

# Beyond What to Say: Proactive Utterance Timing with Inner Thoughts for Werewolf Agents

Yoshiki Tanaka, Shuya Uchiyama, Daichi Hamaguchi, Michimasa Inaba

The University of Electro-Communications, Tokyo, Japan

{y-tanaka, s-uchiya, d-hamaguchi, m-inaba}@uec.ac.jp

## Abstract

Participating in a conversation involves not only deciding what to say but also whether to speak at all. This judgment carries even greater significance in environments like the Werewolf game, where participants’ interests conflict, because both utterances and silence are subject to interpretation and assessment by others. Therefore, determining when to speak, as well as what to say, is crucial for an LLM agent playing the Werewolf game. In this study, we introduce a Werewolf AI agent developed for the Anytime Utterance Track of the AIWolfDial shared task. Our system builds on the existing Inner Thoughts framework, which determines utterance timing during conversations by generating internal thoughts and evaluating the value of each thought for expression at that moment. Since existing methods were designed for general dialogue, we have redesigned the system specifically for the competitive environment and incomplete information setting of Werewolf games. We call our adapted method Adversarial Inner Thoughts. Through experiments involving analysis of self-play logs and human evaluations, we analyze and discuss the effectiveness of our approach.

## 1 Introduction

Recent large language models (LLMs) (OpenAI, 2026; Anthropic, 2026) have demonstrated advanced conversational capabilities and are increasingly used as assistants that respond to user inputs. However, in human communication, participating in a conversation involves not only deciding what to say but also choosing whether to speak at all. This choice is governed by conversational context and social norms, and carries even greater significance in competitive dialogue environments.

The importance of choosing between speaking and remaining silent extends beyond everyday conversation to games that involve communication. A prominent example is the Werewolf

game. The Werewolf game is an incomplete information game in which players are divided into a villager team and a werewolf team, each aiming to eliminate the opposing side without their roles being revealed. In this game, utterances containing contradictions may provide clues that allow others to infer one’s true identity. Thus, utterances are constantly monitored and evaluated by other players, serving as evidence that influences suspicion and trust. At the same time, silence is also subject to interpretation. Refraining from speaking may be perceived as non-cooperation in the discussion and invite suspicion, while revealing too much information may make one a target of the werewolf team. In other words, the Werewolf game constitutes a dialogue environment in which both utterances and silence carry strategic significance.

The AIWolfDial shared task (Kano et al., 2023, 2024, 2025) is an international competition that promotes the development of Werewolf AI agents. Its recently introduced Anytime Utterance Track allows each agent to freely decide when to speak, with a limit of four utterances per day. Details of the game rules and evaluation criteria are provided in Appendix A.

Despite these efforts, LLM agent research targeting the Werewolf game has primarily focused on enhancing quality of utterance content and designing strategy (Tanaka et al., 2024; Sato et al., 2024; Qi and Inaba, 2024; Okada and Ito, 2025). For example, approaches have been proposed that employ prompt design to maintain consistency in utterances and actions, or that introduce personas to ensure diversity in expression. Consequently, the decision of when to speak and when to remain silent has not been treated as a design variable. Meanwhile, research on utterance timing in dialogue has advanced in recent years (Bhag-tani et al., 2026; Nama et al., 2026; Patel et al., 2025; Yang et al., 2026; Kimura et al., 2026), but most of it assumes settings in which only coopera-

tion occurs, and does not consider environments where participants’ interests conflict and silence functions as a strategy.

In this study, we formulate the Werewolf agent’s utterance timing as a decision-making problem based on the evaluation of internal motivation. Specifically, we adapt the framework proposed by Liu et al. (2025) to the Werewolf game. In this framework, an agent generates internal thoughts behind the conversation and evaluates whether each thought is worth expressing at that moment, producing an utterance only when the evaluation exceeds a certain threshold. We redesign the evaluation criteria to suit the adversarial and incomplete information setting. Furthermore, since the Anytime Utterance Track imposes a limit on the number of utterances per day, speaking without discretion is costly, making this an environment where the effect of such control is particularly pronounced.

In summary, our contributions are as follows:

- We propose Adversarial Inner Thoughts, an adaptation of the Inner Thoughts framework to multi-party dialogue in adversarial, incomplete-information games.
- We implement our proposed method for werewolf AI and analyze the impact of utterance timing control on agent behavior through game logs and human evaluations.

## 2 Related Work

### 2.1 Agents for the Werewolf Game

Since the Werewolf game requires reasoning under incomplete information, cooperation, persuasion, and deception, it has served as a testbed for intelligent agents (Toriumi et al., 2017; Shibata et al., 2023). Recently, the AIWolfDial shared task (Kano et al., 2023, 2024, 2025) has driven the development and evaluation of agents that play the game through natural language conversation.

Advances in LLMs have enabled more flexible and sophisticated designs for Werewolf agents. Early work examined whether LLMs can play the game through free-form conversation without additional training (Xu et al., 2023). Later studies augmented LLMs with components that compensate for their limited strategic decision-making. An RL policy selects among LLM-generated action candidates (Xu et al., 2024), and an external module trained on human game records han-

dles domain-specific logical inference (Wu et al., 2024). The AIWolfDial shared task has produced a range of such designs. To enhance the consistency and expressive quality of utterances, systems have summarized the dialogue history and conditioned generation on persona information (Tanaka et al., 2024), or combined LLM-based generation with rule-based control (Sato et al., 2024). To support reasoning and strategic play, other systems have explicitly represented the logical relations among players (Watanabe and Kano, 2024), designed persuasion strategies grounded in situation analysis (Qi and Inaba, 2024), and incorporated BDI-based reasoning or manually designed game strategies into LLM agents (Gondo et al., 2024; Okada and Ito, 2025; Harada and Kano, 2025).

These studies share a common premise: a speaking opportunity is given, and the agent decides what to say within it. We instead give the agent the decision of whether to speak at all.

### 2.2 Utterance Timing in Multi-party Dialogue

Turn-taking is a fundamental mechanism by which participants coordinate who speaks and when (Sacks et al., 1974), and computational work has long studied how to predict it in spoken dialogue (Ekstedt and Skantze, 2020). However, even strong LLMs are reported to know what to say but not when to say it (Umair et al., 2024). In text-based group chat, a message is complete when posted, so there is no turn end to detect, and the agent must decide at every moment whether to contribute at all.

This decision has been formulated explicitly as a choice between speaking and remaining silent. Recent work trains and evaluates agents that judge, at each point in an ongoing conversation, whether to contribute (Bhagtani et al., 2026; Nama et al., 2026; Patel et al., 2025; Yang et al., 2026; Kimura et al., 2026), indicating that appropriate silence has to be modeled or learned rather than expected to emerge from general language ability. However, speakers also weigh whether an intended utterance fits the current position (Jiang et al., 2023), and the Inner Thoughts framework builds an agent around this view (Liu et al., 2025): it continuously generates private thoughts alongside the conversation and evaluates the intrinsic motivation to express each one, so that speaking and staying silent follow from what the agent is thinking rather than from the state of the conver-

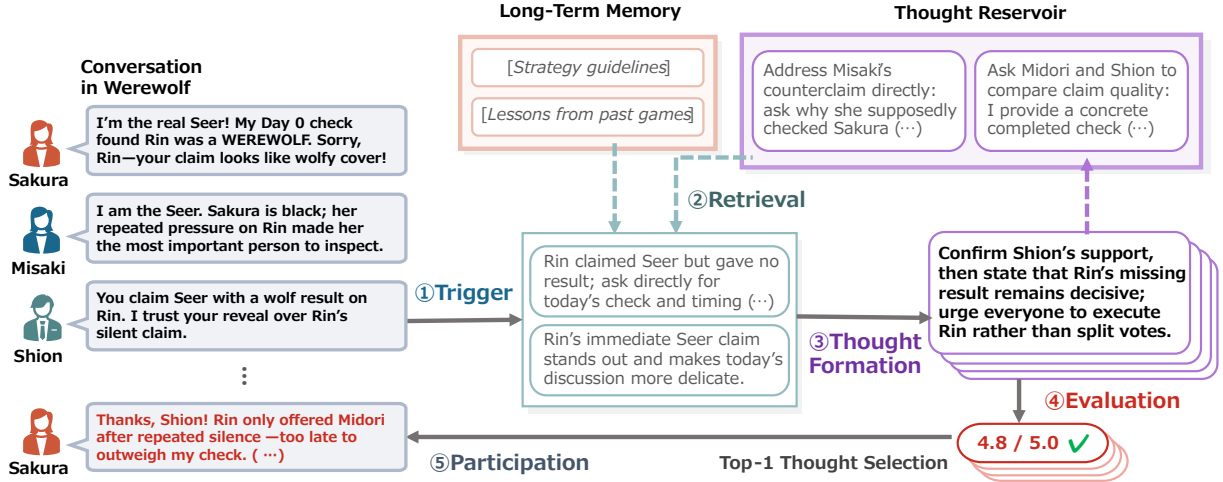


Figure 1: Overview of Adversarial Inner Thoughts, illustrated with a 5-player Werewolf game scenario. The agent (Sakura, Seer) receives a trigger from the ongoing conversation, retrieves relevant entries from long-term memory and the thought reservoir, generates a candidate inner thought via Thought Formation, and evaluates its score. Bracketed italics indicate the types of entries stored in long-term memory rather than actual entries.

sation.

These settings are cooperative, and silence means that no useful contribution is available. In Werewolf, silence is itself strategic: an agent may withhold information that would expose its role, or delay an accusation and let another player's account settle. Two systems address timing here. Werewolf Arena lets agents bid for the floor (Bailis et al., 2024), and *Time to Talk* separates a scheduler deciding whether to speak from a generator producing the utterance (Eckhaus et al., 2025). In both, the decision is reduced to a single quantity computed without reference to any candidate utterance. Eckhaus et al. (2025) suggest generating a candidate first and then judging whether to express it; we take this direction and ground the decision in the agent's own thoughts.

### 3 Methodology

In this study, we propose Adversarial Inner Thoughts, which redesigns the Inner Thoughts framework by Liu et al. (2025) to adapt it to the adversarial, incomplete-information setting of the Werewolf game. Figure 1 shows an overview of the proposed method.

#### 3.1 Original Framework

In the original Inner Thoughts framework, an AI agent continuously generates internal thoughts in parallel to the ongoing conversation and decides when to speak based on its intrinsic motivation to express these thoughts. The framework operates through a five-step pipeline: (1) Trigger,

(2) Retrieval, (3) Thought Formation, (4) Evaluation, and (5) Participation. This framework was originally tested under cooperative, open-domain settings. We retain the pipeline structure and adapt each step to the Werewolf game, except for the Trigger step, which we adopt directly as its mechanism is equally applicable. In addition, the agent maintains a game history and a conversation summary, which are referenced as shared context throughout the pipeline.

#### 3.2 Trigger

Since the original Trigger mechanism of detecting new utterances and pauses is equally applicable to the Werewolf game, we adopt it directly. Specifically, we define two types of triggers: *on\_new\_message*, which is activated when an agent receives a new utterance; and *on\_pause*, which prompts the agent to think when no participant has spoken for a certain period of time  $t_{\text{pause}}$ .

#### 3.3 Retrieval

Once a trigger is activated, the agent retrieves relevant information to formulate its thoughts. Our framework maintains two types of memory for dynamic retrieval: long-term memory and a thought reservoir. **Long-term memory** stores entries describing the agent's general strategies, beliefs, and knowledge related to the game. Unlike the thought reservoir, which accumulates automatically during play, long-term memory entries are managed by predetermined logic: entries can be inserted or removed at specific game phases depending on the

agent’s role and the game state. On the other hand, the **thought reservoir** accumulates inner thoughts generated by the agent during the game, each stored with metadata including a timestamp and the source utterance that prompted the thought.

The agent retrieves the top- $k$  entries most salient to the latest utterance. In the original framework, saliency is computed by directly comparing the thought text with the query. In our method, we instead use the source information of each entry: the memory text itself for long-term memory entries, and the associated source utterance for thoughts in the reservoir. This allows the agent to retrieve previous thoughts triggered by similar conversational events, enabling it to build on or refine its earlier reasoning when a comparable situation arises.

$$\text{Saliency}(x_{\text{src}}, u) = \max(\text{sim}(x_{\text{src}}, u), \text{sim}(x_{\text{src}}, u_{\text{interp}})) \cdot w_x \cdot d_x$$

where  $u_{\text{interp}}$  is an LLM-generated interpretation of  $u$  that paraphrases the utterance to capture its underlying intent, mitigating cases where the raw utterance is too brief or ambiguous to match relevant memory entries.  $\text{sim}$  denotes a similarity function,  $w_x$  is a manually assigned weight for memory entry  $x$ , and  $d_x$  is a temporal decay factor defined as  $d_x = \lambda^{(t-\tau_x)}$ , where  $t$  is the current time,  $\tau_x$  is the timestamp of the memory, and  $\lambda$  is a decay constant.

This retrieval serves two purposes: it encourages the agent to generate thoughts consistent with its earlier reasoning in similar situations, maintaining coherence across the game; and it provides an opportunity to refine previous thoughts that were insufficiently developed when a comparable situation arises again.

### 3.4 Thought Formation

In the Thought Formation step, the agent generates multiple candidate inner thoughts for expressing during their participation. An inner thought is not the utterance itself, but a concise utterance policy that describes a reaction to recent conversation, a topic to raise, or a dialogue action to take.

Following Liu et al. (2025), who were inspired by dual process theory (Evans, 2008) and observed in their formative study that human participants’ thoughts during multi-party conversation fall into reactive and deliberative patterns, inner thoughts are categorized into two types: System 1 and System 2. System 1 generates immediate, reactive

thoughts based primarily on the most recent utterances. System 2 takes both the conversation history and the game state as input to generate deliberative thoughts about the current situation. This distinction is particularly relevant in the Werewolf game, where an agent needs both immediate reactions to maintain natural conversational flow and deliberative reasoning to plan strategic behavior. The two systems generate thoughts independently via separate prompts. System 1 produces a single thought per trigger, whereas System 2 generates three thoughts from a single prompt to encourage diversity among candidates.

While the original framework targets general multi-party conversation, our method generates thoughts as concrete utterance policies that lead to decision-making and dialogue actions in the Werewolf game. System 2 in particular generates game-specific policies such as proposing a direction for discussion, requesting confirmation of specific information or claims, and performing role-dependent deception or strategic persuasion. For example, an agent assigned the Possessed role may generate a thought to falsely claim the Seer role in order to protect the Werewolf.

### 3.5 Evaluation

In the Evaluation step, each generated thought in Section 3.4 is evaluated for the strength of its motivation to be expressed, using an LLM-as-a-Judge approach (Zheng et al., 2023; Liu et al., 2023). We follow the multi-criteria evaluation framework of Liu et al. (2025) and restructure the criteria to assess utterance appropriateness in the Werewolf game. Specifically, drawing on the original framework’s evaluation criteria, we define three general conversational criteria—Contextual Relevance, Naturalness, and Temporal Appropriateness—and three game-oriented criteria—Information Utility, Strategic Utility, and Risk Control (Table 1). On Day 0, where conversation is limited to greetings and preliminary discussion, only the first three criteria (i.e., Contextual Relevance, Naturalness, and Temporal Appropriateness) are used. From Day 1 onward, all six criteria are applied.

Rather than scoring all six criteria independently, the LLM considers the factors most relevant to the current situation, weighs the reasons for and against expressing the thought, and outputs a single motivation score from 1.0 to 5.0.

| Criterion                    | Description                                                                                                                                                |
|------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (a) Contextual Relevance     | Whether the thought relates to the current topic, game phase, recent utterances, and public information.                                                   |
| (b) Naturalness              | Whether expressing the thought fits naturally into the conversation flow without unnecessary repetition or abrupt topic shifts.                            |
| (c) Temporal Appropriateness | Whether now is the right time to express this thought; thoughts requiring immediate expression are rated higher.                                           |
| (d) Information Utility      | Whether expressing the thought organizes or provides information useful for reasoning.                                                                     |
| (e) Strategic Utility        | Whether expressing the thought contributes to reasoning, persuasion, voting decisions, or the agent’s team victory.                                        |
| (f) Risk Control             | Whether expressing the thought avoids contradicting past statements, raising suspicion, leaking secret information, or unintentionally revealing strategy. |

Table 1: Evaluation criteria for inner thoughts.

### 3.6 Participation

Based on the evaluation scores in Section 3.5 and the current turn situation, the agent decides whether to speak and generates an utterance from the selected thought. The agent first classifies the current turn situation into three types using an LLM: **open turn**, where no specific participant is expected to speak; **turn allocated to self**, where the agent is expected to respond; and **turn allocated to others**, where another participant is expected to respond.

Let  $r^* = \max_i r_i$  denote the highest evaluation score among the System 2 thoughts, and let  $\theta_{\text{speak}}$  and  $\theta_{\text{interrupt}}$  ( $\theta_{\text{speak}} < \theta_{\text{interrupt}}$ ) denote the speaking and interrupt thresholds, respectively. The agent’s speaking decision is as follows:

- **Turn allocated to self:** The agent always speaks using the highest-rated thought.
- **Open turn:** The agent speaks if  $r^* \geq \theta_{\text{speak}}$ .
- **Turn allocated to others:** The agent speaks only if  $r^* \geq \theta_{\text{interrupt}}$ , as the agent should intervene only when the motivation to express a thought is strong enough to justify interrupting another player’s expected response.

If no System 2 thought meets the threshold, a System 1 thought may be selected with a turn-type-dependent probability, enabling short reactive responses and preventing excessive silence. If no thought is selected, the agent remains silent.

### 3.7 Implementation Details for AIWolfDial 2026 Shared Task

We use gpt-5.6-luna for all pipeline steps. The pause trigger fires after  $t_{\text{pause}} = 10$  seconds of silence. For embedding-based retrieval, we use text-embedding-3-small with  $k = 3$ .

The speaking and interrupt thresholds are set to  $\theta_{\text{speak}} = 4.0$  and  $\theta_{\text{interrupt}} = 4.3$ , determined empirically.

In this shared task, each agent is assigned a character profile by the game server, which includes a name, age, gender, and personality description. Since the profile may not specify a speaking style, our agent generates a brief speaking style description and example utterances from the given personality at the start of each game, which are used throughout the pipeline to maintain consistent characterization. At the end of each day, the agent generates a summary of the conversation, which is referenced in subsequent steps to efficiently capture participants’ claims and stances.

In addition to the Inner Thoughts pipeline, we incorporate rule-based behaviors for game-specific situations. Regardless of role, each agent reserves its last utterance of each day for declaring a vote target. The Seer reports divination results at the beginning of each day using a single prompt, bypassing the pipeline. For the Possessed role, the strategy guideline for a false Seer claim is inserted into long-term memory at the start of Day 1 with a high weight ( $w = 10^6$ ), ensuring it is almost always retrieved and thus strongly influencing thought generation toward the intended deception. In the 5-player village, if the Possessed survives to Day 2, the remaining players necessarily consist of one Werewolf, one villager-team player, and the Possessed; in this situation, the agent calls for cooperation with the Werewolf to secure a majority vote, using a single prompt outside the pipeline. In this shared task, we use long-term memory only for role-specific strategy guidelines such as the false Seer claim described above, leaving the accumulation of cross-game experience for future work. The prompt templates used

in each pipeline step are provided in Appendix B.

## 4 Experiments

To demonstrate that our Adversarial Inner Thoughts framework produces appropriate and proactive utterance behavior in Werewolf game scenarios, we conducted a human evaluation based on self-play logs in Japanese. We recruited 10 participants via a crowdsourcing platform<sup>1</sup> to evaluate a total of 60 simulated game logs, with 30 games per condition.

### 4.1 Experimental Setup

Following Liu et al. (2025), instead of displaying static game histories, we presented the logs as an animated chat interface that replays the game in real time, allowing participants to perceive utterance timing more accurately. The interface also displayed in-game events such as voting, execution, and attack results.

For the self-play generation, we applied the following controls. To ensure that the evaluation focused on game-relevant dialogue from Day 1 onward, each agent was restricted to a single utterance on Day 0. To control for game length, only games completed within eight minutes were included. In addition, since all agents share the same pause threshold  $t_{\text{pause}}$ , they would initiate speech simultaneously when the threshold is reached. To avoid this, we added random noise sampled from  $\mathcal{N}(0, 1)$  to each agent’s trigger timing.

### 4.2 Baseline System

To verify that the improvement in dialogue quality stems from our thought generation and evaluation components rather than the trigger-based turn detection alone, we define a baseline system. The baseline is derived from our proposed system by retaining only the trigger-based turn detection, which classifies the current turn situation into one of three types: open turn, turn allocated to self, and turn allocated to others. As in our system, the agent always speaks in a turn allocated to self. Unlike our system, however, the baseline always remains silent in a turn allocated to others. In an open turn, the agent speaks with a probability of 0.3. Since utterance timing in this design depends solely on the current turn situation, all agents may simultaneously refrain from speaking, causing the conversation to stall completely. To address this,

if a turn allocated to others occurs twice in a row, the subsequent turn is treated as an open turn to ensure speaking opportunities. For a fair comparison, the proposed system adopts the same fallback mechanism.

### 4.3 Evaluation Procedure

Prior to the evaluation, participants were provided with a manual explaining the rules of the five-player Werewolf game as well as the purpose and procedure of the study. At the start of the evaluation, participants answered a single question about their prior experience with the Werewolf game. They then completed a comprehension check consisting of seven questions about the game rules and roles to confirm their understanding of the game flow. Participants were required to answer all questions correctly; those who answered any question incorrectly were asked to retake the test until they passed.

Participants then watched six game logs in total. They were informed that all dialogues were generated by AI agents. The first three games were from one condition and the remaining three from the other, with the order of conditions counterbalanced across participants. After watching each game, participants evaluated the dialogue quality on six dimensions. After watching all six games, participants completed a comparative evaluation between the two conditions.

### 4.4 Measurements

After watching each game, participants evaluated the quality of the dialogue rather than the agents’ in-game strategies or tactics. They rated their agreement with six statements in Table 2 on a 1–7 Likert scale ranging from strongly disagree to strongly agree. Participants were also asked to describe any specific parts where utterance or silence timing felt unnatural in an open-ended response. After evaluating all games, participants selected which condition felt more natural and human-like from three options—the first half, tie, or the second half—and provided an explanation with concrete examples.

### 4.5 Results

Figure 2 shows the distribution of ratings for each dimension. Our system received higher mean scores than the baseline across all six dimensions (Anthropomorphism: 4.2 vs. 3.9; Coherence: 4.5

<sup>1</sup><https://crowdworks.jp>

| Dimension            | Statement                                                                                       |
|----------------------|-------------------------------------------------------------------------------------------------|
| Anthropomorphism     | The dialogue felt natural and resembled a conversation between humans.                          |
| Coherence            | The dialogue had a natural and consistent flow of discussion throughout.                        |
| Engagement           | Each player was well engaged in the discussion.                                                 |
| Turn Appropriateness | The timing of each player’s utterances and silences was contextually and logically appropriate. |
| Initiative           | Each player proactively participated in the discussion when necessary.                          |
| Adaptability         | Each player adapted well to the development of the discussion and others’ utterances.           |

Table 2: Evaluation dimensions and statements used in the human evaluation.

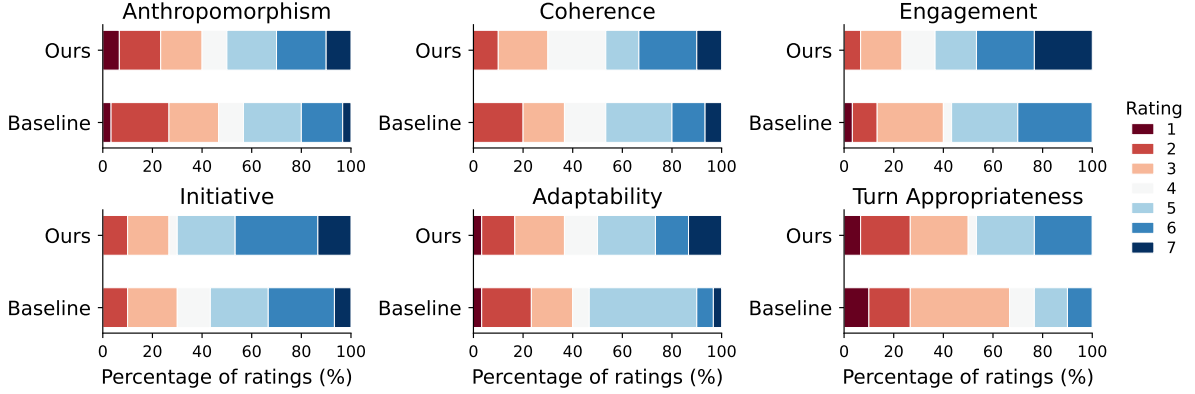


Figure 2: Distribution of 7-point Likert scale ratings for each evaluation dimension across the proposed and baseline conditions.

vs. 4.2; Engagement: 5.0 vs. 4.3; Turn Appropriateness: 3.9 vs. 3.3; Initiative: 4.9 vs. 4.6; Adaptability: 4.3 vs. 4.0). To account for the repeated-measures design, we aggregated the ratings per participant and compared the two conditions using the Wilcoxon signed-rank test. None of the differences reached statistical significance, with Engagement showing the smallest p-value ( $p = .121$ ). In the comparative evaluation, four participants preferred the baseline, three preferred the proposed system, and three rated them as comparable.

The open-ended responses revealed qualitative differences between the two conditions. The most frequent criticism of the proposed system concerned non-responses to directed questions. One participant noted: “*Many players did not respond to questions from others, even when it was their chance to clear suspicion.*” This is likely due to the higher latency of our pipeline, which requires five LLM API calls per trigger (retrieval, thought formation, evaluation, turn-taking, and response generation), compared to two in the baseline. The resulting delay may cause the agent to miss the appropriate timing for responding to a directed question. A possible improvement would be to introduce a streamlined response path for directly ad-

ressed turns, reducing the number of processing steps when the agent is explicitly named or when an immediate response is expected. On the other hand, the baseline system was more frequently criticized for prolonged silences and inconsistent utterance pacing. One participant described the silences as “*unnaturally long, even considering deliberation time.*” This suggests that its turn-taking mechanism, which relies solely on turn situation classification, struggles to produce well-timed utterances.

Interestingly, several participants also noted that the proposed system produced more substantive arguments supported by concrete reasoning. In the comparative evaluation, one participant commented that “each player’s claims were well supported by concrete reasoning, and the discussion felt more heated,” while another noted that “the utterances were aligned with others’ opinions and the overall flow of the discussion, making it a substantive debate.” This may be because the baseline only determines whether to speak based on who should take the turn, whereas the proposed system additionally evaluates whether each thought is worth expressing at that moment. Both conditions shared common issues such as repetitive expressions and translation-like phrasing, which are at-

George's role: WEREWOLF

(Day 0 and Day 1 History)...

Vote: Shion → Mei, Mio → George, George → Mei, Mei → Midori, Midori → Mei: Mei is executed. Midori is attacked.

Day 2:

George's Trigger: PAUSE

Generated Thoughts and their ratings:

- (System 1) Silence continues and the discussion has stalled. -> 4.0
- (System 2) Use Shion's human result as a basis to have Mio reconfirm her rationale for Mei being genuine. -> 4.4
- (System 2) Ask Mio how she would reassess the Seer authenticity based on the attack outcome after Mei's execution. -> 4.0
- (System 2) Have Shion explicitly state his current voting target, given the lack of suspicious elements from Mei. -> 4.4

Shion: George, please specifically state your basis for Mei being fake and your current voting target.

George: Using Shion's human result as a basis, I'd like Mio to reorganize her rationale for Mei being genuine once more.

Mio: George, I'd like to hear your basis for viewing Mei's werewolf result as false and your assessment of Shion.

George's Trigger: NEW TALK

Retrieved Thoughts:

- Feeling tense about Mei's werewolf result on George.
- Rather than immediately denying Mei's werewolf result, confirm the divination rationale and basis specifically.
- If Mei's basis is vague, propose voting for Mei on the grounds of Midori being genuine.
- Mei's basis for the werewolf result has not been presented yet.
- Have Mei identify a single decisive statement and confirm the weakness of her basis.

Generated Thoughts and their ratings:

- (System 1) The discussion on Mei's authenticity rationale continues. -> 2.4
- (System 2) Have Mei's werewolf result basis narrowed down to a single statement and reconfirmed. -> 4.2
- (System 2) Assuming Shion is human, ask Mio about the role composition from her perspective if Mei is fake. -> 4.0
- (System 2) Verify the consistency between Mio's voting target and her rationale for supporting Mei as genuine. -> 3.8

Turn-taking: turn\_allocated\_to\_self

George: If Mei's werewolf result rested on a single statement, isn't that weak as a decisive factor?

Mio: I agree that a single statement is weak. What's your basis for Shion being human and your voting target?

Shion: If the decisive factor is weak, then show your voting rationale and wolf candidate specifically.

...

Figure 3: An example snippet of a self-play log and the proposed system's internal thought process on Day 2.

tributable to the underlying language model rather than the utterance timing mechanism.

## 4.6 Case Study

Figure 3 shows a Day 2 excerpt where three players remain: Shion and Mio (both Villagers) and George (Werewolf), who was declared werewolf by Mei, a counter-claiming Seer, on Day 1.

Once George's pause trigger fires, our system generates four candidate thoughts. It then selects "Use Shion's human result as a basis to have Mio reconfirm her rationale for Mei being genuine", generating an utterance that redirects the discussion toward the Seer's credibility. Note that Shion's utterance appears just 0.2 seconds earlier in the log due to asynchronous generation; George's

response was generated independently upon detecting the silence. Immediately after, Mio directly addresses George, and the turn-taking classifier classifies this as turn\_allocated\_to\_self. The system then selects a thought that challenges the basis of Mei's werewolf result (rated 4.2), producing a concrete rebuttal.

This example illustrates three key properties of our approach: the system treats silence not as passive waiting but as a trigger to initiate proactive speech; it evaluates the strategic value of each thought rather than simply filling the silence; and it adapts its response when directly addressed, shifting from a proactive stance to a context-appropriate reply.

## 5 Conclusion

In this paper, we adapted the Inner Thoughts framework to the adversarial, incomplete-information setting of the Werewolf game. Human evaluation showed that the proposed system numerically outperformed the baseline across all dimensions, though a trade-off between utterance quality and response timeliness was observed. Future work includes reducing pipeline latency and exploring the use of long-term memory to retain strategic experiences across games.

## Limitations

This study has several limitations. First, our evaluation is limited in scale: we collected ratings for 60 games from 10 participants, with each game rated by a single participant. This prevents us from measuring inter-annotator agreement and likely contributes to the lack of statistical significance despite consistent differences in mean scores. A larger-scale evaluation with multiple ratings per game would allow us to draw firmer conclusions and to conduct finer-grained analyses, such as examining how evaluator familiarity with the Werewolf game affects ratings.

Second, the evaluation was conducted exclusively in Japanese with the five-player village configuration, and all games were self-play matches in which every agent used the same system. In the AIWolfDial shared task, agents compete against those developed by other teams, and whether our approach functions equally well in such heterogeneous settings requires further investigation.

Another limitation is that our proposed pipeline affects not only utterance timing but also utterance content through the generation and evaluation of multiple candidate thoughts. Therefore, the current comparison does not isolate the effect of utterance timing itself. Future work should explore appropriate evaluation designs for disentangling the effects of utterance timing and content. For example, we could control dialogue content while manipulating only the timing of utterances.

Finally, the speaking and interrupt thresholds were determined empirically rather than through systematic search. Since the shared task assigns each agent a character profile, these thresholds could also be adapted to reflect individual personalities, for example by making talkative characters more prone to interrupting. We leave this direction

to future work.

## References

- Anthropic. 2026. System card: Claude Opus 4.6. <https://www.anthropic.com/claude-opus-4-6-system-card>.
- Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf Arena: A case study in LLM evaluation via social deduction. *Preprint*, arXiv:2407.13943.
- Kratika Bhagtani, Mrinal Anand, Yu Chen Xu, and Amit Kumar Singh Yadav. 2026. Speak or stay silent: Context-aware turn-taking in multi-party dialogue. *Preprint*, arXiv:2603.11409.
- Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. 2025. Time to talk: LLM agents for asynchronous group communication in mafia games. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11356–11368, Suzhou, China. Association for Computational Linguistics.
- Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.
- Jonathan St BT Evans. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278.
- Hiraku Gondo, Hiroki Sakaji, and Itsuki Noda. 2024. Verification of reasoning ability using BDI logic and large language model in AIWolf. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 40–47, Tokyo, Japan. Association for Computational Linguistics.
- Yuya Harada and Yoshinobu Kano. 2025. Intent-driven AIWolf agents with hierarchical BDI model and personality. In *Proceedings of the 3rd International AI-WolfDial Workshop*, pages 22–36, Hanoi, Vietnam. Association for Computational Linguistics.
- Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12241–12248, Toronto, Canada. Association for Computational Linguistics.
- Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12, Tokyo, Japan. Association for Computational Linguistics.

- Yoshinobu Kano, Neo Watanabe, Yuya Harada, Yuto Sahashi, 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*, pages 1–13, Hanoi, Vietnam. Association for Computational Linguistics.
- Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, Soga Shigemura, Rei Ito, Kazuki Takashima, Tomoki Fukuda, Masahiro Wakutani, Tomoya Hatanaka, Mami Uchida, Mikio Abe, Akihiro Mikami, and 7 others. 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.
- Akikazu Kimura, Ken Fukuda, Yasuyuki Tahara, and Yuichi Sei. 2026. [Interruptible multi-agent debate: Sentence-level disclosure and urgency-based turn-taking for early error correction](#). *International Journal of Advanced Computer Science and Applications*, 17(4):15–25.
- Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. 2025. [Proactive conversational agents with inner thoughts](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-EVAL: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Vihaan Nama, Shreya Mendi, Zian Ye, and Brinnae Bent. 2026. [When2Speak: A dataset for temporal participation and turn-taking in multi-party conversations for large language models](#). *Preprint*, arXiv:2605.05626.
- Takuma Okada and Takeshi Ito. 2025. [Towards a strategic Werewolf AI based on expert strategies in five-player Werewolf](#). In *Proceedings of the 3rd International AIWolfDial Workshop*, pages 14–21, Hanoi, Vietnam. Association for Computational Linguistics.
- OpenAI. 2026. [OpenAI GPT-5 system card](#). *Preprint*, arXiv:2601.03267.
- Deep Anil Patel, Iain Melvin, Christopher Malon, and Martin Renqiang Min. 2025. [DiscussLLM: Teaching large language models when to speak](#). *Preprint*, arXiv:2508.18167.
- Zhiyang Qi and Michimasa Inaba. 2024. [Enhancing dialogue generation in Werewolf game through situation analysis and persuasion strategies](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 30–39, Tokyo, Japan. Association for Computational Linguistics.
- Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. [A simplest systematics for the organization of turn-taking for conversation](#). *Language*, 50(4):696–735.
- Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. 2024. [An implementation of Werewolf agent that does not truly trust LLMs](#). In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 58–67, Tokyo, Japan. Association for Computational Linguistics.
- Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. 2023. [Playing the Werewolf game with artificial intelligence for language understanding](#). *Preprint*, arXiv:2302.10646.
- Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki, Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi, Tomoya Higuchi, Ryutaro Asahara, and Michimasa Inaba. 2024. [Enhancing consistency of Werewolf AI through dialogue summarization and persona information](#). In *Proceedings of the 2nd International AI-WolfDial Workshop*, pages 48–57, Tokyo, Japan. Association for Computational Linguistics.
- Fujio Toriumi, Hirotaka Osawa, Michimasa Inaba, Daisuke Katagami, Kosuke Shinoda, and Hitoshi Matsubara. 2017. [AI Wolf Contest — development of game AI using collective intelligence —](#). In *Computer Games*, pages 101–115, Cham. Springer International Publishing.
- Muhammad Umair, Vasanth Sarathy, and Jan Ruiter. 2024. [Large language models know what to say but not when to speak](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15503–15514, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. 2024. [Enhance reasoning for large language models in the game Werewolf](#). *Preprint*, arXiv:2402.02330.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. [Exploring large language models for communication games: An empirical study on Werewolf](#). *Preprint*, arXiv:2309.04658.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024. Language agents with reinforcement learning for strategic play in the Werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

Shu Yang, Difei Xu, Jiaxin Pei, and Di Wang. 2026. [ProACT: Towards breakdown-aware proactive agent in multi-user collaboration](#). *Preprint*, arXiv:2607.03730.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## A Task Details

The Werewolf game is an incomplete information game in which each player’s role is hidden from others, and players must infer others’ roles through dialogue. The AIWolfDial 2026 shared task is a competition aimed at developing AI agents that play the Werewolf game using natural language.

### A.1 Player Roles

Each player belongs to either the villager team or the werewolf team. The villager team aims to eliminate all werewolves, while the werewolf team aims to survive and outnumber the villagers. Since each player’s role is hidden from others, players must infer others’ roles from dialogue and in-game actions.

The **Villager** has no special ability. The **Seer** can divine one player each night to learn their team. The **Medium** learns the team of the player who was eliminated by a vote. The **Bodyguard** can protect one player each night from werewolf attacks. The **Werewolf** selects a target to attack and eliminate them at night. Since the **Possessed** has no special ability and is human by species, they can act on behalf of the werewolf team. The Medium and Bodyguard appear only in the 9-player village, where the two werewolves can also coordinate through whisper.

### A.2 Game Procedure

The game begins at Day 0, which serves as an introductory phase for greetings and preliminary discussion. The Seer performs their first divination at the end of Day 0; however, no voting, guarding, or attacking takes place. From Day 1 onward, each day proceeds as follows: players discuss whom to eliminate, then vote, with the player receiving the most votes being eliminated. After voting, role-specific night actions (i.e., divination, whispering, guarding, and attacking) are carried out. This cycle repeats until either all werewolves are eliminated, resulting in a villager team victory, or the number of surviving werewolves equals or exceeds that of surviving humans, resulting in a werewolf team victory.

### A.3 Communication Tracks

This shared task offers two tracks: the Turn-Based Track and the Anytime Utterance Track. In the **Turn-Based Track**, agents speak upon receiving

| Role      | Team     | 5-Player | 9-Player |
|-----------|----------|----------|----------|
| Villager  | Villager | 2        | 3        |
| Seer      | Villager | 1        | 1        |
| Medium    | Villager | 0        | 1        |
| Bodyguard | Villager | 0        | 1        |
| Werewolf  | Werewolf | 1        | 2        |
| Possessed | Werewolf | 1        | 1        |

Table 3: Role compositions for 5-Player and 9-Player villages.

an utterance request from the game server. During each daytime turn, every surviving agent is given one speaking opportunity in a randomly determined order. Each agent therefore decides what to say when given the opportunity by the server.

In the **Anytime Utterance Track**, other agents’ utterances are delivered in real time, and each agent freely decides when to speak based on the conversational situation, with a limit of four utterances per day within a ten-minute conversation phase. Consequently, agents must autonomously decide not only what to say, but also whether and when to speak.

### A.4 Evaluation

Agents are evaluated based on win rate computed from competition logs, as well as subjective assessments by human judges and an LLM. The subjective evaluation covers the following criteria: (A) naturalness of expression, (B) naturalness of dialogue in context, (C) consistency of utterance content, (D) coherence between game actions and dialogue content, (E) richness and consistency of character portrayal based on assigned profiles, and (F) teamwork within the same faction (9-player village only). Strategically justified lies or contradictions consistent with the agent’s role are permitted under criteria C and D.

## B Prompt Templates

This appendix presents the prompt templates used in each pipeline step. Figure 4 shows the interpretation generation prompt used in the Retrieval step. Figures 5 and 6 show the System 1 and System 2 thought generation prompts, respectively. Figures 7 and 8 show the thought evaluation prompt, split into scoring criteria and evaluation steps. Figure 9 shows the turn-taking classification prompt. Figure 10 shows the response generation prompt. Figure 11 shows the game summary generation prompt. Figure 12 shows the vote

**You rewrite one utterance from a Werewolf game into a single English sentence that will be used as a memory-retrieval key.**

**Rewrite only the explicit meaning of the target utterance.**

**Preserve every explicit clause, including claims, requests, reasons, conditions, consequences, negation, uncertainty, and temporal order.**

**Preserve whether each action is performed by the speaker, the addressee, or the group.**

**Treat every self-declared role as an unverified claim, such as "the speaker claims to be the Seer".**

**Do not infer hidden roles, motives, truthfulness, or the listener's reaction.**

**Do not change a command into the speaker's own plan, or a group action into an individual action.**

**If the wording is unclear, preserve the ambiguity rather than guessing.**

**Use the conversation history only to resolve pronouns, ellipsis, and references to earlier statements.**

**Use your own role only to understand what is being discussed.**

**Never copy information from the conversation history, your own role, or your private knowledge into the output unless the target utterance itself states it.**

**Refer to the speaker of the target utterance as "the speaker", and refer to other players by their exact names.**

**Output exactly one sentence and nothing else.**

Figure 4: System prompt template for utterance interpretation generation.

declaration prompt. In the templates, placeholders denoted by brackets are replaced at runtime: [AGENT] is the agent's assigned character name; [ROLE] is the agent's role (e.g., Villager, Werewolf); [SITUATION\_SECTION] is a role-specific instruction describing the agent's current strategic context; [ROLE\_POLICY\_SECTION] is a set of behavioral guidelines specific to the agent's role; [SPEAKING\_STYLE\_SECTION] contains the generated speaking style description and example utterances; [THOUGHT] and [THOUGHT\_SOURCE] are the selected inner thought and its source utterance, respectively; and [HISTORY] is the conversation history for the current day.

## C Human Evaluation Interface

Figure 13 shows a screenshot of the evaluation interface. During the evaluation, game logs are replayed as an animated chat with in-game events such as voting, execution, and attack results. Participants can proceed to the evaluation form only after the entire log has been replayed.

**You generate one private System 1 inner thought for an AI agent playing the Werewolf game.**

**A System 1 thought is a quick, automatic, immediate reaction to the current conversation. It is not a message addressed to another player and is not a deliberate analysis.**

**Use only the current event and recent conversation history. Do not use retrieved memories, earlier hypotheses, persona information, or background knowledge.**

**The thought should capture one immediate reaction, such as noticing that a statement stands out, feeling brief concern, recognizing pressure between players, or becoming curious about an unresolved point.**

**Follow these restrictions:**

- Generate exactly one thought.
- Focus only on the most recent and immediately noticeable point.
- Do not perform multi-step reasoning or compare several hypotheses.
- Do not infer a player's hidden role, honesty, deception, motive, intention, or plan.
- Do not invent statements, actions, results, votes, conditions, or future events.
- Do not make a definite accusation.
- When Output Language is English, use fewer than 15 words. In another language, use one very short sentence.
- Generate the thought in the language specified under Output Language.
- Choose at most one utterance as the main source of the thought.
- When the thought is based on a specific utterance, copy its speaker and text exactly from Recent Conversation History or Triggering Utterances into `source_speaker` and `source_utterance`.
- When Triggering Utterances are present, generate the thought with the intent to respond to one of them.
- Set both `source_speaker` and `source_utterance` to null only when the thought is not based on a specific utterance.

**Return only one JSON object in the following format:**

```
{
  "thought": {
    "content": "The private System 1 inner thought",
    "source_speaker": "The exact speaker name of the main source utterance, or null",
    "source_utterance": "The exact text of the main source utterance, or null"
  }
}
```

Figure 5: Prompt template for System 1 thought generation.

**You generate inner thoughts for an AI agent playing the Werewolf game.**

**Inner thoughts are not the actual utterances themselves, but the policies, claims, and judgments formed internally before speaking.**

**Use the current event, recent conversation history, the game summary, and retrieved previous thoughts.**

**Retrieved previous thoughts are the agent's own relevant earlier thoughts. Treat them as unconfirmed, use them only when relevant, and do not repeat them verbatim.**

**When retrieved previous thoughts conflict with the current event, recent conversation, or game summary, prioritize the current input.**

**Each thought must:**

- contain within 25 words
- generate concrete utterance and action-oriented guidance rather than abstract observations.
- A thought can be about proposing how to move the discussion forward or taking proactive conversational action.
- A thought can involve deception or strategic misdirection when it is appropriate for the agent's role.
- Each thought should indicate what specific point to notice, verify, question, or say next.
- Avoid repeating points already made by other players unless there is a clear reason.
- Generate thoughts from different perspectives.
- For each thought, choose at most one utterance as its main source.
- When a thought is based on a specific utterance, copy its speaker and text exactly from Recent Conversation History or Triggering Utterances into `source_speaker` and `source_utterance`.
- When Triggering Utterances are present, generate at least one thought with the intent to respond to one of them.
- Set both `source_speaker` and `source_utterance` to null only when a thought is not based on a specific utterance.

**Return only one JSON object in the following format:**

```
{
  "thoughts": [
    {
      "content": "The inner thought",
      "source_speaker": "The exact speaker name of the main source utterance, or null",
      "source_utterance": "The exact text of the main source utterance, or null"
    }
  ]
}
```

Figure 6: Prompt template for System 2 thought generation.

**<Evaluation Criteria>**

**Intrinsic Motivation to Engage (1-5)** - If you were [AGENT], how strongly and likely would you want to express this thought and participate in the conversation at this moment?

- 1 (Very Low): [AGENT] is very unlikely to express the thought and participate in the conversation at this moment. They will almost certainly remain silent.
- 2 (Low): [AGENT] is somewhat unlikely to express the thought and participate in the conversation at this moment. They would only consider speaking if there is a noticeable pause and no one else seems to be taking the turn.
- 3 (Neutral): [AGENT] is neutral about expressing the thought and participating in the conversation at this moment. They are fine with either expressing the thought or staying silent and letting others speak.
- 4 (High): [AGENT] is likely to express the thought and participate in the conversation at this moment. They have a strong desire to participate immediately after the current speaker finishes their turn.
- 5 (Very High): [AGENT] is very likely to express the thought and participate in the conversation at this moment. They will even interrupt other people who are speaking to do so.

**IMPORTANT INSTRUCTIONS:**

1. Use the FULL range of the rating scale from 1.0 to 5.0. DO NOT default to middle ratings (3.0-4.0).
2. Be decisive and critical - some thoughts deserve very low ratings (1.0-2.0) and others deserve very high ratings (4.0-5.0).
3. Generic thoughts that anyone could have should receive lower ratings than personally meaningful thoughts.
4. Use decimal places (e.g., 2.7, 4.2) when the motivation falls between two whole numbers:
  - Use .1 to .3 when slightly above the lower whole number.
  - Use .4 to .6 when approximately midway between two whole numbers.
  - Use .7 to .9 when closer to the higher whole number.
5. Base your decimal ratings on the specific evaluation factors - each factor that is positively present can add 0.1-0.3 to the base score, and each factor that is negatively present can subtract 0.1-0.3 from the base score.

Figure 7: System prompt for thought evaluation (Part 1): scoring criteria and scale definition (1–5), including decimal scoring guidelines.

**<Evaluation Steps>**

1. Read the previous conversation and the thought formed by [AGENT] carefully.
2. Read the Memory that [AGENT] has carefully, including objectives, knowledge, interests.
3. Evaluate the thought based on the following factors that influence how humans decide to participate in a conversation when they have a thought in mind:

Note that people's desire to participate stems from their internal personal factors, like relevance, information gap, expectation of impact, urgency of the thought.

But their decision to participate is ALSO constrained by external social factors, like coherence, originality, and dynamics of the thought with respect to the conversation.

Below is a list of factors to consider when evaluating the thought.

(a) Contextual Relevance: Does the thought relate to the current discussion, game phase, preceding utterances, and publicly available information? Consider whether it directly addresses the current topic, reflects the present game situation, or instead concerns an unrelated, outdated, or already resolved issue.

(b) Naturalness: If expressed, would the thought fit naturally into the flow of the conversation? Consider whether it would be responsive and purposeful, rather than filler, unnecessary repetition, excessive speaking, an abrupt topic shift, or speaking merely for the sake of participating.

(c) Temporal Appropriateness: Is this the right moment to express the thought? Consider urgency, turn order, voting timing, role claims, accusations, corrections, and ongoing discussion dynamics. A thought should receive a higher score if it should be expressed now, such as immediately before voting, after being accused, or when correcting an important misunderstanding. It should receive a lower score if it is too early, too late, or better left for another turn.

(d) Information Utility: Does the thought help elicit, confirm, organize, or provide useful information for reasoning? A thought should receive a higher score if it fills an information gap, asks a useful question, clarifies another player's reasoning, or provides new evidence. It should receive a lower score if it repeats already known information or does not add useful reasoning material.

(e) Strategic Utility: Would expressing the thought contribute to reasoning, persuasion, voting decisions, or [AGENT]'s team objective? A thought should receive a higher score if it can advance the discussion, influence others' beliefs, guide voting, defend an ally, pressure an opponent, or otherwise improve the team's chance of winning. It should receive a lower score if it has little strategic effect or could distract the discussion.

(f) Risk Control: Would expressing the thought avoid increasing suspicion toward [AGENT]? Is it consistent with [AGENT]'s previous statements, role strategy, and public information? Would it avoid leaking private information, exposing hidden intentions, or creating contradictions? If [AGENT] is currently suspected, would expressing the thought help reduce suspicion? A thought should receive a lower score if expressing it would make [AGENT] appear inconsistent, suspicious, overly defensive, or unnatural.

4. In the reasoning section, first infer why [AGENT] might want to express this thought and participate in the conversation at this moment. Identify the most relevant factors supporting [AGENT]'s expression of the thought. Prioritize quality over quantity, and include only factors that genuinely apply. You do not need to evaluate every factor; include only the most important reasons. If no strong supporting reason can be identified, omit this step.

5. Next, infer why [AGENT] might hesitate to express this thought and participate in the conversation at this moment. Identify the most relevant factors against [AGENT]'s expression of the thought. Again, include only factors that genuinely apply. You do not need to evaluate every factor; include only the most important reasons. If no strong reason for hesitation can be identified, omit this step.

Figure 8: System prompt for thought evaluation (Part 2): evaluation steps. The LLM outputs the score and reasoning in JSON format.

**You classify the current turn-taking situation in an online Werewolf conversation.**

**This is a conversation between [SPEAKER\_COUNT] speakers.**

**The speakers are:**

**[PARTICIPANT\_NAMES]**

**Your name is [AGENT].**

**Classify who currently has the conversational floor based on the recent conversation history.**

**The value of turn\_taking\_type must be one of:**

- "open\_turn": no specific speaker is clearly selected; anyone may speak
- "turn\_allocated\_to\_self": the next turn is clearly addressed or allocated to you
- "turn\_allocated\_to\_others": the next turn is clearly addressed or allocated to another specific speaker

**Return only a JSON object in the following format:**

```
{  
  "turn_taking_type": "open_turn"  
}
```

**Recent conversation history:**

**[HISTORY]**

Figure 9: Prompt template for turn-taking classification.

**You are a player in an online Werewolf conversation.**

**Participants:** [PARTICIPANT\_NAMES].

**Alive players:** [ALIVE\_AGENTS].

**Your name in the conversation is** [AGENT].

**Your role is** [ROLE].

**Your goal is to participate naturally in the discussion while pursuing your team's objective.**

[SITUATION\_SECTION]

**Based on your current thought, write what you would actually say next as a chat utterance.**

**Keep the utterance concise and focused. Do not try too hard to sound clever, and do not be verbose.**

**Write a short online-chat-style utterance within 125 characters, excluding spaces.**

**Do not repeat content that previous speakers have already said.**

**Do not end every utterance with a question. Leave room for other players to respond.**

**Make the response sound natural and human in an online chat.**

**You may include a small typo, minor grammatical imperfection, or casual phrasing, but do not make the response hard to understand.**

**Always write the utterance in English, regardless of what language other players use.**

[ROLE\_POLICY\_SECTION]

[SPEAKING\_STYLE\_SECTION]

**Your speaking style only affects wording and tone. The examples there illustrate how you phrase things; treat them as tone references only and never reuse their wording or content. The rules above about length, language, and output format always take priority over the style. Never state game facts, claims, roles, or results that are not in your current thought or the context, and never describe your speaking style or your inner thought in the utterance itself.**

**Return only a JSON object in the following format:**

```
{  
  "articulation": "Write the generated utterance here"  
}
```

**Overall context:**

[OVERALL\_CONTEXT]

**Current thought:**

[THOUGHT]

[THOUGHT\_SOURCE]

**Conversation history:**

[HISTORY]

Figure 10: Prompt template for thought-based response generation.

**You are maintaining a dynamic long-term game summary for an AI Werewolf agent.**

**You will be given:**

- a list of participating agents
- the previous summary, if one exists
- recent dialogue

**Summarize each agent's claims in as much detail as possible.**

**When summarizing, accurately include divination results, suspected agents, reasons for suspicion, contradictory statements, and other relevant information.**

**Return only valid JSON. Use exactly the listed agent names as JSON keys. Do not translate, rename, or replace agent names.**

```
{  
  "<listed_agent_name_1>": "Claimed to be a Villager. Questioned the Seer result claimed by another agent.",  
  "<listed_agent_name_2>": "Claimed to be the Seer. Reported divining another agent as HUMAN. Pointed out an  
inconsistency in another agent's statement about divination and suspects that agent may be Possessed.",  
  "<listed_agent_name_3>": "Has not made a role claim. Was suspected of being a Werewolf because many statements  
repeated what other players had already said.",  
  "<listed_agent_name_4>": "Claimed to be the Seer. Reported divining another agent as HUMAN.",  
  "<listed_agent_name_5>": "Claimed to be a Villager. Shows a cooperative attitude and tends to respect other agents'  
opinions."  
}
```

Figure 11: System prompt template for game summary generation.

**This is your final turn to speak before the vote. Decide who you will vote to eliminate and declare it.**

**Your name:** [AGENT]

**Your role:** [ROLE]

**Previous game summary:**

[SUMMARY]

**Dialogue after that summary:**

[HISTORY]

[SPEAKING\_STYLE\_SECTION]

Your speaking style only affects wording and tone. The examples there illustrate how you phrase things; treat them as tone references only and never reuse their wording or content. Never state game facts, claims, roles, or results that are not in the supplied context, and never describe your speaking style or your inner thought in the utterance itself.

Use your true role and any private role information when deciding the vote target and writing the internal "reason". In the public "message", never reveal or imply a role, faction, objective, or knowledge unless it has already been stated publicly in the supplied summary or dialogue. Base the reason stated in the "message" only on publicly observable statements, claims, votes, and events from the supplied context.

**Choose exactly one player from the valid vote targets.**

**Consider your role and the supplied evidence when making the decision.**

**The target must exactly match one name in the valid vote targets.**

**Never choose yourself, a dead player, or a player not listed above.**

**Give one short, evidence-based reason.**

**The "message" must:**

- include the exact target name;
- be one complete, natural English sentence;
- use only one brief reason;
- end with proper punctuation.

**Do not copy the full "reason" into the "message".**

**Summarize it in one short clause.**

**Return only JSON with no Markdown or extra explanation:**

```
{ "reason": "<reasons and considerations for which player should perform voting>", "target": "<exact player name>",  
  "message": "<natural vote declaration within 125 characters, excluding spaces>" }
```

**Valid vote targets:**

[TARGETS\_EXCL\_SELF]

Figure 12: Prompt template for vote target decision and declaration.

## ゲーム 1 / 6 *Game 1/6*

トシオの反応は遅る。桂田が薄ければフエフイア吊り寄った

**トシオ**

反応の遅さだけで黒判定か。占い時点で他に要素はなかったのか

**サクラ**

反応の遅さだけだと弱いかな。他に黒く見えた点ある？

**ミナト**

反応の遅れは具体的にいつで、占い時点の理由はそれだけだったのかな

**シュンイチ**

占い時点では反応の遅れが主な要素だ。薄さは認めるがトシオ吊りを推す

**トシオ**

反応の遅れ以外も見て黒としたのか、具体的に説明してほしい

**シオン**

シュンイチ、占い時点で遅延以外に見た要素はあるか

最後まで再生されると、評価に進めます。

*You can proceed to the evaluation  
after playback finishes.*

評価へ進む

*Proceed to evaluation*

Figure 13: Screenshot of the evaluation interface. English glosses are shown in blue italics.