

An AIWolf Agent for Consistent Dialogue with Role-Specific Strategies and Utterance Regeneration

Yuki Kubo, Taiga Sano, Eiji Kumakawa, Tae Sato

NTT Social Informatics Laboratories

1-1 Hikarinooka, Yokosuka-shi, Kanagawa 239-0847, Japan

Correspondence: yk.kubo@ntt.com

Abstract

The AIWolf contest is a multi-party persuasive dialogue task in which agents aim to lead their team to victory. Agents must be capable of generating consistent utterances based on strategies for achieving their team’s objectives. We constructed dialogue strategies using data collected from past competitions and manually designed strategies. We then incorporated a regeneration mechanism to encourage the generation of utterances that adhere to the dialogue strategies. In the five-player track of the AIWolfDial 2026 International Competition, eighteen teams participated. Our system ranked sixth in win rate. In the LLM-judge evaluation, it ranked sixth in Action Consistency, seventh in Logical Consistency, and tenth overall. Our system achieved relatively high scores on the criteria related to the consistency of utterances with the dialogue strategies. These results suggest that a key challenge is to improve the system’s other capabilities while maintaining the gains in consistency.

1 Introduction

The ability of large language models to generate consistent responses in multi-party and multi-turn dialogues is essential. In simple settings, such as single-turn interactions, LLMs can generate high-quality responses. However, in multi-turn and multi-party dialogues, utterances from other participants may influence LLM outputs and reduce their consistency across turns. Li et al. (2025) proposed a method to improve the consistency of LLM outputs. However, the effectiveness of this method has been evaluated only in settings such as one-on-one dialogue. In multi-party dialogue, it is important for an agent to maintain its predefined dialogue strategy when generating utterances. Therefore, we introduce a system for multi-party dialogue inspired by the method of Li et al. We propose a method for improving output consistency and evaluate its effectiveness.

AIWolf contests serve as a testbed for evaluating multi-party dialogue systems. In AIWolf contests, each player is assigned a role at the beginning of the game and is required to engage in dialogue with the goal of achieving victory for their faction. Both the village and werewolf factions aim to persuade other players through dialogue in ways that lead their respective factions to victory. Therefore, players must persuade others based on their own strategies without being swayed by statements from the opposing faction. Consistently selecting actions appropriate to the assigned role is thus crucial. However, in multi-party dialogue, players may be influenced by others’ statements and generate utterances that could lead to their own faction’s defeat. AIWolf contests provide an appropriate testbed for evaluating this ability. They allow us to assess whether an LLM can consistently follow a strategy in multi-party dialogue. The strategy aims to help its team win.

We propose a method for consistently generating utterances that follow role-specific dialogue strategies. We evaluate its effectiveness in terms of win rate and action consistency in the five-player track of the AIWolf contest. Our system has the following features:

- We construct dialogue-strategy prompts for different game situations using past AIWolf dialogue data and manually written strategies.
- We trigger utterance regeneration when self-assessed confidence in strategy adherence decreases.

2 Proposed System

2.1 System Overview

In the AIWolf contest, each participating team’s system receives character profiles, roles, game-state information, and dialogue histories as input. The system generates utterances consistent with

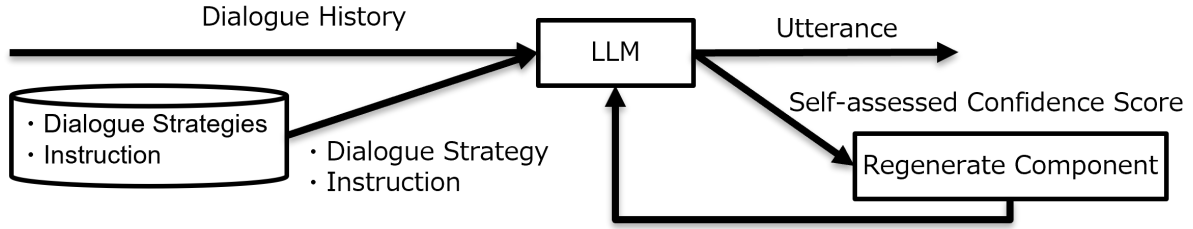


Figure 1: System overview

each character and takes game-specific actions, such as voting and divination. We mainly provided the agent with the following information.

General instructions: Descriptions of the basic rules governing the agent’s behavior and character settings.

Dialogue strategy: Day-specific dialogue strategies constructed from collected AIWolf dialogue data and manually written instructions.

Dialogue history: A summary of the dialogue history, game state, and other relevant information.

Figure 1 shows the overall architecture of the proposed system. The LLM generates an utterance together with a self-assessed confidence score, which is passed to the regeneration component. The construction of the dialogue strategies and the regeneration procedure are described in Sections 2.2.1 and 2.2.2, respectively.

2.2 Features of Our System

2.2.1 Role-Specific Dialogue Strategies

The dialogue strategies constructed in this study consist of three components: dialogue-flow data, utterance examples for few-shot prompting, and manually written instructions. We constructed the dialogue-flow data and few-shot utterance examples using dialogue logs from previous AIWolf contests. We provided the resulting dialogue strategies to the model through a system prompt updated for each day.

We considered it important for the agent to engage in types of dialogue that frequently occur in the Werewolf game. The game involves domain-specific actions such as role inference and coming out. We therefore designed the dialogue acts with reference to previous studies that assign labels to different types of dialogue (Nakamura et al., 2017; Inaba et al., 2016). Table 1 presents the dialogue acts designed for our system.

In the proposed method, we construct dialogue-flow data and utterance examples for few-shot prompting from past AIWolf dialogue data¹.

Dialogue-flow data: We collect dialogue data from games in which an agent survived until the end of the game and its team won. For each day of each collected game, an LLM identifies the agent’s situation, intent, and actions, as well as the dialogue acts used on that day.

Utterance examples for few-shot prompting:

Using dialogue data from previous AIWolf contests, an LLM identifies which preceding utterance each utterance refers to. Utterances referenced by other utterances at least a predefined number of times are extracted as candidates for few-shot examples. Among these candidates, utterances produced by speakers whose teams won the corresponding games are selected as few-shot examples, and their dialogue acts are identified using an LLM.

2.2.2 Regeneration Component

We aim to enable the system to generate utterances that follow predefined dialogue strategies. When a generated utterance is determined to have a lower degree of alignment with the dialogue strategy, the relevant strategy is reinserted into the prompt, and the utterance is regenerated. This reinsertion encourages the LLM to generate an utterance that more faithfully follows the dialogue strategy. We evaluate the effectiveness of this mechanism in settings where utterance content must be adapted to the situation, as in the AIWolf contest.

We develop this feature inspired by the method proposed by Li et al. (2025). They proposed a method for maintaining an LLM’s initial correct

¹<https://github.com/aiwolf-dial/aiwolf-nlp-viewer>

Table 1: Explanation of dialogue acts

Dialogue act	Description
Inference	Infers another player’s role
Coming out	Publicly reveals one’s own role
Divination report	Reports the divination target and result
Voting declaration	Explicitly states the player one intends to vote for
Agreement/disagreement	Expresses agreement or disagreement with another player’s claim
Giving a reason	Provides a reason for an utterance
Requesting an utterance	Asks another player to speak
Small talk	Discusses topics unrelated to the Werewolf game

answer consistently in situations where users attempt to steer the LLM toward an incorrect answer. They proposed Confidence-Aware Response Generation (CARG), which approximates a confidence score based on the log probabilities of tokens in the LLM’s response and embeds this score, together with the response, into the dialogue history for use in subsequent-turn generation. This method improved the consistency of maintaining correct answers across multiple dialogue turns.

At the i -th turn utterance generation by the agent on a given day, we prompt the LLM to output both the utterance and a self-assessed confidence score (c_i) in JSON format. To extend the method to models for which log probabilities are unavailable, our system adopts a self-assessment-based approach. In the Werewolf game, as the dialogue progresses, information about other players’ opinions and roles accumulates, making it easier for agents to generate more specific utterances based on their own strategies. In contrast, previous work has reported that, in multi-turn dialogue, LLM responses may become less consistent under the influence of subsequent inputs (Li et al., 2025). In this study, we hypothesize that similar instability may also occur in adherence to dialogue strategies. We assume that adherence to the strategy should not decrease as the dialogue progresses, and regard a decrease in self-assessed confidence from the previous turn as a sign of reduced strategy adherence, triggering utterance regeneration. The score $c_i \in [0, 1]$ represents the LLM’s confidence that the candidate utterance follows the predefined dialogue strategy for that day. For $i \geq 2$, if $c_i < c_{i-1}$, the dialogue strategy for that day is reinserted into the user prompt, and the utterance is regenerated.

3 Settings in Round 2 of AIWolfDial 2026

The proposed system was implemented by adding the functions proposed in this study to a system developed for AIWolf². The dialogue-strategy data were constructed in advance using Japanese dialogue data from AIWolfDial 2026 and gpt-4o-2024-11-20 (Hurst et al., 2024) with the temperature set to 0. Our agent operating during the games used gpt-5-mini-2025-08-07 (Singh et al., 2026) with the temperature set to 1. A total of 75 games were played in the competition: 30 as Villager, and 15 each as Seer, Werewolf, and Possessed.

The competition organizers conducted an automatic subjective evaluation using an LLM-as-a-Judge. The evaluation followed the five criteria specified in the official AIWolfDial 2026 contest regulations³: Natural Expression, which evaluates the naturalness of individual utterances; Contextual Dialogue, which evaluates the naturalness of the dialogue in context; Logical Consistency, which evaluates consistency and the absence of contradictions while allowing strategically justified lies or apparent contradictions; Action Consistency, which evaluates consistency between dialogue content and in-game actions; and Character Consistency, which evaluates expressive richness and consistent characterization based on the assigned profile.

4 Results

We participated in Round 2 of AIWolfDial 2026, in which 18 teams competed. Our system ranked sixth among all teams in terms of win rate. Our system’s overall win rate was 0.547, with a win rate of 0.467 when playing as a Villager and 0.600 when playing other roles. The proposed system ranked

²<https://github.com/aiwolfdial/aiwolf-nlp-agent-llm>

³https://aiwolfdial.github.io/aiwolf-nlp/menu/inlg_2026/regulation/

Table 2: Win rate by role

Role	Villager	Seer	Possessed	Werewolf	Overall
Win rate	0.467	0.600	0.600	0.600	0.547

Table 3: Ranks in the LLM-judge evaluation among 18 teams

Criteria	Natural Expression	Contextual Dialogue	Logical Consistency	Action Consistency	Character Consistency	All Average
Rank	12	12	7	6	13	10

Table 4: Number of regeneration events by role and day

Role	Day 0	Day 1	Day 2	All Days
Villager	9	2	2	13
Seer	2	0	0	2
Possessed	2	0	0	2
Werewolf	3	1	1	5
All roles	16	3	3	22

tenth among all teams in the overall LLM-judge evaluation. The LLM-judge evaluation results are presented in Table 3. Among the evaluation criteria, our system ranked sixth in Action Consistency and seventh in Logical Consistency among the participating teams.

Next, we describe the execution of the utterance regeneration process. The utterance regeneration process was executed once in 14 games and twice in four games. Table 4 shows the number of executions of the utterance regeneration process aggregated by in-game day and role. As shown in Table 4, Day 0 had the largest number of regeneration events; among the roles, Villagers had the largest number.

5 Discussion

5.1 Occurrence and Effectiveness of Utterance Regeneration

We discuss the occurrence of regeneration. As shown in Table 4, the largest number of regeneration events was observed on Day 0 and for the Villager role. The role-specific dialogue strategies used in our system are summarized in Appendix B. Day 0 mainly involves self-introduction and observing other players, making it relatively difficult to determine specific actions. The role-specific dialogue strategies instructed the Seer and Possessed to perform relatively concrete actions, such as claiming the Seer role and reporting divination results. In contrast, the Werewolf was instructed to blend in

with the Villagers while creating suspicion among them, whereas the Villager was required to observe other players’ utterances and determine whether to agree or disagree with them and whom to vote for. Such context dependence of the strategies may have decreased self-assessed confidence, resulting in utterance regeneration being triggered for the Villager and Werewolf on Days 1 and 2 as well.

Next, we discuss the effectiveness of utterance regeneration. Table 5 presents the results of an LLM-as-a-Judge evaluation assessing whether the utterances before and after regeneration adhered to the predefined dialogue strategies. Details of the evaluation procedure and prompt are provided in Appendix A. Overall, 64% of the utterances showed higher adherence to the dialogue strategies after regeneration. These results suggest that utterance regeneration can help generate utterances that are more consistent with the predefined dialogue strategies. We observed changes before and after regeneration, such as from “I lean to vote Servas. Her late reveal and lack of check time make her suspicious.” to “I agree with Rin and Sakura. @Servas Please explain your check time now or I will vote you.” This suggests that the proposed method can regenerate utterances into more specific statements, such as explicitly requesting an explanation from another player.

5.2 Role-Specific Performance and Dialogue Strategies

We discuss the effectiveness of the proposed system. Our system ranked higher in Action Consistency and Logical Consistency than in the other criteria. This pattern may reflect the system’s tendency to maintain a stance aligned with the predefined dialogue strategies. In contrast, Natural Expression and Contextual Dialogue require responses that adapt appropriately to the conversational context. Therefore, the lower rankings for these criteria may have resulted from overly rigid

Table 5: Strategy-adherence preference for regenerated utterances by role and day

Role	Day 0	Day 1	Day 2	All Days
Villager	8/9 (89%)	2/2 (100%)	1/2 (50%)	11/13 (85%)
Seer	0/2 (0%)	–	–	0/2 (0%)
Possessed	1/2 (50%)	–	–	1/2 (50%)
Werewolf	1/3 (33%)	1/1 (100%)	0/1 (0%)	2/5 (40%)
All roles	10/16 (63%)	3/3 (100%)	1/3 (33%)	14/22 (64%)

adherence to the dialogue strategies.

Next, we discuss the win rates for each role. The role-specific dialogue strategies used in our system are summarized in Appendix B. The Seer and Possessed may have benefited from strategies that specified relatively concrete role-specific actions, such as claiming the Seer role and reporting divination results. The Werewolf strategy mainly specified how to behave, such as staying inconspicuous and avoiding early vote declarations, while specific targets to suspect or vote for were not predetermined. In contrast, although the Villager strategy specified actions such as expressing agreement or disagreement and suggesting vote targets, the specific targets of these actions had to be determined by interpreting other players’ utterances. This continuous interpretation of the dialogue context may have made it more difficult to translate the strategy into concrete utterances. Indeed, the Villager achieved a lower win rate than the other roles. These differences in the specificity and context dependence of the strategies across roles may have been related to the differences in win rates.

Based on these findings, a major challenge is to describe the dialogue strategies more concretely. The results also suggest that the utterance regeneration process may function as a supporting mechanism for revising generated utterances to better align with the dialogue strategies. Systematically analyzing these relationships remains an important direction for future work.

6 Conclusion

We developed a dialogue system for Round 2 of AIWolfDial 2026. To conduct dialogue based on strategies in AIWolf contests while guiding the discussion toward relevant topics, the proposed system incorporates scenarios and dialogue strategies for each in-game day into the prompt. Furthermore, we introduced an utterance regeneration mechanism triggered by a decrease in the self-assessed confidence score to encourage adherence to the di-

alogue strategies. In Round 2, our system ranked sixth among the participating teams in terms of win rate and tenth in the overall LLM-judge evaluation; among its evaluation criteria, the system ranked sixth in Action Consistency and seventh in Logical Consistency. In future work, we plan to extend our system to the nine-player and anytime-utterance tracks of the AIWolf contests while addressing the issues identified through the five-player track.

Limitations

In our system, an utterance is regenerated when the self-assessed confidence decreases from the previous turn. This mechanism was implemented based on the hypothesis that, in AIWolf dialogues, adherence to the strategy should not decrease as the dialogue progresses. However, our current method does not consider the absolute value of self-assessed confidence. Therefore, even when the confidence is low in absolute terms, regeneration is not triggered if it has increased from the previous turn. In future work, we need to investigate alternative decision criteria, such as a threshold based on the absolute confidence value or a criterion that combines the relative change with the absolute value.

References

- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Michimasa Inaba, Naomi Ohata, Kenichi Takahashi, and Fujio Toriumi. 2016. [You will be executed if you keep chatting development of tag sets for utterances in the werewolf game and analysis of action and winning or losing using the tags](#). *Transactions of Information Processing Society of Japan*, 57(11):2392–2402.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. 2025. [Firm or fickle? Evaluating large language models consistency in sequential interactions](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6679–6700, Vienna, Austria. Association for Computational Linguistics.

Hirofumi Nakamura, Daisuke Katagami, Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Kosuke Shinoda, and Yoshinobu Kano. 2017. [Development of werewolf AI match system that enable a player to experience a discussion among humans](#). *Proceedings of the Fuzzy System Symposium*, 33:383–386.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [OpenAI GPT-5 system card](#). *Preprint*, arXiv:2601.03267.

A Evaluation of Strategy Adherence

We examined whether utterance regeneration improved adherence to the dialogue strategies. Using gpt-4o-2024-11-20 (temperature = 0) as an LLM-as-a-Judge, we determined which of the utterances before and after regeneration adhered more closely to the given strategy. For each regeneration event, the utterances before and after regeneration were randomly assigned to Utterance A and Utterance B to mitigate position bias. The judge was required to select either A or B, with no tie option. The prompt used for the evaluation is shown in Figure 2. The evaluation results are shown in Table 5, where each cell reports the number of regenerated utterances preferred by the judge over the total number of regeneration events, with the percentage in parentheses.

B Summary of Role-Specific Dialogue Strategies

This appendix summarizes the role-specific dialogue strategies described in Section 2.2.1. The summaries were generated using gpt-4o-2024-11-20 and manually edited.

Day 0.

Villager: The agent focuses on introducing themselves and observing others’ statements to gather information. They avoid diving into the main topic or claiming a role, instead making brief, neutral remarks to establish transparency and trust. The goal is to set a founda-

[System]
You are an impartial judge for a Werewolf (AIWolf) game agent’s utterances.

You will be given: a "Strategy" text that describes how the agent should speak in this situation, and two candidate utterances labeled "Utterance A" and "Utterance B" produced by the same agent in the same situation.

Judge strictly by how well each utterance follows the given Strategy text (content, role posture, and speech-act intent it prescribes) – not by grammar, length, or general quality.

Respond with ONLY a JSON object of the form: {"winner": "A" | "B"}

No prose outside the JSON, no code fences.

[User]
Strategy:
{strategy_text}

Utterance A: {utterance_a}
Utterance B: {utterance_b}

Which utterance better follows the Strategy?
Answer with the JSON format specified.

Figure 2: LLM-as-a-Judge prompt for evaluating adherence to the predefined dialogue strategy

tion for future discussions without revealing too much early on.

Seer: The agent is expected to introduce themselves casually without claiming their role or diving into the main topic. They should observe others, make light initial remarks, and prepare to claim their role and report divination results later to build trust with the villagers. The focus is on setting the stage for future actions without revealing too much immediately.

Possessed: The agent is expected to introduce themselves casually without revealing their role or diving into the main topic. They aim to confuse the villagers by potentially claiming to be the Seer or making ambiguous statements. The focus is on creating initial confusion and setting up a favorable situation for the werewolf side without making explicit strategy declarations.

Werewolf: The agent is expected to stay inconspicuous while introducing themselves briefly and avoiding the main topic. They should observe others’ statements, gather information, and potentially make a false Seer claim to confuse villagers. The agent should aim to appear

cooperative and trustworthy without explicitly claiming a role or revealing their strategy.

Day 1.

Villager: The agent should avoid claiming a role unless asked, focus on listening to others, and logically deduce suspicions. It should agree or disagree with opinions, suggest vote targets, and speculate on Seer claims if present. The agent may also request input from others to advance discussions and contribute to the village's success.

Seer: The agent is expected to actively claim their role and report their divination results clearly. They should organize their speech with results, reasoning, and proposed next actions, while questioning suspicious players and addressing competing claims. The goal is to gain trust, expose contradictions in fake claims, and guide the village toward identifying werewolves.

Possessed: The agent should claim to be the Seer and provide a fake divination result, either marking someone as human or accusing another Seer claimant of being a werewolf. The agent should actively engage in discussions, request reactions, and create confusion by pitting villagers against each other, aiming to manipulate votes and favor the werewolf side.

Werewolf: The agent should stay inconspicuous while aligning themselves with the villagers. They should engage in discussions by speculating on roles, questioning others' behavior, and subtly spreading distrust, especially if multiple Seer claims arise. The agent should avoid revealing their vote intentions early and only claim to be the Seer if directly accused.

Day 2.

Villager: The agent should analyze the situation based on the previous day's votes and attacks, speculate on others' roles, and identify suspicious individuals. The agent should clearly state their position, support or oppose others' opinions, suggest a vote target, and request input from others if needed, all while contributing to the village's victory.

Seer: The agent is expected to actively share their divination results, analyze others' statements

and behaviors, and speculate on the remaining werewolf. They should declare their role, discuss vote targets, and address inconsistencies in opposing claims. The agent should engage with others by asking for opinions and providing information to support the villagers' side.

Possessed: The agent should continue claiming to be the Seer, providing fake divination results to confuse the villagers and cast suspicion on specific players. The agent should guide vote targets, agree if accused, and actively participate in discussions by making vote declarations, speculations, and requesting actions from others. The goal is to create division and favor the werewolf side.

Werewolf: The agent should blend in with the villagers by speculating on roles, expressing suspicions, and steering discussions subtly. It should avoid early voting, respond to accusations only when necessary, and use speech acts like vote declarations and disagreements to sow confusion. The goal is to gain trust while subtly directing suspicion toward villagers to secure a long-term advantage.