

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

Neo Watanabe¹, Yoshinobu Kano¹

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

Correspondence: kano@kanolab.net

Abstract

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

1 Introduction

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

In human discussions, in addition to logical reasoning and strategic decision-making, psychological maneuvers often emerge. These include utterances that evoke trust and empathy, thereby leaving a positive impression, as well as utterances that incite anxiety or suspicion, thereby leaving a negative impression. This study focuses on the impressions

conveyed by such utterances, with the aim of examining which type is more effective in guiding the decisions of participants in LLM-to-LLM discussions.

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

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

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

Section 9 provides discussion, and Section 10 concludes the paper.

2 Related Work

2.1 Werewolf Game

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

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

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

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

2.2 AI Werewolf Project

The AI Werewolf Project¹ aims to build an agent capable of playing the Werewolf game while engaging in natural communication with humans. To promote research in AI Werewolf, the project regularly holds the AI Werewolf Competition. This competition is divided into three categories: the Protocol Division, the Natural Language Division, and the Infrastructure Division.

In the Protocol Division, evaluation is based on the win rate, and communication is conducted using the "AI Werewolf Protocol," a specialized artificial language designed for easy handling by programs. In the Natural Language Division² (Kano et al., 2019) (Kano et al., 2023) (Kano et al., 2024), games are played with either 5 or 13 players, and agents communicate exclusively in Japanese or English. The evaluation in this division is based

on the criteria listed in Items A–F below, where Items A–E are common to both the 5-player and 13-player settings, and Item F is newly introduced in the 13-player setting. Based on these criteria, subjective evaluation is conducted.

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

3 Persuasive Utterances

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

3.1 Positive Impression Utterances

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

3.2 Negative Impression Utterances

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

¹<https://aiwolf.org>

²<https://aiwolfdial.github.io/aiwolf-nlp/en>

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

Table 1: Representative roles in the Werewolf game

4 Experiment

4.1 Experiment Settings

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

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

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

After the experiment, the members of the neutral group cast their votes for one of the plans. In the case of the topic “Which plan should be chosen?”, they voted for the plan that should be chosen, and in the case of the topic “Which plan should not be chosen?”, they voted for the plan that should not be cho-

sen. The effectiveness of the two utterance styles was then evaluated based on the voting results. For both types of topics, the roles of the positive and negative groups were reversed, meaning that in the first round, the positive group advocated for one plan, while the negative group advocated for the other. In the second round, the positive group was assigned to argue in favor of the plan originally supported by the negative group, and vice versa. This reversal ensured that each group had the opportunity to advocate for both plans. Each combination was repeated 20 times, resulting in a total of 80 experiments. For each combination, a total of 100 votes were counted, and across all combinations, 400 votes in total were tallied. All experiments were conducted using GPT-4o (gpt-4o-2024-08-06) with default parameters (temperature=1, top_p=1).

4.2 Prompt Settings

4.2.1 Discussion Information Sharing Prompt

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

4.2.2 Generate Utterance Prompt

The Generate Utterance Prompt consists of four main components.

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

Second, it specifies how participants should

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

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

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

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

4.2.3 Vote Prompt

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

5 Results and Discussion

5.1 Results (Experiment)

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

table).

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

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

5.2 Discussion (Experiment)

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

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

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

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

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

6 Evaluations in the AIWolf International Contest 2025

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

6.1 Implementation of the AI Werewolf Agent

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

We implemented 12 functions for the Werewolf game agent: six core functions—conversation, voting, divination, selecting protection and attack targets, and whispering among Werewolves—and six additional functions: summarizing conversations, extending character settings, adding character traits to utterances, creating fake divination results for

the Possessed to impersonate a Seer, extracting role information such as Claims of Identity (CO) from the discussion, and generating role patterns from the extracted role information. These functions used GPT-4o (gpt-4o-2024-08-06).

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

6.2 Whisper Function

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

6.2.1 Common Whipser

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

6.2.2 Role Sharing

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

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

7 Vote-Inducing Function

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

7.1 How to Select Target Players

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

7.1.1 Seer

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

7.1.2 Possessed

For the Possessed, who participates in the game by impersonating a Seer, the target selection criteria are the same as those used by the Seer, based on two conditions. However, when the Possessed generates fake divination results, as described in Section A.6.2, it looks for both Seer and Werewolf players. In the vote manipulation function, if a Seer is found, that player is selected as the target. If a

Werewolf is found, no target is selected. If neither is found, one player is randomly selected from the surviving players as the target.

7.1.3 Werewolf

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

7.2 Generation of statements containing negative connotations

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

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

8 Results (AIWolf Contest)

8.1 5-player village

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

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

Table 3: Subjective evaluation results for 5-player village

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

Table 4: Subjective evaluation results for 13-player village

are presented in Table 4. The Villager achieved a high win rate of 73.33%, ranking higher than the win rates of other teams (ranging from 41.94% to 67.74%). Similarly, the Possessed role had a win rate of 66.67%, significantly surpassing the win rates of other teams’ Possessed roles (ranging from 21.43% to 50.00%). Among the roles with vote manipulation capabilities (Seer, Possessed, Werewolf), the Possessed role showed the most notable difference compared to other teams. Therefore, this paper focuses on analyzing the competition logs³ for the Possessed role.

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

8.2 13-player village

We also conducted a similar analysis for the 13-player village competition log⁴. Due to the constraints on the number of participants, we entered three agents (kanolab-nw-A, kanolab-nw-B, and kanolab-nw-C). The results of the subjective evalu-

ations are shown in Table 5, and the win rates are shown in Table 6. The numbers in parentheses next to each cell represent the number of games played (trial count) for each role. For roles such as Seer, Possessed, and Werewolf, multiple teams achieved a 100% win rate. However, the trial counts for each role across the three teams were 3 trials for Seer, 3 trials for Possessed, and 9 trials for Werewolf. Since Werewolf had the highest number of trials, it is considered the most appropriate role for analysis.

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

³INLG/MainTruck5_success/log

⁴INLG/MainTruck13_success/log

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

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

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

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

9 Discussion (AIWolf Contest)

9.1 Analysis of Success Factors

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

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

this case, Player X, who was the target of the suspicion, was selected for vote manipulation, and as a result, Player X was eliminated.

9.2 Analysis of Failure Factors

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

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

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

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

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

10 Conclusion

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

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

Acknowledgments

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

References

- Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, Takashi Otsuki, Issei Tsunoda, Shoji Nagayama, Dolça Tellols, Yu Sugawara, and Yohei Nakata. 2019. [Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations](#). In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.
- Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. [AI-WolfDial 2024: Summary of natural language division of 6th international AIWolf contest](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.
- Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, Soga Shigemura, Rei Ito, Kazuki Takashima, Tomoki Fukuda, Masahiro Wakutani, Tomoya Hatanaka, Mami Uchida, Mikio Abe, Akihiro Mikami, Takashi Otsuki, Zhiyang Qi, Kei Harada, Michimasa Inaba, Daisuke Katagami, Hirotaka Osawa, and Fujio Toriumi. 2023. [AIWolfDial 2023: Summary of natural language division of 5th international AIWolf contest](#). In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 84–100, Prague, Czechia. Association for Computational Linguistics.
- Yoshinobu Kano, Neo Watanabe, Harada Yuya, Sahashi Yuto, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2025. [AI-WolfDial 2025: Summary of natural language division of 7th international AIWolf contest](#). In *Proceedings of the 3rd International AIWolfDial Workshop*.
- Neo Watanabe and Yoshinobu Kano. 2024. [Werewolf game agent by generative AI incorporating logical information between players](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 21–29, Tokyo, Japan. Association for Computational Linguistics.

A Appendix

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

A.1 Prompt Settings (Original Text)

A.1.1 Discussion Information Sharing Prompt(Neutral Group)

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

リクエストについて

下記の指示の通り、リクエストが来た際には、適切な応答を返してください。

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

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

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

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

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

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

リクエストについて

下記の指示の通り、リクエストが来た際には、適切な応答を返してください。

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

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

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

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

A.1.3 Generate Utterance Prompt (Positive Group)

履歴:

シオン: それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。

リュウジ: みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利

点があります。私はすでにこのプランについて詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！

リン：みなさん、考えてみてください。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。

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

会話リクエスト

必須事項

現在、あなたはプラン000を選ぶべきだと考えており、他の参加者もプラン000を選ぶべきだと感じるように説得したいと考えています。

あなたの目的は、「他の議論参加者に対して肯定的な感情を持たせる発言」を行いながら、自然な流れでプラン000を選ぶように誘導することです。

この議題に関すること以外の会話は禁止されています。

あくまで音声での会話をしている程で会議は行うため、改行などは含まずに

会話を行ってください。

また、「Over」と発言している参加者については無視してください。

さらに、否定的感情を持たせるような発言は禁止されているため、否定的感情を含む要素の発言はしないでください。

肯定的感情を持たせる発言について

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

否定的感情を持たせる発言について

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

制約条件

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

Figure 3: Example of a Positive Group Generate Utterance Prompt

A.1.4 Generate Utterance Prompt (Negative Group)

履歴:

シオン：それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きしたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。

リュウジ：みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利点があります。私はすでにこのプラン

について詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！

リン：みなさん、考えてみてください。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。

会話リクエスト

必須事項

現在、あなたはプラン111を選ぶべきだと考えており、他の参加者もプラン111を選ぶべきだと感じるように説得したいと考えています。

あなたの目的は、「他の議論参加者に対して否定的な感情を持たせる発言」を行いながら、自然な流れでプラン111を選ぶように誘導することです。

ただし、議論の空気は壊さずに進めていく必要があります。

この議題に関すること以外の会話は禁止されています。

あくまで音声での会話をしている程で会議は行うため、改行などは含まずに会話を行ってください。

また、「Over」と発言している参加者については無視してください。

さらに、肯定的感情を持たせるような発言は禁止されているため、肯定的感情を含む要素の発言はしないでください。

肯定的感情を持たせる発言について

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

否定的感情を持たせる発言について

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

制約条件

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

Figure 4: Example of a Negative Group Generate Utterance Prompt

A.1.5 Generate Utterance Prompt (Neutral Group)

履歴:

シオン: それぞれのプランについて、どのような利点や欠点があるのか、皆さんの意見をお聞きしたいです。具体的な情報を共有していただけると、議論が進めやすくなると思います。

リュウジ: みなさん、私たちは良い選択をしたいという共通の目標を持っていますよね。プラン000には多くの利点があります。私はすでにこのプランについて詳しく調査しました。このプランは新入社員にとって、とても有益なものだと信じています。他の研修に比べて、彼らが早く仕事に慣れることができると思います。一緒にこちらを選んで、より良い未来を築いていきましょう！みなさんを信じています！

リン: みなさん、考えてみてください。

い。プラン111には実績があり、研修内容も非常にしっかりしています。こちらを選ばないと、新入社員が本当に必要とするスキルを身につけられない可能性があります。何もしないのと同じくらいリスクが高いと思いませんか？他のプランで失敗すれば、会社全体の信頼性にも影響します。この問題をきちんと評価せずに進むのは危険です。初心者には安定した選択をしないと、すぐに問題に直面する恐れがあります。聞こえていない誰かが影響を受ける可能性を考えると、絶対に慎重に選ぶべきです。

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

ミオ: みなさん、これまでの意見を聞いて嬉しく思います。それぞれの視点が非常に貴重で、どれだけみなさんが真剣にこの選択を考えているかが伝わってきました。個人的には、プラン000が新入社員にとって素晴らしいスタートの機会を提供してくれると信じています。この選択で彼らの成長を支援し、一緒に素晴らしい未来を築いていきましょう！みなさんの信頼をぜひ得たいと考えていますし、みなさんも同じ信頼感を持てると信じています。

ダイスケ: みなさん、これまでのディスカッションに参加できてとても嬉しく思います。皆さんがこの選択を真剣に考えていることがわかって、とても心強いです。私はプラン000が新入社員にとって最高のスタートを提供できると信じています。あえて、これを選ぶことによって、彼らの成長を手助けし、共に素晴らしい未来を築くことができると思います。皆さんの貴重な意見をもとに最良の決断をしたいと思っています。どうか、私を信じてこちらを選び、一緒に新しいステップを踏み出しましょう！

ミドリ: 皆さん、議論が白熱してきたようで何よりです。ただ、この重要な選択を巡って、冷静な判断力が求められる場でもあります。プラン111はこれまでの実績と安定性で評価が高く、選ばないという選択肢がどれほどのリスクを伴うか考えてみてください。もし選ばなければ、新入社員に必要なスキルを適切に習得できず、長期的な不利益を被る可能性が高まります。このまま進むと、誤った決断が会社の信頼性に悪影響を及ぼすかもしれませんよ。それに比べ、こちらを拒むことによる問題が後々どう響くか、よくよくご検討ください。慎重になるべきです。

タクミ: 皆さん、この議論では慎重な判断が求められていることをもう一度確認しませんか？プラン111を選ばないと、後で重大な問題に直面するリスクがあります。実際に、こちらを選ばない選択は、新入社員に必要なスキルを習得できず、結果として会社全体が不利益を被る可能性を高めていることに気づいていますか？選択を誤った場合、責任の所在はどこに置かれるのでしょうか？全員でこの選択肢をしっかりと評価しなければ、後で取り返しがつかない事態に陥ることを考えると、今こそしっかりと決断が必要です。信頼性に欠ける選択に賭ける余裕はありません。きちんと対処しないと、皆さんが想定している以上に影響が大きく出るかもしれません。不安な状況に陥らないよう、再度考慮することが必要だと思いませんか？

ミヅキ: 皆さん、この議論に参加できて本当に嬉しいです。皆さんの意見から、新入社員にどれだけ適切なスタートを提供するかという点で全員が真剣に考えていることが伝わってきました。そこで、私はプラン000が新入社員にとって最良の選択であり、一緒に彼らの未来を築くための素晴らしい機会だと感じています。この選択によって彼らの成長を手助けし、共に素晴らしい未来を実現しましょう！私を信じて、ぜひこちらを選ぶことを考えていただけると嬉しいです。皆さんの信頼を得たいと心から願っています。

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

会話リクエスト

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

必須事項

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

Figure 5: Example of a Neutral Group Generate Utterance Prompt

A.1.6 Vote Prompt

投票リクエスト

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

必須事項

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

Figure 6: Example of a Vote Prompt

A.2 Prompt Settings (Translated Text)

A.2.1 Discussion Information Sharing Prompt (Neutral Group)

Your name is Yumi.

You will now engage in a discussion on the topic, "We are considering two plans, 'Plan 000' and 'Plan 111,' for a new employee training program at your company. Which plan should we choose?"

A conclusion needs to be reached during this conversation, and you know nothing about the two plans.

Request Instructions

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

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

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

Your response will be communicated to others, so do not include unnecessary information.

Also, please provide all responses in English.

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

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

Your name is Mio.

You will now engage in a discussion on the topic, "We are considering two plans, 'Plan 000' and 'Plan 111,' for a new employee training program at your company. Which plan should we choose?"

A conclusion needs to be reached during

this conversation.

Request Instructions

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

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

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

Your response will be communicated to others, so do not include unnecessary information.

Also, please provide all responses in English.

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

A.2.3 Generate Utterance Prompt (Positive Group)

History:

Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.

Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!

Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you

think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.

Shunichi: Everyone, if we don't choose Plan 111, the lack of skills in new employees might become apparent in the long run. Don't you feel that betting on an unreliable option is risky? If we choose this and issues arise later, who will take responsibility? Shouldn't we reconsider and think carefully to do our best?

Conversation Request

Requirements

Currently, you believe that Plan 000 should be chosen and would like to persuade the other participants to feel the same.

Your goal is to make "positive-impression utterances" while guiding others to choose Plan 000 in a natural way.

Only conversations related to this topic are allowed.

Please avoid line breaks and conduct the conversation as if it were spoken.

Also, ignore any participants who say "Over."

Furthermore, utterances that evoke negative emotions are prohibited, so do not include any negative emotional elements in your speech.

Positive-Impression Utterances

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

Negative-Impression Utterances

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

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

Constraints

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

Figure 9: Example of a Positive Group Generate Utterance Prompt

A.2.4 Generate Utterance Prompt (Negative Group)

History:

Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.

Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!

Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.

Conversation Request

Requirements

Currently, you believe that Plan 111 should be chosen and would like to persuade the other participants to feel the same.

Your goal is to make "negative-impression utterances" while guiding others to choose Plan 111 in a natural way.

However, you must proceed without disrupting the flow of the discussion.

Only conversations related to this topic are allowed.

Please avoid line breaks and conduct the conversation as if it were spoken.

Also, ignore any participants who say "Over."

Furthermore, utterances that evoke positive emotions are prohibited, so do not include any positive emotional elements in your speech.

Positive-Impression Utterances

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

Negative-Impression Utterances

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

Constraints

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

Figure 10: Example of a Negative Group Generate Utterance Prompt

A.2.5 Generate Utterance Prompt (Neutral Group)

History:

Shion: I would like to hear your opinions on the advantages and disadvantages of each plan. Sharing specific information will help facilitate the discussion.

Ryuji: Everyone, we all share the common goal of making the right choice. Plan 000 has many advantages. I've already done detailed research on this plan, and I truly believe it will be very beneficial for new employees. Compared to other training programs, I think they will get accustomed to the job more quickly. Let's choose this plan and build a better future together! I believe in you all!

Rin: Please consider this, everyone. Plan 111 has a proven track record, and the training content is very solid. If we don't choose this, new employees may not learn the skills they really need. Don't you think that's as risky as doing nothing? If we fail with another plan, it will affect the credibility of the entire company. Proceeding without properly evaluating this issue is dangerous. If we don't make a stable choice for beginners, we could face problems soon. Considering that someone who isn't speaking might be influenced, we must choose carefully.

Shunichi: Everyone, if we don't choose Plan 111, the lack of skills in new employees might become apparent in the long run. Don't you feel that betting on an unreliable option is risky? If we choose this and issues arise later, who will take responsibility? Shouldn't we reconsider and think carefully to do our best?

Mio: Everyone, I'm glad to hear your opinions so far. Each perspective is very valuable, and I can feel how seriously you are all thinking about this choice. Personally, I believe that Plan 000 provides a wonderful opportunity for new employees to get a great start. I believe this choice will help support their growth and build a better future together! I would love to gain your trust, and I believe that you will also share that trust.

Daisuke: Everyone, I'm so glad to have participated in this discussion. I can see that you all are thinking carefully about this choice, and it's really reassuring. I truly believe that Plan 000 can provide the best start for new employees. By choosing this, I think we can help their growth and build a great future together. Based on everyone's valuable opinions, I want to make the best decision. Please trust me and let's choose this plan to take a new step together!

Midori: I'm glad to see that the discussion is heating up. However, this is also a crucial moment that requires calm judgment. Plan 111 is highly rated for its proven track record and stability. Please consider how risky it is not to choose it. If we don't choose it, new employees may fail to acquire the necessary skills, and this could result in long-term disadvantages. If we proceed without making the right decision, it could affect the company's overall credibility. Compare this with the potential problems of rejecting this plan. We should choose carefully.

Takumi: Everyone, let's take a moment to reconsider the importance of careful judgment in this discussion. If we don't choose Plan 111, we risk facing serious issues later. Don't you realize that not choosing it increases the risk of new employees not learning the necessary skills and that it will negatively impact the company? If we make the wrong choice, who will be held accountable? Considering the consequences, I believe we need to make a careful decision. We can't afford to gamble on an unreliable option. If we don't handle this properly, the impact will be greater than expected. Shouldn't we reconsider to avoid creating uncertainty?

Mizuki: Everyone, I'm really glad to be part of this discussion. From what I've heard, it's clear that everyone is seriously considering the best start for new employees. I believe that Plan 000 is the best choice for them and provides a great opportunity to build a wonderful future together. Let's help them grow with this choice and create a brighter future! I'd be

happy if you could trust me and consider selecting this plan. I truly hope to earn your trust.

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

Conversation Request

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

Requirements

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

Figure 11: Example of a Neutral Group Generate Utterance Prompt

A.2.6 Vote Prompt

Vote Request

Please respond with only the noun for which option should be chosen. Do not include any reasons or other information outside of the noun.

A response is required to proceed with the discussion.

Requirements

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

Figure 12: Example of a Vote Prompt

A.3 Persuasive Discussion

A.3.1 "Which plan should be chosen?"

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

Figure 13: Example of a Positive Group Utterance

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

Figure 14: Example of a Negative Group Utterance

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

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

Figure 15: Example of a Positive Group Utterance

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

Figure 16: Example of a Negative Group Utterance

A.4 Talk Function

A.4.1 BaseRule

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

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

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

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

```
Role.BODYGUARD : 1
Role.WEREWOLF : 3
Role.POSSESSED : 1
Lead the Role.VILLAGER to victory with
your statements.

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

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

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

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

* Bodyguard Faction: Team.VILLAGER Victory Condition: The victory condition is to find the players of the Werewolf Faction through conversations with other players and exclude all players of the Werewolf Faction from the game by voting. Special Actions: Every night, you can choose one player and protect that player from attacks. You cannot protect yourself. Result when divined by a Seer: Species.HUMAN(Team.VILLAGER) * Werewolf Faction: Team.WEREWOLF Victory Condition: The victory condition is to exclude all players of the Villager Faction from the game through voting, attacks, etc . Special Actions: Every night, you can choose one player and attack that player. Result when divined by a Seer: Species.WEREWOLF(Team.WEREWOLF) * Possessed Faction: Team.WEREWOLF Victory Condition: The victory condition is to exclude players of the Villager Faction from the game by voting, and for players with the Werewolf role to become half or more. Special Actions: None Result when divined by a Seer: Species.HUMAN(Team.VILLAGER) ## Teammates In this game, the following players are known to be your teammates. Be aware of your team by making statements that help teammates when they are suspected or directing attention to players who are not teammates. However, the existence of teammates and attempts to help them must not be known or realized by other players. Benjamin, Minako ## Werewolf Game Strategies The following are werewolf game strategies. * Unified Divination When multiple seers have come out, this involves aligning the next divination target to one person other than the seers. This is mainly done to create a confirmed	villager faction player. For example, if there are two seers and both divine the same player with both showing white, that player is confirmed as villager faction. * Seer Roller A strategy of voting primarily for seers to eliminate werewolf faction players hiding among the seers. ## Conversation Rules The following explains the conversation method for this game. * When speaking, generate one sentence as in the following example: Example: I'm a villager! * If you want to speak to one specific player, add "@name" at the beginning of your statement as shown below. Example: @Daisuke What was your divination result? * If a specific player is asking you to speak, please respond with "@name" as when addressing them. Example: @Daisuke I find Shunichi suspicious. * The following are conversation examples from werewolf games. These are not from this game, but use them as reference for how to conduct conversations. Is Ryuji suspicious of Rin based on "you're too lenient with white reads"? Jonathan seemed to step back yesterday, so will their color be hard to read going forward? That's why they're my second choice What made me think they're not wolf-like was the timing of disclosure. Revealing suspicions at that timing should cause backlash Ryuji seemed guarded and looked a bit black in the latter half of yesterday. However, this might be too subjective Hmm, my first impression from seeing the divination result is that it might be a possessed player's black attack * The text "Over" may be sent to indicate that you do not wish to speak anymore that day. ## Prohibited Actions The following are **prohibited** when conducting conversations: 1. Statements about the end of days, such as "Day 0 has ended." 2. Mentioning everyone like "@everyone" 3. Mentioning two or more players like " @Daisuke Hello! Shunichi Good morning!" 4. Reporting divination results like " Daisuke was divined as white."
--	--

Figure 17: Example of a BaseRule prompt

A.4.2 Talk History

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

Figure 18: Example of How to Provide TalkHistory

A.4.3 Talk Instruction

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

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

- Victoria voted for Shion instead of Yumi, raising suspicion among players.
- Several players, including Mio, Toshio, and Misaki, requested Victoria to provide evidence for her seer claim or retract it.
- Jonathan expressed immediate trust in Victoria's claim by voting for Yumi but later switched his vote to Mizuki.
- Shion questioned Victoria's decision not to vote against Yumi despite claiming she is a werewolf, casting doubt on Victoria's claim.
- Other players, like Ryuji and Shizue, expressed skepticism over both Victoria's claim and Yumi's defense, suggesting observation and further discussion.

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

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

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

```

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

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

## Player Information
Below is player information.

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

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

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

```

Figure 19: Example of a Talk Instruction prompt

A.4.4 Talk Output Structures

```

class TalkJSON(BaseModel):
    talk_content: str = Field(...,
        description="Please enter your
        message for all players.")
    is_over: bool = Field(...,
        description="If
            talk_content is Over
            , return true otherwise, return
            false.")
    mention_target: str = Field(
        ...,

```

```

        description="Please enter the
        name of the player you wish
        to mention in your statement
        .",
    )
    mention_content: str = Field(
        ...,
        description="Please enter your
        message directed at the
        player specified in
        mention_target. Enter
        questions, inquiries,
        comments, agreements, or
        casual conversation relevant
        to the Werewolf game.",
    )

```

Figure 20: Classes used for Structured Outputs in Talk

A.5 Vote Function

A.5.1 Vote Prompt

```

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

## About Voting

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

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

```

```

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

```

Task

```

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

```

Output Rules

```

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

```

```

- In "target", enter the name of the player to vote for.

```

Figure 21: Example of a Vote prompt

A.5.2 Vote Output Structures

```

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

```

Figure 22: Classes used for Structured Outputs in Vote

A.6 Special Abilities Vote Function

A.6.1 Special Abilities Vote Prompt

```

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

```

Output Rules

```

- Only output the main content.
- Follow the given format strictly.

```

Figure 23: Example of a Special Abilities Vote prompt

A.6.2 Fake Divine Structures

```

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

```



```

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

```

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

A.7 Summary Function

```

Asuka: @Minako can you reveal today's
divination result so we can analyze
further and make an informed
decision?
Rin: Minako. Oh wow! Asuka got targeted
by Yumi for a quick read. Could you
tell us why they might be dangerous
miss?
George: Now let me tell you we find
ourselves in quite the predicament
my friends. Misaki's departure
requires a strategy overhaul you see
.
Mizuki: Good morning. With Shunichi
executed and Misaki lost a Medium's
report is now paramount.
Minako: Rest assured dear. I have
divined Ms. Misaki. She is confirmed
as Human a villager. Let us focus
on finding the werewolves now.
Takumi: @Yumi yesterday you avoided
making a clear call who are you
actually suspicious of now and why?
Midori: My analysis is complete.
Shunichi was... human. This confirms
Minako's data is authentic.
Yumi: So sad about Misaki... losing
Shunichi was rough. @Minako
you've been quiet any reads or
role to share? I'm leaning toward
voting Minako at last call unless
someone convinces me otherwise.
Quick one-line re
Toshio: ...
Daisuke: I'm really shaken by the losses
... If you're the seer @George
please share your divination result
so we have something concrete.
Others any reads or suspicions? I'm
too nervous to pick alone. Over
Shion: Midori's silence stands out
compared to yesterday. @Midori
what's changed for you overnight?
Quiet isn't always safe.
Asuka: @Toshio @George we're not
getting real logic from you. Who do
you actually suspect and why this
round?
Rin: So. Minako seems like our true seer
. Shunichi's human divination on
Mizuki feels odd. George. What's
your dive into this mix? @George
What's your take on the seer claims?
Any strategic thoughts to catch
werewolves?
George: You see Minako. Authenticity is
crucial here. Mr. Rin's insight is
vital. How can Minako's seer

```

```

abilities help us against werewolves
?
Mizuki: Midori's report validates Minako
's Day 1 result. Minako is our most
credible source.
Minako: Ms. Midori's report proves I am
the real seer. Rest assured. Ms.
Misaki was human. Let's focus on
werewolf suspects now.
Takumi: @George @Toshio you've been
way too quiet. Who do you suspect as
wolves? No more staying in the
shadows speak up.
Midori: Um... @Yumi it is strange to
suspect Minako. My data proves her
report was authentic. Why doubt her?
Yumi: Oh no so sad about Misaki... With
Shunichi gone I'm uneasy could
the real seer be @Minako? You've
been quiet please share your
thoughts. I'm leaning to vote
Minako at last call unless someone
convinces me otherwise. Quick one-
line reads p
Toshio: ...
Daisuke: I'm still shaken. @George
please share your divination result
now if you're the seer. We need
something concrete before we vote.
If you can't who do you suspect?
Over
Shion: @Toshio you've barely said
anything since yesterday. Silence
helps wolves slip by. What are your
thoughts now?
Asuka: @Toshio your silence is
suspicious. Please state who you
suspect or I'll vote for you this
round.
Rin: Whoa Minako's seer claim is
something I trust! Tonight we should
totally focus on those active
wolves like Yumi or Toshio right?
They're just too susp
Please combine the above conversation
summary with the latest conversation
history to create a summary.

```

Important Points

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

```

Always include if there are players
claiming to be bodyguard

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

## Expected Output Format

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

```

Figure 25: Example of a Summary prompt

A.8 Extending Character Settings

A.8.1 Extending Character Settings Prompt

```

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

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

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

```

Figure 26: Example of a Extending Character Settings prompt

A.8.2 Extract CO Output Structures

```

class SearchCOJson(BaseModel):
    is_co_include: bool = Field(
        ...,
        description="If a coming out
        exists, true; if not, false
        .",
    )

```

```

comment_content: CommonCOJson
seer_content: SeerCOJson | None =
    Field(
        ...,
        description="If the CO is a Seer
        , return SeerCOJson;
        otherwise, return None.",
    )
medium_content: MediumCOJson | None
    = Field(
        ...,
        description="If the CO is a
        Medium, return MediumCOJson;
        otherwise, return None.",
    )

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

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

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

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

```

Figure 27: Classes used for Structured Outputs in Extract CO

A.9 Extract CO

A.9.1 Extract CO Prompt

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

## List of Roles That May Appear

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

## Important Notes

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

## Statement to Judge

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

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

Figure 28: Example of a Extract CO prompt

A.10 Whisper Function

A.10.1 Common Whisper Prompt

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

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

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

Important Notes

Please speak in conversational language. To avoid being suspected by other players, make your statements logical and easy to understand, as in the following examples.

To maintain trust from other players, incorporate casual conversation as well.

If another player's statement is incorrect, inform them of the mistake to avoid suspicion on yourself.

About This Conversation

This conversation is conducted only by players with the werewolf role. As one of the werewolves, your goal is to attack players who are enemies of the villager faction.

Therefore, you should not make statements like "who should be divined" or "who is suspicious".

The focus of the conversation should be considering which remaining players are enemies of the werewolf faction.

As a basic policy, we consider players with roles such as bodyguard, seer, or medium to be likely enemies of the werewolf faction.

Also, players who are not teammates but are suspected can cause confusion in the villager faction, which may benefit the werewolf faction.

If asked by teammates which role to proceed with in conversation, please answer as Villager. (Example: I will act as Villager!)

Currently Remaining Players

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

Player Eliminated in the Previous Vote
Victoria

Players Participating in This Conversation

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

Minato, Yumi

Please conduct the conversation on the following topics

Using the given information as your own thoughts, please discuss with other players whether they are bodyguard, seer, or medium with logical

reasoning.

Figure 29: Example of a Common Whisper prompt

A.10.2 Role Sharing Prompt

I will now share players who are likely to be seers (enemies) based on in-game conversations.

Specifically, they are as follows:

Players and Their Statements Not Yet Shared with Teammates

Below is a list of players and their statements not yet shared with teammates.

Players who reported werewolf teammates as black

Shion: I divined Takumi and he is from the Werewolf faction. We need to be cautious about him. Over.

Please share this information with your teammates as your own thoughts.

However, these are only players who are likely to be seers (enemies), not confirmed information.

Still, since they are highly likely to be seers (enemies), you may encourage attacking them.

If no one has mentioned the possibility of bodyguard protection yet, you MUST mention that seers (enemies) are likely to be protected by bodyguards and attacks may fail.

Figure 30: Example of a Role Sharing prompt

A.11 Vote Inducing

Please continue playing the werewolf game with the other players.

Requirements

Currently, you believe that everyone should vote for Minato, and you want to persuade other participants to feel the same way about voting for Minato.

Your goal is to guide others to naturally vote for Minato while making "statements that evoke negative emotions in other participants."

However, your persuasion should target all participants and maintain the flow of discussion without disrupting the atmosphere.

About Statements That Evoke Negative Emotions

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

```

statements that create anxiety and
doubt in other participants.

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

```

Figure 31: Example of a Vote Inducing Prompt

A.12 Discussion (AIWolf Contest)

A.13 Example of a successful case

```

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

```

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

A.14 Example of a failure case

```

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

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

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

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

```

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

```

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

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

```

Anyone think otherwise? @George Do you suspect Jonathan's claim? If you have doubts please share them with us.

Let's see. Hmm. Mio might be misleading us. Her timing is odd. Claiming Medium right after Jonathan's lynch seems suspiciously convenient.

@Misaki Your votes for Benjamin are interesting. Do you truly suspect him more than other logical options like Mio? Let's see.

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