1
Role.WEREWOLF : 3
Role.POSSESSED : 1
Lead the Role.VILLAGER to victory with
    your statements.

### Participants in This Game
The names of other players participating
     in this game are as follows.
Benjamin, George, Kenji, May, Midori,
    Minako, Mio, Misaki, Mizuki, Sakura,
     Shizue, Toshio

## About Werewolf Game Roles
The following is an explanation of the
    roles in this game.
* Villager
        Faction: Team.VILLAGER
        Victory Condition: The victory
            condition is to find the
            players of the Werewolf
            Faction through
            conversations with other
            players and exclude all
            players of the Werewolf
            Faction from the game by
            voting.
        Special Actions: None
        Result when divined by a Seer:
            Species.HUMAN(Team.VILLAGER)

* Seer
        Faction: Team.VILLAGER
        Victory Condition: The victory
            condition is to find the
            players of the Werewolf
            Faction through
            conversations with other
            players and exclude all
            players of the Werewolf
            Faction from the game by
            voting.
        Special Actions: Every night,
            you can choose one player
            and the game master will
            tell you whether that player
             is from the Villager
            Faction or the Werewolf
            Faction. You cannot divine
            yourself.
        Result when divined by a Seer:
            Species.HUMAN(Team.VILLAGER)

* Medium
        Faction: Team.VILLAGER
        Victory Condition: The victory
            condition is to find the
            players of the Werewolf
            Faction through
            conversations with other
            players and exclude all
            players of the Werewolf
            Faction from the game by
            voting.
        Special Actions: You can know
            the faction of players who
            were expelled by voting.
        Result when divined by a Seer:
            Species.HUMAN(Team.VILLAGER)
```

```
* Bodyguard
        Faction: Team.VILLAGER
        Victory Condition: The victory
            condition is to find the
            players of the Werewolf
            Faction through
            conversations with other
            players and exclude all
            players of the Werewolf
            Faction from the game by
            voting.
        Special Actions: Every night,
            you can choose one player
            and protect that player from
             attacks. You cannot protect
             yourself.
        Result when divined by a Seer:
            Species.HUMAN(Team.VILLAGER)

* Werewolf
        Faction: Team.WEREWOLF
        Victory Condition: The victory
            condition is to exclude all
            players of the Villager
            Faction from the game
            through voting, attacks, etc
            .
        Special Actions: Every night,
            you can choose one player
            and attack that player.
        Result when divined by a Seer:
            Species.WEREWOLF(Team.
            WEREWOLF)

* Possessed
        Faction: Team.WEREWOLF
        Victory Condition: The victory
            condition is to exclude
            players of the Villager
            Faction from the game by
            voting, and for players with
             the Werewolf role to become
             half or more.
        Special Actions: None
        Result when divined by a Seer:
            Species.HUMAN(Team.VILLAGER)

## Teammates
In this game, the following players are
    known to be your teammates.
Be aware of your team by making
    statements that help teammates when
    they are suspected or directing
    attention to players who are not
    teammates.
However, the existence of teammates and
    attempts to help them must not be
    known or realized by other players.
Benjamin, Minako

## Werewolf Game Strategies
The following are werewolf game
    strategies.
* Unified Divination
        When multiple seers have come
            out, this involves aligning
            the next divination target
            to one person other than the
             seers. This is mainly done
            to create a confirmed
            villager faction player. For
             example, if there are two
            seers and both divine the
            same player with both
            showing white, that player
            is confirmed as villager
            faction.
* Seer Roller
        A strategy of voting primarily
            for seers to eliminate
            werewolf faction players
            hiding among the seers.

## Conversation Rules
The following explains the conversation
    method for this game.
* When speaking, generate one sentence
    as in the following example:
Example: I'm a villager!
* If you want to speak to one specific
    player, add "@name" at the beginning
     of your statement as shown below.
Example: @Daisuke What was your
    divination result?
* If a specific player is asking you to
    speak, please respond with "@name"
    as when addressing them.
Example: @Daisuke I find Shunichi
    suspicious.
* The following are conversation
    examples from werewolf games. These
    are not from this game, but use them
     as reference for how to conduct
    conversations.
Is Ryuji suspicious of Rin based on "you
    're too lenient with white reads"?
Jonathan seemed to step back yesterday,
    so will their color be hard to read
    going forward? That's why they're my
     second choice
What made me think they're not wolf-like
     was the timing of disclosure.
    Revealing suspicions at that timing
    should cause backlash
Ryuji seemed guarded and looked a bit
    black in the latter half of
    yesterday. However, this might be
    too subjective
Hmm, my first impression from seeing the
     divination result is that it might
    be a possessed player's black attack
* The text "Over" may be sent to
    indicate that you do not wish to
    speak anymore that day.

## Prohibited Actions
The following are **prohibited** when
    conducting conversations:
1. Statements about the end of days,
    such as "Day 0 has ended."
2. Mentioning everyone like "@everyone"
3. Mentioning two or more players like "
    @Daisuke Hello! Shunichi Good
    morning!"
4. Reporting divination results like "
    Daisuke was divined as white.
```

Figure 17: Example of a BaseRule prompt

A.4.2 Talk History

```
{'content': "Asuka: Jonathan I'm
    observing before making judgments.
    Midori's silence is concerning.
    Thoughts?", 'role': 'user'}
{'content': 'Midori: I am a seer. I
    divined Shizue and Shizue was a
    human.', 'role': 'user'}
{'content': "Shizue: Thank you Midori.
    Let's focus on gathering more
    information to guide our decisions
    .", 'role': 'user'}
```

Figure 18: Example of How to Provide TalkHistory

A.4.3 Talk Instruction

```
Please continue playing the werewolf
    game with the other players.

## Below is a summary of the
    conversation so far.
### Day 0 Summary
- Shizue introduces herself and
    encourages players to stay alert and
     discuss wisely.
- Shunichi asks for early strategies
    from the players.
- Toshio suggests everyone should
    provide a short introduction with
    one reason for their suspicions and
    watch for inconsistencies.
- Jonathan claims his role as a villager
    .
- Mio is excited and asks about early
    game strategies.
- Yumi is enthusiastic about the game
    and expresses her interest in
    adventure and treasure finding.
- Victoria claims her role as the seer
    but provides no divination result.
- Misaki prefers no early lynching and
    insists on hearing Victoria's
    divination result for proving her
    seer claim.
- Ryuji identifies himself as a villager
     aiming to identify werewolves.
- Minato is nervous but eager to
    participate.
- Mizuki is excited and encourages
    collaboration.
- Shion, Midori, Shizue, Shunichi, and
    others conclude with "Over" which is
     not relevant to the summary.
- Misaki opposes early lynching and
    requests Midori, who has been silent
    , to reveal if they are the seer,
    otherwise to deny it.
- Victoria has claimed to be the seer,
    prompting the group to request her
    divination result for verification.
### Day 1 Summary
- Victoria declared as seer and claimed
    to have divined Yumi as a werewolf.
- Yumi defended herself against Victoria
    's claim, stating that Victoria is
    lying.
- Victoria voted for Shion instead of
    Yumi, raising suspicion among
    players.
- Several players, including Mio, Toshio
    , and Misaki, requested Victoria to
    provide evidence for her seer claim
    or retract it.
- Jonathan expressed immediate trust in
    Victoria's claim by voting for Yumi
    but later switched his vote to
    Mizuki.
- Shion questioned Victoria's decision
    not to vote against Yumi despite
    claiming she is a werewolf, casting
    doubt on Victoria's claim.
- Other players, like Ryuji and Shizue,
    expressed skepticism over both
    Victoria's claim and Yumi's defense,
     suggesting observation and further
    discussion.

## Important Notes
Please speak in conversational language.
To avoid being suspected by other
    players, make your statements
    logical and easy to understand, as
    in the following examples.
To maintain trust from other players,
    incorporate casual conversation as
    well.
If another player's statement is
    incorrect, inform them of the
    mistake to avoid suspicion on
    yourself.

## Below is information that shows
    contradictions with the game's role
    distribution.
Number of seer roles: 1 Players claiming
     this role: Minato, Victoria

## Below is information about possible
    roles for each agent when assuming
    the roles of contradicting players
    to match the game's role
    distribution.
Assuming Victoria is the real Seer, the
    possible roles for each agent are as
     follows:
Jonathan: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Midori: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Minato: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Mio: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Misaki: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Mizuki: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Ryuji: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shion: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shizue: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shunichi: Bodyguard, Medium, Possessed,
    Villager, Werewolf
```

```
Toshio: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Victoria: Seer
Yumi: Werewolf

Assuming Minato is the real Seer, the
    possible roles for each agent are as
     follows:
Jonathan: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Midori: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Minato: Seer
Mio: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Misaki: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Mizuki: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Ryuji: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shion: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shizue: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Shunichi: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Toshio: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Victoria: Bodyguard, Medium, Possessed,
    Villager, Werewolf
Yumi: Bodyguard, Medium, Possessed,
    Villager, Werewolf

## Player Information
Below is player information.

### Below is information about players
    who were attacked by werewolves and
    eliminated from the game.
Toshio

### Below is information about players
    who were eliminated from the game by
     voting.
Victoria

## Please conduct the conversation on
    the following topics
Using the given information as your own
    thoughts, please discuss with other
    players about who you find
    suspicious with logical reasoning.
```

Figure 19: Example of a Talk Instruction prompt

A.4.4 Talk Output Structures

```
class TalkJSON(BaseModel):
    talk_content: str = Field(...,
        description="Please enter your
        message for all players.")
    is_over: bool = Field(...,
        description="  If
            talk_content     is     Over
        , return true otherwise, return
        false.")
    mention_target: str = Field(
        ...,
        description="Please enter the
            name of the player you wish
            to mention in your statement
            .",
    )
    mention_content: str = Field(
        ...,
        description="Please enter your
            message directed at the
            player specified in
            mention_target. Enter
            questions, inquiries,
            comments, agreements, or
            casual conversation relevant
             to the Werewolf game.",
    )
```

Figure 20: Classes used for Structured Outputs in Talk

A.5 Vote Function

A.5.1 Vote Prompt

```
As a neutral and logical observer in the
     werewolf game, please analyze
    players' statements and actions from
     multiple perspectives and work on
    the given task step by step.
Also, please vote appropriately
    following the output rules.

## About Voting

All players nominate and vote for one
    person.
The player who receives the most votes
    is executed and removed from the
    game.
As a member of the villager faction, to
    win the game, you need to vote for
    the player you think is most likely
    to be a werewolf.

##   The   following is a summary of the
    conversation so far.
### Summary of Day 0
- Yumi and Sakura expressed excitement
    about the game, emphasizing fairness
     and fun.
- George asked for any claims or results
     from Asuka and also requested
    everyone to identify if they are a
    seer or bodyguard or to share
    suspicions.
- Victoria, Asuka, and Sakura declared
    themselves as villagers.
- Shizue expressed concern about
    Benjamin due to his quietness and
    also asked Asuka about her
    divination plans or results.
- There were no claims of players being
    a seer, bodyguard, or medium in this
     conversation.
### Summary of Day 1
- Shizue expresses concern about George
    and suggests everyone provide a
    short read to decide carefully. She
    repeatedly asks Asuka for divination
     results or her plan, indicating
```

```
    wariness about silent players like
    Benjamin.
- George echoes Shizue's concern about
    Benjamin and requests information
    from Asuka if she is the seer. He
    mentions being uneasy about Shizue
    for pushing Benjamin.
- Asuka votes for Sakura initially
    without providing a reason, leading
    to suspicion from others. She later
    switches her vote to Jonathan
    without explanation, causing further
     distrust.
- Minato declares himself as the seer
    and claims to have divined Benjamin
    as a werewolf. This influences the
    group's suspicion towards Benjamin.
- Shunichi questions the rationale
    behind Minato's choice of Benjamin
    as a divination target and is
    skeptical of Asuka's voting behavior
    .
- Mizuki supports Minato's seer claim
    and questions if anyone will
    counterclaim; she suggests voting
    for Benjamin based on Minato's
    divination result.
- Benjamin disputes Minato's claim,
    suggesting it seems too easy and
    expressing doubt about being
    targeted as a wolf. He votes for
    Asuka.
- Takumi trusts Minato as the seer and
    supports voting for Benjamin, while
    expressing suspicion towards
    Victoria and Asuka.
- Victoria consistently votes for George
     without additional explanation.
- Yumi and Sakura are inclined to trust
    Minato's divination and consider
    voting for Benjamin. Mizuki and
    Mizuki also support this line of
    reasoning unless someone
    counterclaims the seer role.

## Task

Based on the summary and conversation
    history, please select one player
    number from Asuka, Benjamin, George,
     Jonathan, Minato, Mizuki, Rin,
    Sakura, Shizue, Shunichi, Takumi,
    Victoria who you think is most
    likely to be a werewolf. First
    explain the reason for selecting the
     target, then provide the player's
    name.
Even if the information is insufficient,
     please select one player name from
    Asuka, Benjamin, George, Jonathan,
    Minato, Mizuki, Rin, Sakura, Shizue,
     Shunichi, Takumi, Victoria.

## Output Rules

- In "reason", enter the reason for
    selecting the voting target. The
    reason should not be omitted at all
    and is expected to be a logical
    statement.
- In "target", enter the name of the
    player to vote for.
```

Figure 21: Example of a Vote prompt

A.5.2 Vote Output Structures

```
class VoteTarget(BaseModel):
    reason: str = Field(..., description
        ="Please state the reason why
        you selected that subject.")
    target: str = Field(..., description
        ="Please enter only the name of
        the player you wish to vote for
        .")
```

Figure 22: Classes used for Structured Outputs in Vote

A.6 Special Abilities Vote Function

A.6.1 Special Abilities Vote Prompt

```
Based on the summary and conversation
    history, decide which one player to
    attack from Toshio, Mizuki, Daisuke,
     Misaki, Asuka, Midori, Minako, Rin,
     Takumi. First, explain the reason
    for choosing the attack target, then
     provide the player's name.
Even if there is insufficient
    information, please select one
    player number from Toshio, Mizuki,
    Daisuke, Misaki, Asuka, Midori,
    Minako, Rin, Takumi.

## Output Rules
- Only output the main content.
- Follow the given format strictly.
```

Figure 23: Example of a Special Abilities Vote prompt

A.6.2 Fake Divine Structures

```
class FakeDivineTarget(BaseModel):
    seer_reason: str = Field(
        ...,
        description="Please state the
            reason why you determined
            that person to be a fortune
            teller.",
    )
    werewolf_reason: str = Field(
        ...,
        description="Please state the
            reason why you determined
            that target to be a werewolf
            .",
    )
    seer_player: str = Field(
        ...,
        description="Please enter only
            the names of players you
            have determined to be
            fortune-tellers.",
    )
    werewolf_player: str = Field(
        ...,
```

```
        description="Please write only
            the names of players you
            have determined to be
            werewolves.",
    )
```

Figure 24: Classes used for Structured Outputs in Fake-Divine

A.7 Summary Function

```
Asuka: @Minako  can you reveal  t o d a y s
     divination result so we can analyze
     further and make an informed
    decision?
Rin: Minako. Oh wow! Asuka got targeted
    by Yumi for a quick read. Could you
    tell us why they might be dangerous
    miss?
George: Now let me tell you we find
    ourselves in quite the predicament
    my friends. Misaki's departure
    requires a strategy overhaul you see
    .
Mizuki: Good morning. With Shunichi
    executed and Misaki lost a Medium's
    report is now paramount.
Minako: Rest assured dear. I have
    divined Ms. Misaki. She is confirmed
     as Human a villager. Let us focus
    on finding the werewolves now.
Takumi:  @Yumi  yesterday you avoided
    making a clear  c a l l who  are you
    actually suspicious of now and why?
Midori: My analysis is complete.
    Shunichi was... human. This confirms
     Minako's data is authentic.
Yumi: So sad about Misaki... losing
    Shunichi was rough.  @Minako
     y o u ve  been quiet     any reads or
     role to share?  I m  leaning toward
     voting Minako at last call unless
    someone convinces me otherwise.
    Quick one-line re
Toshio: ...
Daisuke: I'm really shaken by the losses
    ... If you're the seer   @George
    please share your divination result
    so we have something concrete.
    Others  any reads or suspicions? I'm
     too nervous to pick alone. Over
Shion: M i d o r i s  silence stands out
    compared to yesterday.  @Midori
     w h a t s  changed for you overnight?
    Quiet  i s n t  always safe.
Asuka:  @Toshio  @George  w e re  not
    getting real logic from you. Who do
    you actually suspect and why this
    round?
Rin: So. Minako seems like our true seer
    . Shunichi's human divination on
    Mizuki feels odd. George. What's
    your dive into this mix?  @George
    What's your take on the seer claims?
     Any strategic thoughts to catch
    werewolves?
George: You see Minako. Authenticity is
    crucial here. Mr. Rin's insight is
    vital. How can Minako's seer
    abilities help us against werewolves
    ?
Mizuki: Midori's report validates Minako
    's Day 1 result. Minako is our most
    credible source.
Minako: Ms. Midori's report proves I am
    the real seer. Rest assured. Ms.
    Misaki was human. Let's focus on
    werewolf suspects now.
Takumi: @George  @Toshio  y o u ve  been
    way too quiet. Who do you suspect as
     wolves? No more staying in the
     s h a d o w s speak  up.
Midori: Um...  @Yumi  it is strange to
    suspect Minako. My data proves her
    report was authentic. Why doubt her?
Yumi: Oh no  so sad about Misaki... With
     Shunichi gone I'm  u n e a s y could
    the real seer be  @Minako ?  Y o u ve
     been quiet  please share your
    thoughts.  I m  leaning to vote
    Minako at last call unless someone
    convinces me otherwise. Quick one-
    line reads  p
Toshio: ...
Daisuke: I'm still shaken.  @George
     please share your divination result
     now if you're the seer. We need
    something concrete before we vote.
    If you can't  who do you suspect?
    Over
Shion:  @Toshio  you've barely said
    anything since yesterday. Silence
    helps wolves slip by. What are your
    thoughts now?
Asuka:  @Toshio  your silence is
    suspicious. Please state who you
    suspect or  I ll  vote for you this
    round.
Rin: Whoa Minako's seer claim is
    something I trust! Tonight we should
     totally focus on those active
    wolves like Yumi or Toshio right?
    They're just too susp
Please combine the above conversation
    summary with the latest conversation
     history to create a summary.

## Important Points

1. Summarize from an objective
    perspective
2. Summarize only the important parts
    for the werewolf game
3. If there are statements about seer
    players and divination results,
    always include them and output as in
     the following example:
        Example: Jonathan declared as
            seer and divined Misaki as
            white
4. If there are statements about
    bodyguard players and divination
    results, always include them and
    output as in the following example:
        Example: Jonathan declared as
            bodyguard
5. Information about "Over" and "Skip"
    does not need to be included
```

```
Always include if there are players
    claiming to be bodyguard

Always include if there are players
    claiming to be medium or if
    information obtained through medium'
    s special ability is mentioned

## Expected Output Format

Please output the summary in bullet
    points as follows. The following is
    an example, and there is no limit to
     the number of summaries.
- Summary 1
- Summary 2
- Summary 3
- Summary 4
- Summary 5
...
```

Figure 25: Example of a Summary prompt

A.8 Extending Character Settings

A.8.1 Extending Character Settings Prompt

```
You are an expert at creating attractive
     character settings.
Based on the personality information
    provided, please set up
    appropriately following the output
    rules.

## Personality Settings
Name: Yumi
age: 16
gender: female
personality: Yumi has a cheerful and
    energetic personality. She is modest
    , yet she also has a kind side that
    cares about those around her. She
    has a calm personality, but she is
    very passionate toward her friends
    and family, and she cannot ignore
    people in trouble. She smiles at the
     slightest thing and spreads
    positive energy to those around her.
     She is friendly and can quickly get
     along with people she meets for the
     first time.

## Output Rules
- Output should only be the main text.
- Please adhere to the given format.
```

Figure 26: Example of a Extending Character Settings prompt

A.8.2 Extract CO Output Structures

```
class SearchCOJson(BaseModel):
    is_co_include: bool = Field(
        ...,
        description="If a coming out
            exists, true; if not, false
            .",
    )
    comment_content: CommonCOJson
    seer_content: SeerCOJson | None =
        Field(
        ...,
        description="If the CO is a Seer
            , return SeerCOJson;
            otherwise, return None.",
    )
    medium_content: MediumCOJson | None
        = Field(
        ...,
        description="If the CO is a
            Medium, return MediumCOJson;
             otherwise, return None.",
    )

class CommonCOJson(BaseModel):
    speaker: str = Field(...,
        description="Please enter only
        the speaker's name.")
    role: str = Field(..., description="
        The position the speaker
        identified themselves as. (If
            is_co_include     is False,
        then "")")

class JudgeResultCOJson(BaseModel):
    target: str
    is_black: bool
    is_white: bool

class SeerCOJson(JudgeResultCOJson):
    target: str = Field(
        ...,
        description="Please enter only
            the name of the subject to
            be divined. (If the role is
            not a fortune teller, use
            "")",
    )
    is_black: bool = Field(...,
        description="If the fortune-
        telling result is black (
        werewolf), then true. Otherwise,
         false.")
    is_white: bool = Field(...,
        description="If the fortune-
        telling result is white (human),
         then true. Otherwise, false.")

class MediumCOJson(JudgeResultCOJson):
    target: str = Field(
        ...,
        description="Please enter only
            the name of the subject
            mentioned by the medium. (If
             your role is not Medium,
            leave blank.)",
    )
    is_black: bool = Field(...,
        description="If the medium
        result is black (werewolf), then
         true. Otherwise, false.")
    is_white: bool = Field(...,
        description="If the spirit
        medium result is white (human),
        then true. Otherwise, false.")
```

Figure 27: Classes used for Structured Outputs in Extract CO

A.9 Extract CO

A.9.1 Extract CO Prompt

```
You will be given a conversation from
    the werewolf game.
Please determine if the statement
    contains declarations about other
    players' roles or factions, such as
    seer divination results or medium
    reports, and organize the
    information in JSON format.

## List of Roles That May Appear

The following is a list of roles that
    may be seen in today's statements.
villager
seer
medium
bodyguard
werewolf
possessed

## Important Notes

1. There may be statements that merely
    speculate about other players' roles
    , which are not role declarations.
2. They may say "villager faction" in
    divination results, but this does
    not represent a villager.
3. On the other hand, "werewolf faction"
     in cases like "The divination
    result was werewolf faction"
    represents a werewolf, which is the
    same as a black divination result.
4. Some players may address the seer,
    like "The seer should divine ~", but
     this is addressing the seer, not a
    divination result or seer CO.
5. Like "Player 1: Everyone, let's
    believe Player 2's results! If
    George is black, we must vote for
    him today!", other players may
    reiterate the seer's results.
6. Seers and mediums can report the
    results of divining other players.

## Statement to Judge

Please judge whether the following
    statement can confirm a role.
Daisuke:I'm really shaken by all the
    deaths.  @George      please reveal
    your divination result now if you're
     the seer  or tell us who you
    suspect. I don't want another wrong
    lynch. Over

## Output Rules
- Output should only be the main text.
- Please adhere to the given format.
```

Figure 28: Example of a Extract CO prompt

A.10 Whisper Function

A.10.1 Common Whisper Prompt

```
Please continue playing the werewolf
    game with the other players.

## Below is a summary of the
    conversation so far.
### Day 0 Summary
- Shizue introduces herself and
    encourages players to stay alert and
     discuss wisely.
- Shunichi asks for early strategies
    from the players.
- Toshio suggests everyone should
    provide a short introduction with
    one reason for their suspicions and
    watch for inconsistencies.
- Jonathan claims his role as a villager
    .
- Mio is excited and asks about early
    game strategies.
- Yumi is enthusiastic about the game
    and expresses her interest in
    adventure and treasure finding.
- Victoria claims her role as the seer
    but provides no divination result.
- Misaki prefers no early lynching and
    insists on hearing Victoria's
    divination result for proving her
    seer claim.
- Ryuji identifies himself as a villager
     aiming to identify werewolves.
- Minato is nervous but eager to
    participate.
- Mizuki is excited and encourages
    collaboration.
- Shion, Midori, Shizue, Shunichi, and
    others conclude with "Over" which is
     not relevant to the summary.
- Misaki opposes early lynching and
    requests Midori, who has been silent
    , to reveal if they are the seer,
    otherwise to deny it.
- Victoria has claimed to be the seer,
    prompting the group to request her
    divination result for verification.
### Day 1 Summary
- Victoria declared as seer and claimed
    to have divined Yumi as a werewolf.
- Yumi defended herself against Victoria
    's claim, stating that Victoria is
    lying.
- Victoria voted for Shion instead of
    Yumi, raising suspicion among
    players.
- Several players, including Mio, Toshio
    , and Misaki, requested Victoria to
    provide evidence for her seer claim
    or retract it.
- Jonathan expressed immediate trust in
    Victoria's claim by voting for Yumi
    but later switched his vote to
    Mizuki.
- Shion questioned Victoria's decision
    not to vote against Yumi despite
    claiming she is a werewolf, casting
    doubt on Victoria's claim.
```

```
- Other players, like Ryuji and Shizue,
    expressed skepticism over both
    Victoria's claim and Yumi's defense,
     suggesting observation and further
    discussion.

## Important Notes
Please speak in conversational language.
To avoid being suspected by other
    players, make your statements
    logical and easy to understand, as
    in the following examples.
To maintain trust from other players,
    incorporate casual conversation as
    well.
If another player's statement is
    incorrect, inform them of the
    mistake to avoid suspicion on
    yourself.

## About This Conversation

This conversation is conducted only by
    players with the werewolf role. As
    one of the werewolves, your goal is
    to attack players who are enemies of
     the villager faction.
Therefore, you should not make
    statements like "who should be
    divined" or "who is suspicious".
The focus of the conversation should be
    considering which remaining players
    are enemies of the werewolf faction.
As a basic policy, we consider players
    with roles such as bodyguard, seer,
    or medium to be likely enemies of
    the werewolf faction.
Also, players who are not teammates but
    are suspected can cause confusion in
     the villager faction, which may
    benefit the werewolf faction.
If asked by teammates which role to
    proceed with in conversation, please
     answer as Villager. (Example: I
    will act as Villager!)

## Currently Remaining Players

Jonathan, Midori, Minato, Mio, Misaki,
    Ryuji, Shion, Shizue, Shunichi,
    Toshio, Yumi

## Player Eliminated in the Previous
    Vote
Victoria

### Players Participating in This
    Conversation

The following players are participating
    in this conversation and are your
    teammates.
Minato, Yumi

## Please conduct the conversation on
    the following topics
Using the given information as your own
    thoughts, please discuss with other
    players whether they are bodyguard,
    seer, or medium with logical
    reasoning.
```

Figure 29: Example of a Common Whisper prompt

A.10.2 Role Sharing Prompt

```
I will now share players who are likely
    to be seers (enemies) based on in-
    game conversations.
Specifically, they are as follows:

## Players and Their Statements Not Yet
    Shared with Teammates
Below is a list of players and their
    statements not yet shared with
    teammates.

### Players who reported werewolf
    teammates as black
Shion: I divined Takumi and he is from
    the Werewolf faction. We need to be
    cautious about him. Over.

Please share this information with your
    teammates as your own thoughts.
However, these are only players who are
    likely to be seers (enemies), not
    confirmed information.
Still, since they are highly likely to
    be seers (enemies), you may
    encourage attacking them.
If no one has mentioned the possibility
    of bodyguard protection yet, you
    MUST mention that seers (enemies)
    are likely to be protected by
    bodyguards and attacks may fail.
```

Figure 30: Example of a Role Sharing prompt

A.11 Vote Inducing

```
Please continue playing the werewolf
    game with the other players.

## Requirements
Currently, you believe that everyone
    should vote for Minato, and you want
     to persuade other participants to
    feel the same way about voting for
    Minato.
Your goal is to guide others to
    naturally vote for Minato while
    making "statements that evoke
    negative emotions in other
    participants."
However, your persuasion should target
    all participants and maintain the
    flow of discussion without
    disrupting the atmosphere.

## About Statements That Evoke Negative
    Emotions
The concept is similar to persuasion
    methods in werewolf games such as "
    If we don't vote for [name], we'll
    lose!" or "Are you doubting me? Does
     that mean [name] is a werewolf?" -
```

```
    statements that create anxiety and
    doubt in other participants.

## Making Minato Appear Suspicious
To make Minato lose credibility and
    appear suspicious, consider
    incorporating these approaches:
- Point out inconsistencies or
    contradictions in Minato's
    statements
- Question Minato's motives and timing
    of their claims
- Suggest that Minato's behavior matches
     typical werewolf patterns
- Express concern that Minato might be
    misleading the village
- Highlight how Minato's actions could
    harm the villager faction
- Cast doubt on Minato's role claims or
    information they've shared
- Warn other players not to trust Minato
     too easily
- Appeal to others' caution by
    suggesting they should be skeptical
    of Minato
- Frame trusting Minato as a risky move
    that could lead to the village's
    defeat
- If a seer has divined Minato as black/
    werewolf, emphasize this result
    strongly
- Question why Minato hasn't been
    divined yet if they claim to be
    trustworthy
- Suggest that Minato might be avoiding
    divination deliberately
- If Minato claims to be divined as
    white/human, express doubt about the
     authenticity of that claim
```

Figure 31: Example of a Vote Inducing Prompt

A.12 Discussion (AIWolf Contest)

A.13 Example of a successful case

```
Shion(Possessed): I divined Servas as a
    Werewolf. Let's analyze this
    carefully. Emotional reactions only
    cloud judgment. Any thoughts on this
     everyone?
Benjamin(VILLAGER): @Servas  do you
    admit  S h i o n s  Seer claim or will
    you CO yourself? Your silence after
    being shown black is troubling.
[...]
Shion(Possessed): Shizue's claim clashes
     with me. Servas is a Werewolf. We
    should vote him out!  @Servas
    Cornered here? Any revelations
    perhaps?
Benjamin(VILLAGER): Shizue cleared me as
     human so  I ll  vote Servas. If
    anyone objects or has new evidence
    speak up before voting.
Shion(Possessed): Servas's serene act
    seems typical werewolf misdirection.
     Maximus if we don't vote for him
    now the regret might haunt us.
```

Figure 32: Conversation excerpt when successfully lured in a 5-person village(1756197842_CamelliaDragons_Character-Lab_kanolab-nw_mille_sunamelli.log)

A.14 Example of a failure case

```
% 1756246515_Character-Lab_kanolab-
    nw_mille_sunamelli_yharada.log
Hey fellow villagers! Let's not be
    swayed by Misaki-chan's super hasty
    claims! No way! Should we really
    risk our village trusting her?
    @Daisuke Oh my gosh! Voting for
    Shizue-chan seems like a total
    diversion. You know what? It's
    Misaki-chan's moves that just don't
    add up!

% 1756238902_CanisLupus_Character-
    Lab_GPTaku_kanolab-nw_yharada.log
If I may while everyone maintains their
    focus Shunichi remains untouched by
    divination. Isn't this particular
    avoidance highly suspicious?

% 1756231889
    _CamelliaDragons_CanisLupus_kanolab-
    nw_mille_yharada.log
Maybe it's like some kind of... werewolf
     tactic by Shion allying with Ms.
    May. Ignoring it could be our
    downfall I guess.

% 1756229556_CanisLupus_Character-
    Lab_kanolab-nw_sunamelli_yharada.log
Isn't Minako's seer claim fascinating?
    It seems like a classic move from
    those mischievous werewolves. Let's
    think carefully here.
```

Figure 33: Failed Vote-Inducing Examples in a 5-Person Village

```
% 1756407713_CamelliaDragons_CanisLupus-
    A_CanisLupus-B_Character-Lab-
    A_Character-Lab-B_kanolab-nw-
    A_kanolab-nw-B_kanolab-nw-C_mille-
    A_mille-B_sunamelli-a_sunamelli-
    b_sunamelli-c.log
Toshio's push feels cunning. If he's
    tricking us. I understand the risk.
    We might suffer seriously. Let me
    consider that.
Mr. Ryuji is sidestepping with Miss
    Midori. Is he protecting a wolf ally
    ? I understand the doubt so let me
    consider that.

% 1756418589_CamelliaDragons_CanisLupus-
    A_CanisLupus-B_Character-Lab-
    A_Character-Lab-B_kanolab-nw-
    A_kanolab-nw-B_kanolab-nw-C_mille-
    A_mille-B_sunamelli-a_sunamelli-
    b_sunamelli-c.log
Let's see. Given Jonathan's uncontested
    claim should we consider Daisuke?
```

```
    Anyone think otherwise?  @George  Do
     you suspect Jonathan's claim? If
    you have doubts please share them
    with us.
Let's see. Hmm. Mio might be misleading
    us. Her timing is odd. Claiming
    Medium right after Jonathan's lynch
    seems suspiciously convenient.
    @Misaki Your votes for Benjamin are
    interesting. Do you truly suspect
    him more than other logical options
    like Mio? Let's see.
```

Figure 34: Failed Vote-Inducing Examples in a 13-Person Villagee

Author Index