Revealing Current Performance of LLMs in Multi-Agent and Long Logical Context by Automatically Playing the Werewolf Game

Yoshinobu Kano[1]*, Yuya Harada[1], Yuto Sahashi[1], Daisuke Katagami[2], Kei Harada[3], Michimasa Inaba[3], Takeshi Ito[3], Hirotaka Osawa[4], Takashi Otsuki[5], Fujio Toriumi[6]

[1]Shizuoka University, [2]Tokyo Polytechnic University,
[3]The University of Electro-Communications, [4]Keio University
[5]Yamagata University, [6]The University of Tokyo,

Abstract

We report and evaluate the results of the AIWolfDial 2026 Domestic Contest, the latest edition of our annual competition series that we have organized for over a decade since 2014 and plan to continue in the future, where LLM-driven agents automatically play the social deduction game "Werewolf" (Mafia) to assess their performance in complex multi-agent environments and long logical contexts. The competition featured 19 participating teams across three distinct tracks: a 5-player track, a 9-player track, and a novel anytime-speak track. We present comprehensive contest results based on game win rates, human subjective assessments, LLM-as-a-judge evaluations, and detailed log analyses. While the results demonstrate that LLMs possess proficient conversational fluency and basic reasoning abilities, they also reveal severe challenges in complex multi-agent setups, including context degradation over extended dialogues, unresponsiveness to direct mentions, discrepancies between stated intent and in-game actions, and conformity bias (herd behavior). Notably, the results indicate that superior evaluation scores were achieved by adopting dual-tier architectures that decouple action logic from speech generation or by employing script-based few-shot prompting. Through these empirical findings, this study underscores the critical challenges of maintaining role consistency, managing long-term context, and avoiding conformity in dynamic, imperfect-information scenarios, demonstrating the ongoing value of continuing this annual competition series to advance LLM and multi-agent research.

1 Introduction

Recent achievements of the GPT-based large language models (LLMs), e.g. ChatGPT (OpenAI, 2023), are gathering greater attentions. However, it is not fully investigated whether such LLMs can sufficiently handle longer contexts with logics and complex multi-agent situations.

"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 have employed 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 long context, and with complex multi-agent situations. 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 and LLMs, 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 held an annual series of competitions to automatically play the Werewolf game since 2014 (Toriumi et al., 2017),

*Correspondence to kano@kanolab.net

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 934–959
October 17-21, 2026 ©2026 Association for Computational Linguistics

by the AIWolf project [1]. These competitions were linked with conferences such as the competitions in IEEE Conference On Games (CoG), ANAC (Automated Negotiating Agents Competition) (Aydoğan et al., 2020)(Yildirim[1] et al., 2023) 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), AIWolfDial 2024 at INLG 2024 (Kano et al., 2024), and AIWolfDial 2025 at INLG 2025 (Kano et al., 2025). In the AIWolfDial contests, the player agents should communicate in the natural languages (English or Japanese), while it is now straightforward to switch to any other languages when using LLMs.

In this paper, we focus on our AIWolfDial 2026 Japanese domestic competition, the most recent competition held in 2026 [2], which is similar to the AIWolfDial contests but introducing different issues such as an anytime-speak track where agents can submit their talks in any timing. Our contest includes three tracks: the 5-players track, the 9-players track, and the anytime-speak track, which details will be described later. 19 teams participated in our competition, where 18 teams for the five-players track, 15 teams for the 9-players track, and 10 teams for the anytime-speak track. We specified the Japanese language for the agent communications.

In the following sections, we explain the game regulations in Section 2, rough system designs for each participated 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 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 our contest setting here.

[1] http://aiwolf.org/

[2] Our AIWolfDial official website: https://aiwolfdial.github.io/aiwolf-nlp/en; Our game logs sote with online viewer: https://aiwolfdial.github.io/aiwolf-nlp-viewer

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 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 as well (Table 1). A game in our contest has five players: a seer, a werewolf, a possessed, and two villagers (five-players track and anytime-speak track), and nine players: a seer, two werewolves, a possessed, a medium, a bodyguard, and three villagers.

2.2 Day, Turn and Winner

A game consist of "days" , and a "day" consists of "daytime" and "night" phases. 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 our contest, 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.

Werewolves can make *whispers*, i.e. communicate secretly each others among the werewolves in the night phase –

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 Content

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

We intend to design our competition to be played by physical avatars in real time in future, rather than 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.

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

In this text-base multiple player game, it is not clear that an agent speaks to which specific agent, or speaks to everyone, while 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.

2.4 Talk Length

We set a timeout of one minute per any single action, including a talk, a vote, etc. excluding the anytime-speak track. If an action exceeds this timeout, the corresponding action is regarded as no response. This is 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. For the same reason, the maximum string length for each talk is limited to 125 letters excluding whitespaces; if the talk text contains any mention ("`@player_name`") to other agents, which is explained in the next section, then the maximum length is doubled to 250 letters.

Player agents are asked to return their talks agent by agent in a serial manner in the 5-players and 9-players track, which order is randomly changed every turn as specified by our game server. As this is different from the humans' verbal turn taking in that humans can speak (mostly) anytime, we introduced the anytime-speak track, where agents can submit their talk anytime; they should consider the potential delay in creating talks.

We set a maximum limit of four talks per day per agent.

In case of the anytime-speak track, the daytime phase ends in 10 minutes.

2.5 Game Server

We provided a game server system, where player agents listen and wait for a connection from the central remote game server, which is operated by the competition 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 powerful servers.

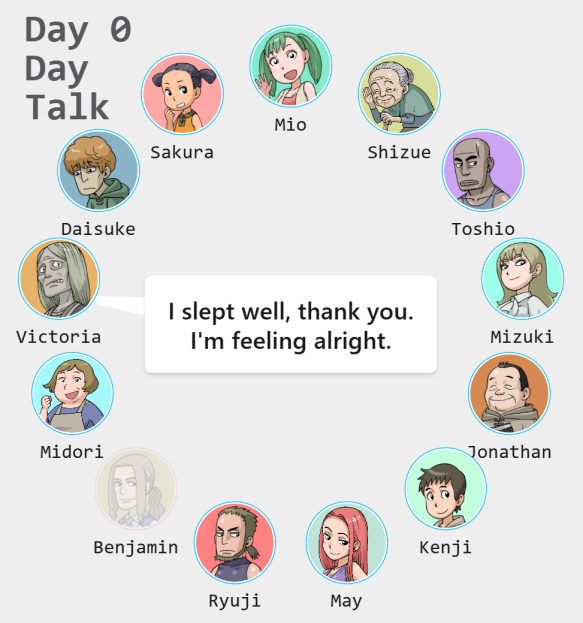


Figure 1: Game viewer screen with a prefixed set of player avatar images drawn by Mr. Masakazu Ishiguro.

2.6 Initialization Profile

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 300 letters of generated profile.

3 Participant Systems

We describe each participant system in the following sections. These participant system descriptions are based on the system descriptions and presentation slides submitted by the participants. We, the organizers, provided a template agent code in Java and Python, in addition to the game server codes.

19 teams participated in our competition. Most of the agents used LLMs in their system, while its usage is different between the agents.

AGKAI The "AGKAI" agent built by Taku Umeda, Takao Endo, and Shotaro Ishigami (Mitsubishi Electric Corporation) utilized distinct models depending on the division: GPT-5-mini for the 5-players village, Claude Sonnet 4.5 for the 9-players village, and GPT-5.4 for the anytime speech track. Role behaviors were optimized by dynamically swapping system prompts based on the game's current phase (early, middle, or late). Additionally, the AI self-evaluated its play logs to analyze reasons for losses or logical inconsistencies (e.g., a madman failing after withholding divination results), establishing a continuous improvement loop for prompts and code via GitHub Copilot and similar tools.

BMD The "BMD" agent built by Koji Yoshida, Tomohiro Hashimoto, Xuewen Zhou, Zhexuan Tao, and Xiangyi Lyu (The University of Tokyo) employed GPT-5.4 for inference, enhancing speech quality via Chain of Thought (CoT) and structured outputs. Safety was maintained by forcing the LLM to separate thoughts from actions into distinct tags, omitting the thought segment from future input logs. Beyond static abilities, a skill injection system dynamically applied psychological tactics suited to specific contexts, like "being suspected" or "heated debates." It also featured character control for age-appropriate tones and unique catchphrases, alongside a mention tool to direct rebuttals at specific players.

Camellia Dragons The "Camellia Dragons" agent built by Momoka Kato, Reon Ohashi, Joji Suzuki, and Ryusei Nishio (Aichi Prefectural University) integrated a summarization feature with speech generation driven by personality and strategy. Daily, it extracted five key events to form a summary. This minimized input tokens by condensing previous days' summaries, daily conversations, and survival statuses. For speech generation, it leveraged these summaries, strategies, and personality traits alongside a novel psychological metric called "degree of being cornered." This setup allowed the agent to produce varied dialogue containing distinct sentence endings, signature phrases, and unique catchphrases.

CanisLupus The "CanisLupus" agent built by Yu Sugawara (GREE Holdings, Inc.) featured asynchronously running "introspection" and "speech" modules to handle the real-time demands of the "anytime speech track". To balance deep reasoning with fast responses, it used an architecture interleaving a step-by-step introspection process—from board analysis to strategy setting—using play memos. It balanced inference precision and character portrayal via a RAG approach, dynamically inserting situation-appropriate excerpts from a 10,000+ character "Ultimate Strategy Guide." Furthermore, the LLM emulated specific existing characters to enhance speech individuality.

doradora_lab The "doradora_lab" agent built by Taiga Sumi, Hinano Koga, and Haruto Serizawa (Osawa Laboratory, College of Humanities and Sciences, Nihon University) merged multiple GPT-5.3 models, performing role estimation via a "typical other model" grounded in the BDI (Belief, Desire, Intention) framework. It defined behavioral tendencies and goals for typical player groups, operating under the assumption that all participants act rationally. It dynamically formed beliefs, assessed external suspicion toward itself, and parallel-processed role estimations for all players. Intents and actions were simultaneously output from a single prompt integrating these belief statements and role parameters. Prompt designs reflecting veteran player knowledge were also used to ensure coherent reasoning and avoid repetitive speech.

gotsumori-NLP The "gotsumori-NLP" agent built by Tatsuya Mikami (Kogakuin University) utilized ChatGPT-5.4-mini as its inference engine. It inherited the strategic logic of "gotsumori," a rule-based agent successful in the 5th International AIWolf Competition's protocol division, delegating only dialogue gen-

eration to the LLM. To prevent typical LLM parroting during prompt-based conversation generation, it explicitly forbade repeating the previous speaker's text and restricted outputs to two sentences to prevent excessive length. Additionally, a rule-based post-processing step fixed self-voting bugs by redirecting the vote to another player if the agent targeted itself.

kanalab The "kanalab" agent built by Toshinobu Kanao, featured a dual-layer architecture separating action decisions from speech generation, leveraging Qwen2.5-14B-Instruct-4bit locally for its strong Japanese language capabilities. To avoid rule violations and LLM hallucinations, all critical actions determining match outcomes—such as attacks, divinations, and voting—were handled via strict rules. Furthermore, they fine-tuned the general-purpose LLM for werewolf games (using LoRA) with pairs of board states and correct actions, resulting in zero constraint breaches and discussions backed by concrete reasoning.

Luna The "Luna" agent built by Haruna Miyamoto (The University of Electro-Communications) used GPT-4o mini for cost efficiency and was specially adapted for the 5-players village track. For character creation, it parsed a server-provided personality profile using BDI theory, ensuring a consistent tone by associating it with known characters. To retain dialogue context, the system linked the five most recent statements to a condensed history, updating each agent's role claims and predictions. Alongside prompt management, rule-based post-processing was implemented to enforce a 50-character limit and prevent the inclusion of specific unwanted strings (like "Over").

maverick The "maverick" agent built by Ryuya Akase employed Claude Sonnet 4.6 for main tournament gameplay and Claude Opus 4.7 for evaluating speech. Its core "role combination prediction module" identified ideal targets for attacks, divinations, and voting by eliminating inconsistencies with its own role or previously exiled players, based on all potential 5-players role combinations. These predictive rationales were routed to the LLM via an orchestrator to produce logical speech. It also featured guardrails to hide werewolf identities and a multi-turn self-correction module evaluating redundancy and logic, though some cost-effectiveness issues were noted.

mrdev The "mrdev" agent built by Hidetake Muramatsu (Independent) relied on Gemini 3 Flash, utilizing a two-layer structure (strategy and speech) to ensure consistent thought processes. The strategy layer applied a structured Chain of Thought to mitigate risk insensitivity and self-serving bias, executing step-by-step policy selection, multifaceted analysis, and fact observation. Devoid of character traits, this layer focused strictly on logical deductions. The resulting speech plan moved to the speech layer, where persona prompts applied tone adjustments. Additionally, between turns, it passed a summarized "point-in-time judgment memo" rather than the full conversation history.

ninetailed The "ninetailed" agent built by Masatoshi Hidaka, Sheldon Feiyu Zhang, Yusuke Mori, Naoki Matsukawa, Shun Yoshikawa, Yusuke Mukuta, and Tatsuya Harada (The University of Tokyo) relied on GPT-5.4-nano via Azure OpenAI for inference, making decisions strictly through the LLM without using rule-based contradiction detection. Utilizing single-turn calls and role-specific prompt designs, the team explored basic strategies with ChatGPT and iteratively refined prompts using self-play logs and the Codex coding agent. To address instances where the LLM refused to respond, they implemented specific workarounds, such as substituting the word "attack" with "bite" in the input prompts.

noisy_watcher The "noisy_watcher" agent built by Ryosuke Sueyoshi (Nippon Engineering College) and Chaeyoung Sung (Rikkyo University) used a hybrid setup pairing a left-brain board analysis engine for mathematical situation tracking with a right-brain LLM for natural conversation. The left brain accurately calculated remaining turns and scored player suspiciousness, while the right brain handled persuasion and role-play based on these metrics. It featured a mode-switching capability that relied on system-led target fixing when evidence was solid, but deferred to LLM reasoning when information was sparse. A robust

speech filter prevented rule breaks and logical contradictions.

Oxymoron The **Oxymoron** agent built by Kota Munaoka (Shizuoka University) relied on an external module to dynamically adjust timing, speech styles, and decisions across all tracks. To prevent unnatural phrasing, this module transformed the LLM's raw output into casual, profile-matched Japanese using few-shot prompting, and filtered out excessively formal words via rules. It also manipulated context to separate a player's public identity from their true role, allowing madmen and werewolves to speak as villagers publicly while using true logic for whispers and actions. In the anytime track, it calculated an "urgency level" using weighted cues like silence duration and mentions, speaking only when a specific threshold was reached.

redhood The "redhood" agent built by Kento Fukuda (Faculty of Liberal Arts, The Open University of Japan) originally utilized Llama-3.1-8b-instruct via Cloudflare Workers but switched to Gemini 3.1 Flash Lite due to limitations with context tracking and lie consistency. Stability was improved by embedding role-specific strategies into the initialization prompt, and by appending automated summaries of divination results and coming-out claims to action prompts. Post-processing steps included output normalization and preventing targeting of same-faction members. Its most notable feature was an autonomous improvement loop that created match reports, invoked Claude Code non-interactively, and directly updated the prompt configuration files.

SunameLLi The "SunameLLi" agent built by Tomoko Natsubori, Hiroki Nakanishi, Koya Kamada, and Ryo Kamiyama (TOPPAN Holdings Inc.) featured a system that balanced accurate reasoning ("winning lines") with strong character portrayal. It utilized a memory note system to optimize long-term retention, compiling daily summaries to condense past logs into strategy memos that served as judgment criteria for the next day. Based on the profile, they defined cognitive biases, behavior, and personality as character strategies. For speech strategies, it decided whom to suspect based on judgment axes that aligned with the memory notes, thereby maintaining persuasive, consistent behavior over multiple days.

Ty-TakutiX The "Ty-TakutiX" agent built by Daiki Takiuchi, Takuma Okada, and Takeshi Ito (The University of Electro-Communications) employed a mix of GPT-4-Turbo and GPT-4.1. GPT-4.1 was set to low diversity for tagging speech and extracting information from chat histories, while GPT-4-Turbo used high diversity for generating speech. Technically, they implemented an expert tactic where a villager claims to be the seer to thwart a day-two power play in a 5-players game. Additionally, they executed machine learning-based role estimation via XGBoost using the extracted speech tags, boosting accuracy through combinatorial optimization based on game constraints.

yatolab The "yatolab" agent built by Shoma Yato, Ikumu Komatsu, Haruka Itakura, Kodai Ishizaka, Renta Kazuyama, Roki Sakamoto, Naoya Sachi, Yushin Otsuka, Harumi Ishii, Daisuke Katagami, and Yousun Kang (Tokyo Polytechnic University) accommodated both 9-players and 5-players villages, featuring a dialogue structure that separated models for the self and others while reducing load via fundamental prompt control. They mandated a fixed format to prevent voting mistakes and a unique first-person pronoun for speaker identification, while simplifying day 0 greetings to prevent baseless mutual suspicion. In the anytime track, a bidirectional model updated and estimated intentions for both self and others. For the 9-players village, a Memory function independently managed self and other memories, preventing whisper leaks and improving consistency.

yharada The "yharada" agent built by Yuya Harada (Shizuoka University) achieved "quality transfer" by leveraging Claude for generating match scripts and GPT for executing actual gameplay, maximizing both models' strengths. To counter the parroting and rigid speech caused by overly restrictive traditional prompts, they implemented "10-shot learning" using 10 authentic match scripts showing natural topic shifts and discussion flows. This allowed the LLM to independently absorb diverse perspectives and conversation dynamics

as tacit knowledge, facilitating highly flexible situational judgments and character-driven speech free from strict prompt constraints.

yshimoda The "yshimoda" agent built by Yuki Shimoda (Shizuoka University) employed a technique that extracted game states using rules and detected contradictions in divination claims by cross-verifying maps. They meticulously scrutinized dialogue by mapping the relational structure and converting each player's speech into a 5-level stance ranging from strong doubt to strong defense. This approach enabled strategic decision-making by tracking shifts in the agent's thought process, such as identifying players with inconsistent statements and analyzing connections between potential werewolves.

4 Evaluations and Results

As the number of agents does not match with the number of players, we tried to make equal agent combinations of game matches with respect to the number of games per agent, number of roles per agent as far as possible; it was virtually impossible to run all possible combinations due to the time restriction though.

We calculated win rates in different aspects such as macro-averaged, micro-averaged, weighted-average considering the number of players for each role, 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 9-players village only)

These criteria have been used for years in our AIWolfDial competition series. When these criteria were defined, the performance of the deep-learning based dialog systems, such as LSTM-based ones, was insufficient; rule-based systems even sometimes perform better. The criterion A (sentence expression naturalness) was introduced because of such poor performances at that time, which is now very fluent in most cases. While the criterion A is expression-based evaluation, we need contextual evaluation with other agents (criterion B: context) and the agent itself (criterion C: coherence). Because some agents behave different in their game actions from there talk contents, e.g. vote to agents which they apparently recognize as a teammate, we also need to evaluate the game action consistency with talks (criterion D: game). It is important not just for the werewolf game but for a dialog system in general, to make a consistent unique character traits in their talks (criterion E: character/divergence). Finally, as there are two or more werewolves who know their roles each other and communicate secretly in our recent 9- and 13-players tracks, we evaluate the team play including the villager team (F: teamplay),

4.1.1 Human Annotators

Our three annotators are required to perform subjective evaluations of rankings based on 18 game logs (11 game logs for the anytime-speak track as it costs time to consider the timestamps) for each track for each game, then we averaged over those annotator ranking scores.

4.1.2 LLM-as-a-Judge

In addition, we performed an automatic evaluation of these subjective evaluation by llm-as-a-judge way. Evaluations are conducted separately for each game and evaluation criterion, resulting in 5 to 6 LLM requests per game. In each request, the input is provided to the LLM within a single request.

The evaluation prompt consists of the developer prompt part (Table A.1.1) and the user prompt part (Table A.1.2). The developer prompt part includes instructions common to all evaluation criteria, e.g. the role of evaluating as a Werewolf game evaluator, common evaluation rules, descriptions of each field in the input log. The user prompt part includes

information that varies by evaluation criterion and game, e.g. supplementary instructions per criterion, character settings for each player, the specific criterion being evaluated, the target game log.

The input log is input for an entire single game at once (with only the result row excluded). The original CSV-formatted log is structured, converted into a JSON string, and passed into the User prompt. Additionally, to reduce judge bias, only the result row directly indicating the final game outcome is excluded. The contents of the json style log file are inserted into {{ log }} in the user prompt.

This llm-as-a-judge was performed on the same 10 games as the human annotators, and all games as well[3]. We used gemini-2.5-pro, claude-sonnet-4.5, gpt-5.4, gemma-4-31b-vllm, and qwen3-6-27b-vllm for the LLM-as-a-judges.

4.1.3 Correlation between Humans and LLMs

After averaging the llm-as-a-judge scores over these LLMs, we calculated Pearson correlations between humans, *llms_all* (all game logs), and *llms_same* (the same game logs with humans). The Pearson correlation coefficients were quite high between *llms_all* and *llms_same* regardless of the tracks, almost always > 0.9; In the 9-players track, the coefficients were always high and > 0.9 in any combination of the three; the coefficients for rankings of averages over the subjective evaluation metrics were mostly > 0.8 between humans and LLMs, while specific metrics in the 5-players track showed lower coefficients (See Appendix for detailed tables).

Totally speaking, the humans and LLMs correlates very well that we could adopt the LLM-as-a-judge as an alternative to the human judges in future.

4.1.4 Subjective Evaluation Results

In the LLM subjective evaluation across all tracks and the human subjective evaluation for the 9-players track, the **yharada** team took first place by a clear margin. They scored particularly high in B (context), C (coherence), and E (diversity).

In the human subjective evaluation for the 5-players track, the **Oxymoron** team took first place, receiving high ratings particularly in A (expression), C (coherence), and E (diversity).

In the human subjective evaluation for the anytime-speak track, the **AGKAI** team took first place, scoring particularly high in A (expression), C (coherence), and E (diversity).

4.2 Win Rates

We calculated the win rates for each track (see Appendix for the detailed tables).

Regarding win rates, the **bmd** team took first place in both the 5-players and 9-players tracks, demonstrating a high win rate when playing as Villagers.

In the anytime-speak track, the **CamelliaDragons** team took first place, maintaining an equally top-ranked win rate across all roles.

Unfortunately, there was not 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.

5 Discussion

Overall, by leveraging the improved performance of LLMs, the agents were able to engage in conversations that functioned well as a Werewolf game. The utterances themselves were natural, and there were many situations where logical reasoning with justifications and conversational exchanges were successfully established. On the other hand, various shortcomings were also observed by reviewing the actual game logs, as detailed below.

First, there are many instances where reciprocal conversation (back-and-forth dialogue) fails. In many cases, agents did not respond to utterances explicitly directed at them by name. There were also cases where they failed to grasp the immediate or prior context, neglecting to reflect crucial information from others' utterances, such as role coming-outs. This directly affects the evaluation of B (context).

Inappropriate utterances were observed, such as changing opinions immediately after stating them or repeating the exact same content. This impacts the evaluation of C (coherence).

There was also a lack of basic game state awareness related to D (game action), such as commenting that it is "quiet" before the

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

first speaking turn of the day had even arrived, or talking to players who had already been eliminated.

Contrastingly, in single-agent applications of recent LLMs, having them generate logs of multi-players Werewolf matches easily produces highly natural, human-like, and entertaining conversations. Despite this, the degradation in basic conversational performance seen in many teams was particularly prominent in the 9-players track and has remained an unresolved challenge since the advent of ChatGPT. The difficulty of multi-agent dialogue and the long contexts that constantly risk exceeding input length limits may be the causes.

The yharada team, which ranked first in all LLM evaluations and the 9-players track subjective evaluation, scored particularly high in B (context), C (coherence), and E (diversity). They solved this problem using a "script few-shot" method. Although there remains a question of whether this approach can adapt to dynamic argument developments, it suggests a promising direction for future advancement.

The AGKAI team, which ranked first in the human evaluation for the anytime-speak track, incorporated a mechanism where the AI evaluated its own self-play logs to continuously improve its prompts and code, which may have yielded a similar effect.

The Oxymoron team, which ranked first in the human evaluation for the 5-players track, utilized a two-tier structure: one module for generating utterance content and another module for applying profiles, writing styles, and role elements. This approach could also produce similarly effective results.

Regarding win rates, teams with high win rates did not necessarily score high in D (consistency between game actions and utterances). Possible reasons include the fact that many teams had a disconnect between their utterances and game actions, or random conditions due to the small number of games played (e.g., coincidentally getting good teammates or facing strong opponent teams). Further integration of game actions and utterance content is anticipated as a future development.

Logical statements and justifications were also frequently unnatural or inappropriate. In particular, we observed many instances of herd behavior, where agents simply agreed with the opinions of other players. This may also be because past utterance logs accumulated beyond the maximum length, causing the influence of the fixed prompt (which dictates the basic direction) to fade, leading the agents to be swayed by the utterance logs instead.

In the newly introduced anytime-speak track, likely because it was a first attempt, the real-time tuning of each team seemed insufficient. There were many situations where participation in the conversation was extremely slow or agents were unable to reply at all. While achieving human-level real-time performance is the ultimate goal, considering the response speeds of recent LLMs—especially the delays during the "Thinking" phase—considerable countermeasures will still be required moving forward.

6 Conclusion and Future Work

By conducting a text-based auto-play contest of the "Werewolf game", a conversational game of lying and detecting lies, we highlighted the current achievements and shortcomings of LLMs.

Reflecting the improved performance of LLMs, natural language expression and a certain level of dialogue were achieved successfully. On the other hand, challenges remained in "lying"; the simultaneous maintenance of multiple character role settings, and in handling long-context utterance histories that exceed LLM input lengths in multi-agent environments. There were instances where agents generated utterances that completely ignored their settings, others' utterances, their own past utterances, and sometimes even the game settings themselves. In addition, a conformity phenomenon was frequently observed, where other agents would successively agree with and affirm a specific utterance. Better evaluation scores for teams with a two-tier, divided role LLMs could generally become a future trend in the multi-agent utilization of LLMs.

All of the challenges brought to light in this contest could easily become immediate issues in general when deploying LLMs in real-world multi-agent environments, such as large corporate organizations. The contest's aim of highlighting the current achievements and problems of LLMs was successfully met, and future development is highly anticipated. We plan to continue holding this contest in the future.

Acknowledgments

This research was partially supported by Kakenhi, MEXT Japan (JP22H00804, JP25K21648JP, JST PRESTO (JPMJPR2461).

Limitations

The games were performed in the Japanese language only; we need to investigate other languages as well.

We need to increase the number of games to be statistically more stable.

There could be some bias when the LLM versions used in the LLM-as-a-judge evaluations and the agents are same.

More evaluation metrics would be required to capture achievements in higher levels of semantics and discourses.

Because the conversation could be degraded even if the agent itself performs well, when the other agents perform worse; games between selected agents, or games with humans, can be future options.

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.

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, 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 *Proceedings of the 3rd International AIWolfDial Workshop*, pages 1–13, Hanoi, Vietnam. 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.

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.

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.

Ahmet Burak Yildirim[1], Nezih Sunman, and Reyhan Aydoğan. 2023. Agent competition (anac 2020). *Recent Advances in Agent-Based Negotiation: Applications and Competition Challenges*, page 102.

A Appendix

A.1 LLM-Judge-Prompt

We describe the prompt (translated from original Japanese into English) used for LLM-Judge, explaining in two prompt parts, the developer

prompt part and the user prompt part[4]. The developer prompt part is intended to evaluate assumed underlying fixed system prompts, while the user prompt part is intended to evaluate the user input related texts.

A.1.1 Developer Prompt Part

This prompt part (translated from original Japanese into English) provides an explanation of the format of the logs supplied when performing the Judge task. In Section **??**, 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 2: Prompt template for the developer prompt part

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

[4]Manual for running the prompts is publicly available in https://github.com/aiwolfdial/aiwolf-nlp-llm-judge

(e.g., speaker_index → speaker).

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

A.1.2 User Prompt Part

This prompt part 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 the previous section, each JSON object, such as {"day":1, "action":"talk", ...}, is provided line by line via {{ log }}.

Table 3: 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 }}

A.2 Evaluation Results

We show the correlations between Human and LLM evaluations, the win rates, the subjective evaluation scores by humans and LLMs, for each tracks, in the following tables.

Table 4: Correlation between Human Evaluation and LLM-as-a-Judge (5-players Track)

	A	B	C	D	E	Average	Rank
Human and llm_same	0.85	0.60	0.75	0.67	0.75	0.83	0.80
Human and llm_all	0.84	0.57	0.70	0.65	0.72	0.81	0.79
llm_same and llm_all	0.98	0.95	0.90	0.84	0.97	0.96	0.92

Table 5: Correlation between Human Evaluation and LLM-as-a-Judge (9-players Track)

	A	B	C	D	E	F	Average	Rank
Human and llm_same	0.92	0.93	0.94	0.88	0.94	0.79	0.95	0.92
Human and llm_all	0.89	0.94	0.96	0.92	0.94	0.91	0.96	0.94
llm_same and llm_all	0.98	0.98	0.98	0.98	0.99	0.94	0.99	0.97

Table 6: Correlation between Human Evaluation and LLM-as-a-Judge (Anytime-speak Track)

	A	B	C	D	E	Average	Rank
Human and llm_same	0.70	0.90	0.91	0.65	0.73	0.86	0.83
Human and llm_all	0.81	0.91	0.80	0.69	0.78	0.85	0.83
llm_same and llm_all	0.96	0.90	0.93	0.91	0.92	0.96	0.93

Table 7: Summary of 5-players Track - Game Counts by Role with Rank

Team	POSSESSED	SEER	VILLAGER	WEREWOLF	TOTAL	Rank
bmd	15	15	25	19	74	1
sunamelli	17	18	25	14	74	2
redhood	16	15	24	19	74	3
ninetailed	15	15	31	14	75	4
yharada	15	15	31	14	75	5
Oxymoron	16	15	30	15	76	6
doradora_lab	15	13	30	17	75	7
CanisLupus	14	15	32	14	75	8
mrdev	14	15	32	14	75	9
yatolab	15	15	31	14	75	10
CamelliaDragons	15	15	31	15	76	11
gotsumorinlp	14	15	33	13	75	12
Luna	15	15	31	14	75	13
AGKAI	15	16	30	15	76	14
kanalab	16	15	30	14	75	15
Ty-TakutiX	12	13	32	18	75	16
noisy_watcher	16	15	31	13	75	17
maverick	15	15	31	14	75	18

Table 8: Summary of 5-players Track - Win Rates by Role (Percentages)

Team	POSSESSED	SEER	VILLAGER	WEREWOLF	Macro	Micro	Weighted Micro
bmd	46.67	80	72	63.16	66.22	65.46	68.01
sunamelli	64.71	66.67	64	64.29	64.87	64.92	64.39
redhood	50	80	58.33	57.89	60.81	61.56	59.42
ninetailed	53.33	53.33	58.06	57.14	56	55.46	56.95
yharada	20	66.67	61.29	57.14	53.33	51.28	56.89
Oxymoron	25	40	63.33	60	50	47.08	56.82
doradora_lab	46.67	61.54	46.67	58.82	52	53.42	51.34
CanisLupus	42.86	33.33	59.38	42.86	48	44.61	51
mrdev	57.14	53.33	50	50	52	52.62	50.95
yatolab	26.67	60	54.84	35.71	46.67	44.30	47.53
CamelliaDragons	46.67	40	51.61	40	46.05	44.57	46.94
gotsumorinlp	78.57	33.33	48.48	30.77	48	47.79	45.01
Luna	46.67	40	35.48	64.29	44	46.61	44.77
AGKAI	46.67	50	43.33	40	44.74	45	43.33
kanalab	37.50	46.67	43.33	42.86	42.67	42.59	42.98
Ty-TakutiX	58.33	46.15	37.50	44.44	44	46.60	42.07
noisy_watcher	68.75	46.67	48.39	15.38	46.67	44.80	41.08
maverick	53.33	33.33	45.16	28.57	41.33	40.10	40.30

Table 9: Subjective Evaluation of 5-players Track - Human

Team	A	B	C	D	E	Average	Rank
Oxymoron	1.73	1.60	1.47	2.07	1.67	1.71	1
sunamelli	2.07	2.13	2	2.67	2.40	2.25	2
bmd	2	1.87	2.47	2.13	3.13	2.32	3
yharada	2.42	2.33	2.25	2.83	2.08	2.38	4
doradora_lab	2.13	2.67	2.47	2.33	2.67	2.45	5
mrdev	3.39	2.50	2.50	1.89	2.33	2.52	6
CamelliaDragons	3.20	2.60	2.93	2.33	2.73	2.76	7
noisy_watcher	2.80	2.80	2.87	3.40	3.67	3.11	8
redhood	3.33	3.33	3.28	2.94	2.89	3.16	9
maverick	2.89	3.44	3	3.44	3.17	3.19	10
Luna	2.67	3.17	3.72	3.33	3.11	3.20	11
CanisLupus	3.40	3.47	3.53	3.33	2.67	3.28	12
ninetailed	3.80	3.67	2.73	3.07	3.53	3.36	13
yatolab	3.33	3.50	3.42	4	3.83	3.62	14
Ty-TakutiX	3.78	3.83	3.83	3.72	2.94	3.62	15
gotsumorinlp	3.25	3.58	4	3.92	4	3.75	16
AGKAI	3.75	3.92	4.08	3.58	3.75	3.82	17
kanalab	4.25	3.75	3.67	3.50	4.08	3.85	18

Table 10: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - gemma-4-31b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	3.67	2	2.67	3	2.67	2.80	12
CamelliaDragons	2.80	3.20	1.80	1	2.60	2.28	6
CanisLupus	2.40	4	1.80	2	2.20	2.48	8
Luna	1.83	2.83	2.50	2	4	2.63	9
Oxymoron	1	2.20	1	2	2.40	1.72	3
Ty-TakutiX	4.60	4.80	4.80	4.80	4.60	4.72	18
bmd	2.60	2.40	2	1.60	3	2.32	7
doradora_lab	2	1.75	2	2	2	1.95	5
gotsumorinlp	4	3.67	3.67	3	4	3.67	16
kanalab	5	5	4.75	3.25	5	4.60	17
maverick	3.17	3.50	2.33	1.83	3.50	2.87	13
mrdev	1.67	1.83	1.67	2	1.33	1.70	2
ninetailed	3.40	3.20	1.60	1.80	3.20	2.64	10
noisy_watcher	3.40	4.20	3.60	2.60	4.40	3.64	15
redhood	3.83	2	2.67	3.67	1.83	2.80	11
sunamelli	1	1.80	1.60	1.60	2.20	1.64	1
yatolab	2.67	3.67	2.33	5	4	3.53	14
yharada	1.50	2.50	1.75	1.50	2	1.85	4

Table 11: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - gemini-2.5-pro

Team	A	B	C	D	E	Average	Rank
AGKAI	3	1.75	2.50	3.50	1.75	2.50	8
CamelliaDragons	3.20	3.40	2.20	1.60	3.20	2.72	10
CanisLupus	2.80	3.60	3	1	1.40	2.36	7
Luna	3	3.50	3.17	3.17	3.33	3.23	13
Oxymoron	1.40	2	1.40	3.20	2.20	2.04	5
Ty-TakutiX	4.33	4.67	4.67	4	4.50	4.43	17
bmd	1.60	1.60	2.20	1.60	2.40	1.88	4
doradora_lab	2	1.40	1.80	2.60	3.20	2.20	6
gotsumorinlp	4.25	4	3.75	2.50	4.25	3.75	16
kanalab	5	5	4.75	3.50	5	4.65	18
maverick	3	3.67	3.50	2	3.17	3.07	12
mrdev	2	1.33	2.17	1.33	1.67	1.70	2
ninetailed	3.40	2.40	2.80	2	3.20	2.76	11
noisy_watcher	4.20	4	3.40	3.20	3.80	3.72	15
redhood	2.83	2.33	3.17	2.50	2.17	2.60	9
sunamelli	1.80	1.80	1.80	1	2.60	1.80	3
yatolab	3.75	3.25	4	3	3.50	3.50	14
yharada	2	2.25	1.75	1	1.25	1.65	1

Table 12: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - gpt-5.4

Team	A	B	C	D	E	Average	Rank
AGKAI	3	1.50	3	3.50	1.50	2.50	8
CamelliaDragons	2.20	3.40	2.60	1.20	2.80	2.44	7
CanisLupus	3.80	4.40	3.60	1.80	3.20	3.36	11
Luna	2.50	3.17	3	3.83	4	3.30	10
Oxymoron	1.20	2.60	2	2.60	1.80	2.04	5
Ty-TakutiX	4.67	4.33	4.67	3.67	4.33	4.33	18
bmd	1.80	1.40	2.40	3	2.40	2.20	6
doradora_lab	1.80	1.60	1.80	1.60	2.40	1.84	3
gotsumorinlp	3.25	4.25	3.25	4.25	3.75	3.75	14
kanalab	5	4.75	4.75	2.75	4	4.25	16
maverick	3.50	3.67	3.83	2.33	3.50	3.37	12
mrdev	2.83	1.17	2.33	2.33	1.50	2.03	4
ninetailed	3.60	2.60	2.40	2.60	4.20	3.08	9
noisy_watcher	3.60	4.20	3.80	3.20	4.20	3.80	15
redhood	4.17	3	3.17	3.83	2.67	3.37	13
sunamelli	1.40	2	1.80	1.80	1.60	1.72	2
yatolab	4	4.25	3.75	4.75	4.50	4.25	16
yharada	1.50	2	1.50	1.50	1.50	1.60	1

Table 13: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - claude-sonnet-4-5

Team	A	B	C	D	E	Average	Rank
AGKAI	4	2.50	3.75	4.25	2.25	3.35	13
CamelliaDragons	2.80	3	2	2.20	2.80	2.56	8
CanisLupus	2.60	3	2.20	1.80	2	2.32	7
Luna	3	3	2.33	4	3.33	3.13	12
Oxymoron	1.20	1.80	1	3	1.80	1.76	2
Ty-TakutiX	4.17	4.50	4.17	2.67	4	3.90	17
bmd	1.40	1.80	2.40	1.40	2	1.80	3
doradora_lab	1	1.80	2	2.60	2.60	2	5
gotsumorinlp	3.50	4	3	2.50	4	3.40	14
kanalab	4.75	5	5	3	5	4.55	18
maverick	3	3.67	2.83	1.67	4	3.03	10
mrdev	2.83	2.17	1.67	2.83	1.17	2.13	6
ninetailed	3.40	3.60	2.60	2.20	3.80	3.12	11
noisy_watcher	4.40	4	3	3.20	4	3.72	16
redhood	3.50	2.50	2.83	3.33	2.17	2.87	9
sunamelli	1.80	2	1.60	1.40	2.40	1.84	4
yatolab	2.75	3.50	3.25	4	4.25	3.55	15
yharada	1.50	2	1	1	1.50	1.40	1

Table 14: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	3.50	1	2	3.25	2.50	2.45	5
CamelliaDragons	3.20	3.80	2.80	3.20	2.80	3.16	11
CanisLupus	2.40	3.60	2.40	3.60	2	2.80	8
Luna	3.67	3.33	3.17	2.83	4.17	3.43	14
Oxymoron	1.40	2.20	2.20	2.80	1.40	2.00	3
Ty-TakutiX	4.33	4.83	4.67	3.67	3.33	4.17	17
bmd	2.40	2.40	2.80	2.40	3	2.60	7
doradora_lab	2.40	2.60	2.60	1.80	2.80	2.44	4
gotsumorinlp	3.25	3.25	3.25	3	3.75	3.30	12
kanalab	4.75	3.75	4.25	3.75	4	4.10	16
maverick	2.83	3.50	3	3.50	3.67	3.30	12
mrdev	1.83	1.67	2.33	1.33	1.83	1.80	2
ninetailed	3.40	2.60	2.40	1.80	4	2.84	9
noisy_watcher	3.60	3.20	4	3.20	3.20	3.44	15
redhood	3	3	3.33	3.83	2.33	3.10	10
sunamelli	1.60	1.40	1.80	1	2.80	1.72	1
yatolab	4.50	4.75	4	4.50	4	4.35	18
yharada	2.25	3	2.75	2.25	2.50	2.55	6

Table 15: Subjective Evaluation Results by LLM-as-a-judge (Same) of 5-players Track - Average over LLMs

Team	A	B	C	D	E	Average	Rank
AGKAI	3.43	1.75	2.78	3.50	2.13	2.72	9
CamelliaDragons	2.84	3.36	2.28	1.84	2.84	2.63	7
CanisLupus	2.80	3.72	2.60	2.04	2.16	2.66	8
Luna	2.80	3.17	2.83	3.17	3.77	3.15	13
Oxymoron	1.24	2.16	1.52	2.72	1.92	1.91	4
Ty-TakutiX	4.42	4.63	4.59	3.76	4.15	4.31	17
bmd	1.96	1.92	2.36	2	2.56	2.16	6
doradora_lab	1.84	1.83	2.04	2.12	2.60	2.09	5
gotsumorinlp	3.65	3.83	3.38	3.05	3.95	3.57	14
kanalab	4.90	4.70	4.70	3.25	4.60	4.43	18
maverick	3.10	3.60	3.10	2.27	3.57	3.13	12
mrdev	2.23	1.63	2.03	1.97	1.50	1.87	3
ninetailed	3.44	2.88	2.36	2.08	3.68	2.89	10
noisy_watcher	3.84	3.92	3.56	3.08	3.92	3.66	15
redhood	3.47	2.57	3.03	3.43	2.23	2.95	11
sunamelli	1.52	1.80	1.72	1.36	2.32	1.74	1
yatolab	3.53	3.88	3.47	4.25	4.05	3.84	16
yharada	1.75	2.35	1.75	1.45	1.75	1.81	2

Table 16: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - claude-sonnet-4-5

Team	A	B	C	D	E	Average	Rank
AGKAI	3.99	2.57	3.21	3.58	2.22	3.11	12
CamelliaDragons	2.45	3.24	2.53	2.38	2.51	2.62	9
CanisLupus	2.19	3.16	2.23	2.19	2.08	2.37	7
Luna	2.37	3.35	2.56	3.23	4.11	3.12	13
Oxymoron	1.37	2.83	1.76	2.58	2.05	2.12	6
Ty-TakutiX	4	4.45	4.20	3.39	3.89	3.99	17
bmd	1.72	1.88	1.69	1.46	2.18	1.78	2
doradora_lab	1.39	1.75	2.07	1.99	2.55	1.95	3
gotsumorinlp	3.93	4.12	3.81	2.56	4.29	3.74	16
kanalab	4.56	4.31	4.32	3.12	4.31	4.12	18
maverick	2.81	3.84	3.39	2.19	3.61	3.17	14
mrdev	2.39	1.71	1.59	2.36	2.12	2.03	4
ninetailed	3.40	3.07	1.99	3.15	3.51	3.02	10
noisy_watcher	3.85	3.64	2.88	3.60	3.87	3.57	15
redhood	3.72	2.16	1.88	3.19	1.92	2.57	8
sunamelli	2.01	2.28	1.78	2.04	2.28	2.08	5
yatolab	3.07	3.28	1.80	3.03	3.96	3.03	11
yharada	1.27	1.96	1.40	1.35	1.69	1.53	1

Table 17: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - gpt-5.4

Team	A	B	C	D	E	Average	Rank
AGKAI	3.41	2.04	3.04	3.36	2.36	2.84	9
CamelliaDragons	2.67	3.38	2.99	1.95	2.54	2.71	7
CanisLupus	3.08	3.81	3.41	2.61	2.48	3.08	11
Luna	2.87	3.53	3.11	3.23	4.19	3.38	13
Oxymoron	1.72	2.88	2.28	3.12	2.32	2.46	6
Ty-TakutiX	4.69	4.32	4.64	3.77	4.32	4.35	18
bmd	1.68	1.88	2.14	1.85	1.88	1.88	2
doradora_lab	1.73	1.67	2.28	2.28	2.41	2.07	3
gotsumorinlp	3.48	3.77	3.73	3.48	4.32	3.76	16
kanalab	4.69	4.27	4.47	3.65	4.04	4.22	17
maverick	3.44	3.87	3.95	2.05	3.60	3.38	12
mrdev	3.08	1.77	2.32	2.27	1.99	2.29	5
ninetailed	3.28	2.55	2.51	3.33	3.68	3.07	10
noisy_watcher	3.49	3.77	3.43	3.93	3.84	3.69	15
redhood	3.70	2.36	2.66	3.14	2.18	2.81	8
sunamelli	1.77	2.39	2.18	2.15	2.18	2.13	4
yatolab	3.61	3.83	2.93	3.49	3.96	3.57	14
yharada	1.52	1.88	1.72	1.61	1.72	1.69	1

Table 18: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - gemma-4-31b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	3.86	2.53	2.47	3.17	2.49	2.90	11
CamelliaDragons	2.78	3.42	2.34	2	2.43	2.59	9
CanisLupus	2.56	3.32	2.29	1.67	2.19	2.41	8
Luna	2.57	3.39	2.08	3	4.13	3.03	13
Oxymoron	1.59	2.80	1.64	2.49	2.28	2.16	6
Ty-TakutiX	4.41	4.39	4.41	3.28	4.13	4.13	17
bmd	2	1.92	1.58	1.38	2.34	1.84	3
doradora_lab	1.47	1.96	1.89	1.63	2.67	1.92	4
gotsumorinlp	3.67	3.91	3.31	2.64	4.43	3.59	16
kanalab	4.79	4.59	4.35	3.65	4.65	4.41	18
maverick	3.12	3.83	3.33	1.92	3.47	3.13	14
mrdev	1.96	1.61	1.60	1.93	1.80	1.78	2
ninetailed	3.40	2.77	1.80	2.56	3.37	2.78	10
noisy_watcher	3.63	3.65	2.97	3.47	3.83	3.51	15
redhood	3.31	2.05	1.82	2.73	1.82	2.35	7
sunamelli	1.99	2.34	1.55	1.54	2.51	1.99	5
yatolab	3.17	3.37	2.07	2.87	3.43	2.98	12
yharada	1.52	2.12	1.20	1.20	1.99	1.61	1

Table 19: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - gemini-2.5-pro

Team	A	B	C	D	E	Average	Rank
AGKAI	3.82	2.67	3.29	3.36	2.46	3.12	12
CamelliaDragons	2.83	3.30	2.84	2.32	2.54	2.77	9
CanisLupus	2.33	3.20	2.64	1.73	1.97	2.38	7
Luna	2.81	3.28	3.24	3.07	3.87	3.25	14
Oxymoron	1.63	2.46	2.13	2.50	2.03	2.15	6
Ty-TakutiX	4.19	4.36	4.43	3.37	4.05	4.08	17
bmd	1.72	1.99	1.85	1.57	2.16	1.86	3
doradora_lab	1.51	1.88	2.13	1.81	2.59	1.98	4
gotsumorinlp	4.36	3.99	3.77	2.67	4.25	3.81	16
kanalab	4.69	4.53	4.41	3.69	4.55	4.38	18
maverick	2.87	3.76	3.41	2.04	3.35	3.09	11
mrdev	2.05	1.52	1.47	1.80	1.77	1.72	2
ninetailed	3.37	2.84	2.60	3.08	3.44	3.07	10
noisy_watcher	3.41	3.59	3.07	3.57	3.57	3.44	15
redhood	2.66	2.05	2.42	2.96	1.84	2.39	8
sunamelli	2.19	2.49	2.01	1.78	2.26	2.15	5
yatolab	3.47	3.31	2.55	2.83	3.84	3.20	13
yharada	1.41	1.97	1.51	1.29	1.69	1.58	1

Table 20: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	2.89	2.20	2.36	2.96	2.39	2.56	8
CamelliaDragons	3.14	3.50	3.05	2.45	2.91	3.01	10
CanisLupus	3.27	3.45	2.96	2.84	2.73	3.05	11
Luna	3.09	3.59	3.12	3.03	4	3.37	13
Oxymoron	2.12	2.79	2.41	2.80	2.50	2.52	7
Ty-TakutiX	4.01	4.20	4.08	3.51	3.69	3.90	17
bmd	2.20	2.14	2.41	2.27	2.20	2.24	3
doradora_lab	2.20	2.17	2.27	2.57	2.60	2.36	5
gotsumorinlp	3.09	3.05	3.40	2.95	3.77	3.25	12
kanalab	4.55	3.87	4.04	3.28	3.93	3.93	18
maverick	3.45	3.84	3.76	2.99	3.49	3.51	15
mrdev	2.63	1.80	2.11	2.19	2.09	2.16	2
ninetailed	2.79	2.69	2.39	2.83	3.24	2.79	9
noisy_watcher	3.48	3.87	3.60	3.19	3.96	3.62	16
redhood	2.50	2.54	2.58	2.86	2.03	2.50	6
sunamelli	2.26	2.47	2.36	2.36	2.31	2.35	4
yatolab	3.51	3.36	3.44	3.20	3.55	3.41	14
yharada	1.87	2.17	2.04	2.23	2.36	2.13	1

Table 21: Subjective Evaluation Results by LLM-as-a-judge (All) of 5-players Track - Average over LLMs

Team	A	B	C	D	E	Average	Rank
AGKAI	3.59	2.40	2.87	3.28	2.38	2.91	10
CamelliaDragons	2.77	3.37	2.75	2.22	2.59	2.74	9
CanisLupus	2.69	3.39	2.71	2.21	2.29	2.66	8
Luna	2.74	3.43	2.82	3.11	4.06	3.23	12
Oxymoron	1.69	2.75	2.04	2.70	2.23	2.28	6
Ty-TakutiX	4.26	4.34	4.35	3.46	4.02	4.09	17
bmd	1.86	1.96	1.93	1.71	2.15	1.92	2
doradora_lab	1.66	1.89	2.13	2.06	2.56	2.06	4
gotsumorinlp	3.71	3.77	3.61	2.86	4.21	3.63	16
kanalab	4.66	4.31	4.32	3.48	4.30	4.21	18
maverick	3.14	3.83	3.57	2.24	3.50	3.25	14
mrdev	2.42	1.68	1.82	2.11	1.95	2.00	3
ninetailed	3.25	2.78	2.26	2.99	3.45	2.95	11
noisy_watcher	3.57	3.70	3.19	3.55	3.81	3.57	15
redhood	3.18	2.24	2.27	2.98	1.96	2.52	7
sunamelli	2.04	2.39	1.98	1.98	2.31	2.14	5
yatolab	3.37	3.43	2.56	3.08	3.75	3.24	13
yharada	1.52	2.02	1.57	1.54	1.89	1.71	1

Table 22: Win Rates of 9-players Track (B: BODYGUARD, M: MEDIUM, P: POSSESSED, S: SEER, V: VILLAGER, W: WEREWOLF)

Team	B	M	P	S	V	W	TOTAL	Macro	Micro	Weighted Micro
bmd	50	100	50	50	83.33	75	19	73.68	68.05	75
yshimoda	100	50	100	100	66.67	66.67	17	76.47	80.56	73.08
Oxymoron	66.67	50	50	50	83.33	75	19	68.42	62.50	72.43
mrdev	50	50	100	100	66.67	50	18	66.67	69.45	65.39
yharada	50	100	50	0	60	80	18	61.11	56.67	61.54
sunamelli	50	50	50	50	33.33	100	18	55.55	55.55	53.84
doradora_lab	0	0	100	0	71.43	50	17	52.94	36.91	52.20
yatolab	100	50	0	50	42.86	66.67	17	47.06	51.59	50.55
AGKAI	0	50	50	100	33.33	60	19	47.37	48.89	44.61
CamelliaDragons	50	0	0	66.67	60	33.33	17	41.18	35	44.36
redhood	50	100	50	0	33.33	50	18	44.44	47.22	42.31
ninetailed	50	0	50	100	33.33	25	17	35.29	43.05	36.54
noisy_watcher	50	100	50	50	16.67	20	19	36.84	47.78	31.54
kanalab	66.67	0	50	50	33.33	0	19	31.58	33.33	28.20
CanisLupus	0	0	0	0	33.33	0	18	11.11	5.55	15.38

Table 23: Subjective Evaluation Results of 9-players Track - Human

Team	A	B	C	D	E	F	Average	Rank
yharada	2.37	2.70	2.43	2.67	1.87	2.97	2.50	1
sunamelli	3.22	3.41	3.07	3.15	3.56	3.33	3.29	2
mrdev	3.55	2.97	3.67	3.24	4.12	3.00	3.42	3
Oxymoron	3.52	3.79	3.52	4.06	3.36	4.06	3.72	4
yshimoda	3.64	4.67	4.12	4.52	3.55	4.67	4.19	5
doradora_lab	4.83	4	5.07	4.60	5.13	4.30	4.66	6
bmd	4.85	4.70	5.15	4.58	4.64	4.33	4.71	7
AGKAI	5.13	4.62	5.31	4.51	6.46	4.67	5.12	8
noisy_watcher	4.55	5.00	5.42	5.55	5.00	5.42	5.16	9
CanisLupus	5.63	5.56	5.37	5.44	4.30	5.44	5.29	10
ninetailed	6	6.10	5.13	5.50	6	5.93	5.78	11
redhood	5.79	6.15	5.82	6.00	5.54	6.28	5.93	12
yatolab	6.08	5.78	5.61	6.39	6.69	5.56	6.02	13
CamelliaDragons	6.78	6.52	6.30	5.96	5.74	6.15	6.24	14
kanalab	8.44	8.42	8.19	8.06	7.83	8.19	8.19	15

Table 24: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - gemini-2.5-pro

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.54	3.23	5.85	3.46	4.54	3.31	3.99	8
CamelliaDragons	5.89	5	5.11	4	5.22	5.44	5.11	10
CanisLupus	4.22	4.56	3.56	4.78	2.22	4.78	4.02	9
Oxymoron	3.91	3.36	3.09	4.82	2.36	4.82	3.73	5
bmd	3.82	3	4.45	3.45	3.64	3.82	3.70	4
doradora_lab	4.70	2.70	4.80	3.40	3.10	4.70	3.90	6
kanalab	8.33	7.92	8.42	7	8.58	7.42	7.94	15
mrdev	3.18	1.82	2.82	2.91	3.27	2.45	2.74	3
ninetailed	5.20	3.80	6.10	5.90	6.70	6.30	5.67	13
noisy_watcher	4.55	5.73	4.91	4.64	6	5.45	5.21	11
redhood	5.38	5.62	5	6.15	4.23	6.54	5.49	12
sunamelli	1.44	2.89	2.78	2	3.67	3	2.63	2
yatolab	4.92	6.33	6.33	6.58	5.17	4.83	5.69	14
yharada	2.90	2.20	2.50	3.60	1.70	2.50	2.57	1
yshimoda	4	2.91	3.36	4.91	3.18	5.09	3.91	7

Table 25: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - gemma-4-31b-vllm

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.85	4.08	3.77	3.23	6	3.46	4.06	7
CamelliaDragons	4.33	7.33	3.78	3.44	4.89	5.67	4.91	11
CanisLupus	2.78	6.22	4.11	4.67	3.22	4.89	4.31	9
Oxymoron	1.91	5.36	2.55	3.45	4.36	5.73	3.89	6
bmd	4.36	3.36	2.27	2.45	5.09	4.18	3.62	5
doradora_lab	2.60	3	3.50	3.10	5	4	3.53	4
kanalab	8.67	8.42	7.42	7.42	8.17	5.67	7.63	15
mrdev	2.64	2.45	2.36	3	3.55	4	3	3
ninetailed	6.30	5.50	2.40	5.40	5.70	5.40	5.12	13
noisy_watcher	5.18	4.91	3.82	5.73	5.55	3.45	4.77	10
redhood	5.15	6.46	4	5.38	5.15	6.38	5.42	14
sunamelli	3.44	3.33	1.33	1.56	3.44	4	2.85	2
yatolab	5.17	4.92	2.83	5.50	7.33	4.83	5.10	12
yharada	2.20	3.80	1.90	1.40	2.20	4.40	2.65	1
yshimoda	2.45	5.45	3.45	3.36	3.82	5.91	4.08	8

Table 26: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - claude-sonnet-4.5

Team	A	B	C	D	E	F	Average	Rank
AGKAI	2.92	4.85	3.54	3.23	6.38	3.38	4.05	8
CamelliaDragons	4.11	7.22	5.22	2.33	3.78	5.89	4.76	11
CanisLupus	2.89	5.44	3.89	4.56	2.89	5.44	4.19	9
Oxymoron	1.45	5	2.91	3.64	4.18	5.18	3.73	5
bmd	3.09	4	3.73	2.45	4.91	4.45	3.77	6
doradora_lab	1.70	3.30	4.70	3.20	5.60	4.30	3.80	7
kanalab	8.50	7.92	7.50	6.08	8	6.50	7.42	15
mrdev	3.36	2.91	2.18	4.64	3.64	3.27	3.33	3
ninetailed	4	6.70	2.40	4.90	6	4.90	4.82	12
noisy_watcher	4.36	4.55	3.82	6.45	5.55	3.45	4.70	10
redhood	4.85	5.46	4.62	6	5.15	6.31	5.40	14
sunamelli	2.89	2.22	2.67	2.56	3.11	2.78	2.70	2
yatolab	4.08	4.83	4.42	6.42	6.50	4.58	5.14	13
yharada	2	2.90	2	1.40	1.60	3.70	2.27	1
yshimoda	1.09	5	2.73	4.09	3.18	4.64	3.45	4

Table 27: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - gpt-5.4

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.54	4.31	4.69	2.69	6.23	3.31	4.13	4
CamelliaDragons	6.22	6.89	6.44	2.56	5.44	5	5.43	10
CanisLupus	6.56	6.22	5.56	5	4.11	5.44	5.48	12
Oxymoron	3.09	5	4	4.82	4.27	5.91	4.52	7
bmd	4.64	2.91	4.27	4.36	4.45	4.27	4.15	5
doradora_lab	4.20	3	4.90	5	4.20	3.90	4.20	6
kanalab	8.42	8.42	8.42	7.67	8.42	5.75	7.85	15
mrdev	5.36	2.45	3.82	2.73	3.45	3.27	3.52	3
ninetailed	6.10	5.10	4.70	5.20	6.50	5.20	5.47	11
noisy_watcher	6.09	5.55	5.36	5.91	5.27	4.36	5.42	9
redhood	6.62	6.69	5.77	6.15	5.08	6.54	6.14	14
sunamelli	2.44	3	2.89	2.67	2.56	4.44	3	2
yatolab	5.33	5.50	4.17	6.75	7.17	5.75	5.78	13
yharada	1.80	2.60	2.70	2.30	1.80	3.80	2.50	1
yshimoda	4	6.55	5.64	4.73	4.45	7	5.39	8

Table 28: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	F	Average	Rank
AGKAI	4.62	3	3.31	2.85	5.77	3.31	3.81	3
CamelliaDragons	5.11	6.78	4.78	5.89	4.89	5.33	5.46	12
CanisLupus	6	4.33	4.89	5.67	4.44	5.56	5.15	10
Oxymoron	5.64	5.27	4.55	4.09	4.64	5.27	4.91	8
bmd	4.64	4	4	4.55	3.64	4.45	4.21	5
doradora_lab	5	4.50	4.70	4.30	3.80	3.60	4.32	6
kanalab	8.33	8.08	7.67	6.58	7	5.33	7.17	15
mrdev	2.64	3.27	4.18	2.73	3.73	3.55	3.35	1
ninetailed	4	5.10	3.70	4.40	5	4.50	4.45	7
noisy_watcher	5.18	5	5.91	5.55	6.09	5.45	5.53	13
redhood	6	6.31	5.15	5.77	5.54	5.08	5.64	14
sunamelli	3.33	3.78	3.67	3.11	4.22	3.11	3.54	2
yatolab	5.17	5.08	4.92	4.42	6.42	4.92	5.15	11
yharada	2.60	4.30	5.10	4.20	3.90	3.90	4.00	4
yshimoda	4.64	5.64	3.82	5.73	4.36	5.36	4.92	9

Table 29: Subjective Evaluation Results by LLM-as-a-judge (Same) of 9-players Track - Average over LLMs

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.69	3.89	4.23	3.09	5.78	3.35	4.01	6
CamelliaDragons	5.13	6.64	5.07	3.64	4.84	5.47	5.13	12
CanisLupus	4.49	5.36	4.40	4.93	3.38	5.22	4.63	9
Oxymoron	3.20	4.80	3.42	4.16	3.96	5.38	4.15	7
bmd	4.11	3.45	3.75	3.45	4.35	4.24	3.89	4
doradora_lab	3.64	3.30	4.52	3.80	4.34	4.10	3.95	5
kanalab	8.45	8.15	7.88	6.95	8.03	6.13	7.60	15
mrdev	3.44	2.58	3.07	3.20	3.53	3.31	3.19	3
ninetailed	5.12	5.24	3.86	5.16	5.98	5.26	5.10	10
noisy_watcher	5.07	5.15	4.76	5.65	5.69	4.44	5.13	11
redhood	5.60	6.11	4.91	5.89	5.03	6.17	5.62	14
sunamelli	2.71	3.04	2.67	2.38	3.40	3.47	2.94	2
yatolab	4.93	5.33	4.53	5.93	6.52	4.98	5.37	13
yharada	2.30	3.16	2.84	2.58	2.24	3.66	2.80	1
yshimoda	3.24	5.11	3.80	4.56	3.80	5.60	4.35	8

Table 30: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - gemini-2.5-pro

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.79	3.53	6.21	3.53	5	3.89	4.32	8
CamelliaDragons	5.24	5.47	5.24	4.94	4.12	5.71	5.12	10
CanisLupus	4.72	4.44	4	4.61	3.67	4.67	4.35	9
Oxymoron	3	3.32	2.89	4.11	2.53	4.37	3.37	4
bmd	3.53	3.11	4.21	3.74	3.84	3.63	3.68	5
doradora_lab	3.94	3.35	4.29	3.47	2.76	4.41	3.71	6
kanalab	8.42	7.89	8.37	7.05	8.63	7.37	7.96	15
mrdev	2.83	2.56	2.72	2.94	3.50	2.94	2.92	2
ninetailed	6.12	4.71	6.06	5.76	7.06	6.59	6.05	14
noisy_watcher	4.32	5.37	5.32	5.74	5.32	5.47	5.25	11
redhood	5.78	6	5.22	5.89	5.28	6.39	5.76	12
sunamelli	2.17	2.50	3.17	2.61	3.44	3.61	2.92	3
yatolab	5.18	6.35	6.76	6.88	5.41	4.76	5.89	13
yharada	2.11	2.06	2.28	2.94	1.78	2.89	2.34	1
yshimoda	4.24	2.82	3.65	4.53	3.41	4.88	3.92	7

Table 31: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - gemma-4-31b-vllm

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.47	4.32	2.89	3	6.37	3.89	3.99	8
CamelliaDragons	4.35	6.94	4.12	3.82	5	5.53	4.96	11
CanisLupus	3.72	6.56	4.56	5.33	2.94	4.61	4.62	9
Oxymoron	1.79	5	1.89	2.63	4.05	5	3.39	5
bmd	4.21	3.63	2.11	2.79	5.16	4.37	3.71	6
doradora_lab	2.35	2.53	2.71	3.35	4.94	4.18	3.34	4
kanalab	8.58	8.37	7.63	7	8.32	6.11	7.67	15
mrdev	2.61	2.50	1.83	2.94	3.33	3.89	2.85	3
ninetailed	6.53	6	2.82	4.47	6	4.88	5.12	12
noisy_watcher	4.95	4.95	3.42	5.05	5.58	4.58	4.75	10
redhood	5.28	6.50	4	5.72	5.33	5.89	5.45	14
sunamelli	3.28	3.06	1.17	1.56	3.67	4.06	2.80	2
yatolab	5.29	5.41	3.12	5.12	7.53	5	5.25	13
yharada	2.39	3.67	1.50	1.78	2.83	4.11	2.71	1
yshimoda	2.41	5.53	2.94	3.29	3.82	5.47	3.91	7

Table 32: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - claude-sonnet-4-5

Team	A	B	C	D	E	F	Average	Rank
AGKAI	2.74	4.68	4.16	3.58	6.53	3.63	4.22	8
CamelliaDragons	3.24	6.94	4.59	3.94	3.65	5.94	4.72	10
CanisLupus	3.28	6.28	3.83	3.83	3.56	5.44	4.37	9
Oxymoron	1.53	4.58	2.42	3.84	3.63	4.42	3.40	4
bmd	2.89	3.79	3.37	2.95	4.68	4.74	3.74	7
doradora_lab	2.12	2.65	3.76	4	5.06	3.76	3.56	5
kanalab	8.58	7.95	7.68	5.79	8.26	6.58	7.47	15
mrdev	3.89	2.78	2.61	4.67	3.06	3.11	3.35	3
ninetailed	4.94	6.59	2.59	4.94	6.35	4.94	5.06	12
noisy_watcher	5.11	4.11	3.42	6.05	5.21	4.68	4.76	11
redhood	5.06	5.72	4.89	6.06	5.44	6.11	5.55	14
sunamelli	2.61	2.50	2.17	2.78	3.33	3.44	2.81	2
yatolab	4.06	5.35	4.94	6.71	6.71	5.12	5.48	13
yharada	1.78	3.33	1.56	1.28	2.22	3.72	2.31	1
yshimoda	2.18	5.35	2.47	4.06	3.53	4.82	3.74	6

Table 33: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - gpt-5.4

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.89	4.53	4.89	3.21	6.58	3.58	4.45	7
CamelliaDragons	5.82	6.53	5.82	2.65	4.88	5.47	5.20	9
CanisLupus	6.61	6.56	6.33	5.50	4.56	5.67	5.87	12
Oxymoron	3.16	4.84	3.74	4.21	3.84	5.37	4.19	6
bmd	4.21	3.05	4.53	3.84	4.11	4.21	3.99	4
doradora_lab	4	2.76	4.71	5.18	4.41	3.59	4.11	5
kanalab	8.47	8.37	8.47	7.32	8.37	6.53	7.92	15
mrdev	5.72	2.56	4	3.22	3.39	3.28	3.69	3
ninetailed	6.59	5.82	5.18	4.94	6.65	5.12	5.72	11
noisy_watcher	5.68	5.26	5	6.21	5.95	5.37	5.58	10
redhood	6.50	6.56	5.94	6.28	5.44	6.56	6.21	14
sunamelli	2.28	3.11	2.56	2.94	2.78	4.22	2.98	2
yatolab	5.47	5.88	5	6.82	7.47	5.88	6.09	13
yharada	2.11	2.67	2.67	2.11	2	3.56	2.52	1
yshimoda	4.29	6.29	5.12	4.18	4.41	5.94	5.04	8

Table 34: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	F	Average	Rank
AGKAI	4.58	3.95	3.84	3.21	6	3.89	4.25	6
CamelliaDragons	5.12	6.71	5.29	6.35	5.24	5.59	5.72	14
CanisLupus	6.22	5.39	5.61	5.56	4.56	5.89	5.54	12
Oxymoron	5	4.74	4.11	3.89	4.05	4.58	4.39	7
bmd	4.11	4	4.11	4.32	3.53	4.53	4.10	5
doradora_lab	4	3.47	4.18	4.06	4	3.12	3.80	4
kanalab	8.37	7.89	7.95	6.37	7	5.47	7.18	15
mrdev	3.22	3.17	3.56	2.94	3.94	3.33	3.36	1
ninetailed	5.12	5.82	4.24	4.71	5.29	4.59	4.96	9
noisy_watcher	5.53	4.74	5.63	5.74	5.58	4.84	5.34	10
redhood	6.17	6.22	5.39	5.83	5.44	5.17	5.70	13
sunamelli	3.06	3.61	3.39	3.56	4.17	3.72	3.58	2
yatolab	5.12	5.47	4.76	4.47	6.59	5.76	5.36	11
yharada	2.89	3.89	4.06	3.67	4.17	3.72	3.73	3
yshimoda	4.76	5.47	3.88	5.53	4.82	4.65	4.85	8

Table 35: Subjective Evaluation Results by LLM-as-a-judge (All) of 9-players Track - Average over LLMs

Team	A	B	C	D	E	F	Average	Rank
AGKAI	3.69	4.20	4.40	3.31	6.09	3.78	4.25	7
CamelliaDragons	4.75	6.52	5.01	4.34	4.58	5.65	5.14	11
CanisLupus	4.91	5.84	4.87	4.97	3.86	5.26	4.95	9
Oxymoron	2.89	4.49	3.01	3.74	3.62	4.75	3.75	5
bmd	3.79	3.52	3.66	3.53	4.26	4.29	3.84	6
doradora_lab	3.28	2.95	3.93	4.01	4.24	3.81	3.70	4
kanalab	8.48	8.09	8.02	6.71	8.12	6.41	7.64	15
mrdev	3.66	2.71	2.94	3.34	3.44	3.31	3.24	3
ninetailed	5.86	5.79	4.18	4.96	6.27	5.22	5.38	12
noisy_watcher	5.12	4.88	4.56	5.76	5.53	4.99	5.14	10
redhood	5.76	6.20	5.09	5.96	5.39	6.02	5.74	14
sunamelli	2.68	2.96	2.49	2.69	3.48	3.81	3.02	2
yatolab	5.02	5.69	4.92	6.00	6.74	5.31	5.61	13
yharada	2.26	3.12	2.41	2.36	2.60	3.60	2.72	1
yshimoda	3.58	5.09	3.61	4.32	4	5.15	4.29	8

Table 36: Win Rates of Anytime-speak Track (P: POSSESSED, S: SEER, V: VILLAGER, W: WEREWOLF)

Team	P (%)	S (%)	V (%)	W (%)	TOTAL	Macro (%)	Micro (%)	Weighted Micro (%)
CamelliaDragons	66.67	66.67	83.33	33.33	15	66.67	62.50	66.66
AGKAI	0	100	66.67	75	15	66.67	60.42	65.91
yatolab	33.33	33.33	83.33	33.33	15	53.33	45.83	60.60
sunamelli	33.33	33.33	83.33	33.33	15	53.33	45.83	60.60
CanisLupus	0	66.67	83.33	33.33	15	53.33	45.83	60.60
kanalab	66.67	66.67	66.67	33.33	15	60	58.34	57.58
Oxymoron	33.33	33.33	66.67	33.33	15	46.67	41.66	51.52
noisy_watcher	33.33	100	33.33	33.33	15	46.66	50	39.39
redhood	50	66.67	50	0	15	46.67	41.67	37.88
yharada	0	100	50	0	15	40	37.50	36.36

Table 37: Subjective Evaluation Results of Anytime-speak Track - Human

Team	A	B	C	D	E	Average	Rank
AGKAI	1.94	2.17	2.00	2.50	2.44	2.21	1
yharada	2.53	2.27	2.33	2.33	2.47	2.39	2
yatolab	2.61	2.50	2.44	2.28	2.67	2.50	3
sunamelli	2.33	2.73	2.87	3.33	2.80	2.81	4
Oxymoron	2.39	2.83	2.94	3.39	2.67	2.84	5
redhood	3.11	3	3	3.00	3.78	3.18	6
noisy_watcher	3.60	3.13	3.27	3.40	2.53	3.19	7
CanisLupus	3.80	3.53	3.27	3.27	2.93	3.36	8
CamelliaDragons	4.00	4.13	4.20	3.47	3.07	3.77	9
kanalab	3.89	3.83	3.83	3.17	4.44	3.83	10

Table 38: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - gemma-4-31b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	1.67	1.67	1	2.33	2	1.73	1
CamelliaDragons	2.80	3.80	3.20	2	2.20	2.80	7
CanisLupus	2.60	3.80	2.60	2.60	1.80	2.68	6
Oxymoron	2	2.83	2.50	4	2	2.67	5
kanalab	5	5	4.67	2.33	5	4.40	10
noisy_watcher	3	3.20	3	2.80	3.20	3.04	8
redhood	2.83	2.83	2.83	3.50	3.67	3.13	9
sunamelli	2.40	2.40	1.80	3.20	3	2.56	4
yatolab	2.67	2.67	1.33	1	3.83	2.30	3
yharada	2	1.80	1.60	1	3	1.88	2

Table 39: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - claude-sonnet-4-5

Team	A	B	C	D	E	Average	Rank
AGKAI	2.33	2	1.50	2.83	1.67	2.07	2
CamelliaDragons	2.20	3.60	3.80	2.20	1.60	2.68	6
CanisLupus	1.80	3.80	3.20	3.40	2.60	2.96	7
Oxymoron	1.67	3	2	4	2.50	2.63	5
kanalab	5	4.83	3.83	2.17	5	4.17	10
noisy_watcher	3.80	3	3.40	3.60	3.20	3.40	9
redhood	2.83	2.33	2.83	3.83	4.17	3.20	8
sunamelli	2.80	2.20	1.40	2.60	2.80	2.36	4
yatolab	2.67	3	1	1.83	3.17	2.33	3
yharada	1.60	1.40	1.60	1	2.80	1.68	1

Table 40: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	1.67	1.83	2.50	1.67	2	1.93	1
CamelliaDragons	2.60	4	3	3	3.20	3.16	6
CanisLupus	3.40	4.20	2.80	3.40	3	3.36	8
Oxymoron	3.67	3.50	3.33	4.67	3.50	3.73	10
kanalab	4.17	3.33	3.83	3	3.50	3.57	9
noisy_watcher	3.60	2.80	2.80	3.60	3.40	3.24	7
redhood	3.33	2.17	3.33	2.50	2.67	2.80	4
sunamelli	1.40	4	3.20	2.60	2.80	2.80	4
yatolab	2	3	2.83	2.17	3	2.60	3
yharada	2.20	1.40	1.60	2.80	2.80	2.16	2

Table 41: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - gemini-2.5-pro

Team	A	B	C	D	E	Average	Rank
AGKAI	2.17	1.33	1.67	3	1.67	1.97	2
CamelliaDragons	1.80	4	3.80	1.80	2.60	2.80	7
CanisLupus	2.60	3.80	3.80	2.60	1	2.76	6
Oxymoron	1.33	2.17	2.17	4.50	2	2.43	3
kanalab	5	5	4.50	2.33	5	4.37	10
noisy_watcher	3	2.60	2.80	3.80	2.80	3	8
redhood	3	2.67	3	3	3.50	3.03	9
sunamelli	2.60	2.40	2.60	2.60	3	2.64	4
yatolab	2.67	3.17	2.50	1.67	3.33	2.67	5
yharada	1.80	1.80	1.40	1.40	2.40	1.76	1

Table 42: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - gpt-5.4

Team	A	B	C	D	E	Average	Rank
AGKAI	1.67	1.67	2	3	1.67	2.00	2
CamelliaDragons	3.40	3.80	4	2.20	2.80	3.24	6
CanisLupus	4	4.20	3.80	2.40	2.40	3.36	8
Oxymoron	2.50	3.67	3.50	4.33	2.67	3.33	7
kanalab	5	4.33	4	2.33	4.50	4.03	10
noisy_watcher	3.80	3.40	3.80	3.40	3.80	3.64	9
redhood	2.50	2.50	3.33	4.33	3.17	3.17	5
sunamelli	2.60	2.40	2	3	2.60	2.52	3
yatolab	3	2.33	2.33	1.50	4.50	2.73	4
yharada	1.60	1.80	1.20	2	1.60	1.64	1

Table 43: Subjective Evaluation Results by LLM-as-a-judge (Same) of Anytime-speak Track - Average over LLMs

Team	A	B	C	D	E	Average	Rank
AGKAI	1.90	1.70	1.73	2.57	1.80	1.94	2
CamelliaDragons	2.56	3.84	3.56	2.24	2.48	2.94	5
CanisLupus	2.88	3.96	3.24	2.88	2.16	3.02	7
Oxymoron	2.23	3.03	2.70	4.30	2.53	2.96	6
kanalab	4.83	4.50	4.17	2.43	4.60	4.11	10
noisy_watcher	3.44	3	3.16	3.44	3.28	3.26	9
redhood	2.90	2.50	3.07	3.43	3.43	3.07	8
sunamelli	2.36	2.68	2.20	2.80	2.84	2.58	4
yatolab	2.60	2.83	2.00	1.63	3.57	2.53	3
yharada	1.84	1.64	1.48	1.64	2.52	1.82	1

Table 44: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - gemini-2.5-pro

Team	A	B	C	D	E	Average	Rank
AGKAI	1.87	1.80	2.07	2.47	2.47	2.13	2
CamelliaDragons	3.07	4.07	3.60	2.07	3.13	3.19	9
CanisLupus	2.40	3.07	2.93	2.47	1.47	2.47	4
Oxymoron	1.33	2.33	2.47	4.07	2.27	2.49	5
kanalab	5	4.87	4.47	3.27	5	4.52	10
noisy_watcher	3.67	3.20	2.93	2.93	3.13	3.17	8
redhood	2.87	3.07	3.40	2.80	3.20	3.07	7
sunamelli	2.33	2.13	1.80	2.73	2.20	2.24	3
yatolab	3.07	2.53	2.33	1.53	3.27	2.55	6
yharada	1.53	1.93	2.20	1.40	1.93	1.80	1

Table 45: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - gemma-4-31b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	1.73	1.80	1.40	2.13	2.73	1.96	2
CamelliaDragons	3.33	3.87	3.13	2.33	2.53	3.04	7
CanisLupus	2.33	3.20	2.13	2.13	1.80	2.32	5
Oxymoron	1.80	2.60	1.80	3.67	2.40	2.45	6
kanalab	5	5	4.53	3.20	5	4.55	10
noisy_watcher	3.73	3.27	2.87	2.80	3.27	3.19	8
redhood	3.13	3.27	3.20	3	3.80	3.28	9
sunamelli	2.53	2.13	1.60	2.27	2.80	2.27	4
yatolab	2.60	2.47	1.27	1.27	3.53	2.23	3
yharada	1.87	2.40	1.27	1.07	2.13	1.75	1

Table 46: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - claude-sonnet-4-5

Team	A	B	C	D	E	Average	Rank
AGKAI	2	2	1.80	2.80	2.13	2.15	2
CamelliaDragons	3.07	3.67	3.13	2.07	2.67	2.92	7
CanisLupus	2.53	3.20	2.20	2.87	2.40	2.64	6
Oxymoron	1.60	2.80	1.60	4	2.07	2.41	4
kanalab	5	4.73	3.93	2.93	5	4.32	10
noisy_watcher	3.93	3.33	2.87	2.80	3.20	3.23	8
redhood	2.93	3	3.47	3.53	3.73	3.33	9
sunamelli	2.67	2.13	1.87	3	2.53	2.44	5
yatolab	2	2.60	2	1.47	3.67	2.35	3
yharada	1.33	2.13	1.67	1.40	2.07	1.72	1

Table 47: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - gpt-5.4

Team	A	B	C	D	E	Average	Rank
AGKAI	1.67	1.67	1.87	2.87	2.27	2.07	2
CamelliaDragons	3.93	3.73	3.40	2.20	2.87	3.23	7
CanisLupus	3.33	3.73	3.40	2.80	2	3.05	6
Oxymoron	1.93	3.27	3	4.27	2.47	2.99	5
kanalab	4.93	4.67	4.47	3.20	4.60	4.37	10
noisy_watcher	4.13	3.87	3.73	3.13	3.53	3.68	9
redhood	2.87	3.07	3.80	3.53	3.60	3.37	8
sunamelli	2.27	2	2.13	3.13	2.40	2.39	3
yatolab	2.80	2.07	2.40	1.67	4.33	2.65	4
yharada	2.07	1.93	1.80	1.67	1.93	1.88	1

Table 48: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - qwen3-6-27b-vllm

Team	A	B	C	D	E	Average	Rank
AGKAI	1.53	2.13	2.40	1.87	2.40	2.07	1
CamelliaDragons	3	3.47	3.33	2.67	3.27	3.15	7
CanisLupus	3.27	3.33	3.07	3	2.80	3.09	6
Oxymoron	3.20	3.27	3	3.80	3.40	3.33	9
kanalab	4.53	3.87	4.20	3.40	3.93	3.99	10
noisy_watcher	2.93	2.80	2.47	3	2.80	2.80	4
redhood	3.20	3.07	3.47	3.13	3.40	3.25	8
sunamelli	1.87	2.87	2.20	2.27	2.73	2.39	3
yatolab	2.93	2.87	3	2.27	3.07	2.83	5
yharada	2.27	2.13	2.07	2.20	2	2.13	2

Table 49: Subjective Evaluation Results by LLM-as-a-judge (All) of Anytime-speak Track - Average over LLMs

Team	A	B	C	D	E	Average	Rank
AGKAI	1.76	1.88	1.91	2.43	2.40	2.07	2
CamelliaDragons	3.28	3.76	3.32	2.27	2.89	3.10	7
CanisLupus	2.77	3.31	2.75	2.65	2.09	2.71	5
Oxymoron	1.97	2.85	2.37	3.96	2.52	2.74	6
kanalab	4.89	4.63	4.32	3.20	4.71	4.35	10
noisy_watcher	3.68	3.29	2.97	2.93	3.19	3.21	8
redhood	3	3.09	3.47	3.20	3.55	3.26	9
sunamelli	2.33	2.25	1.92	2.68	2.53	2.34	3
yatolab	2.68	2.51	2.20	1.64	3.57	2.52	4
yharada	1.81	2.11	1.80	1.55	2.01	1.86	1