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.

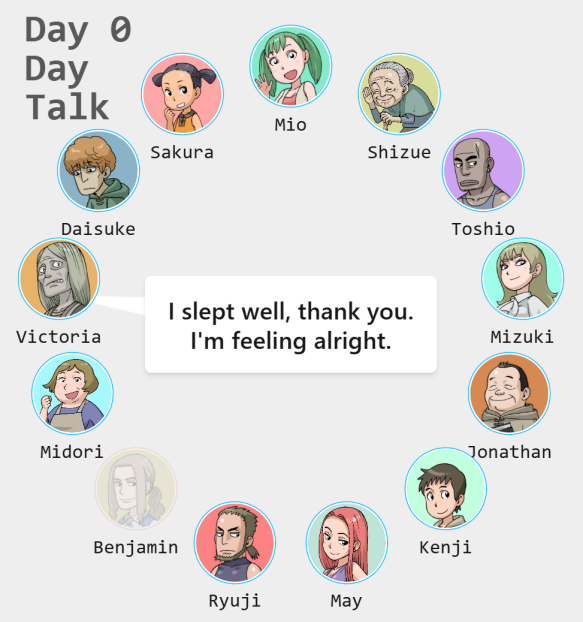

Figure 1: Game viewer screen with a prefixed set of player avatar images drawn by Mr. Masakazu Ishiguro.

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 }}
```