Towards a Strategic Werewolf AI Based on Expert Strategies in Five-Player Werewolf

Takuma Okada, Takeshi Ito
The University of Electro-Communications

Abstract

AI has surpassed humans in perfect information games, yet imperfect information games like Werewolf remain difficult due to uncertainty and persuasion. This study focuses on Five-Player Werewolf and proposes a strategic agent that models expert play. Key features include Villager CO (Coming Out) and first-day utterances designed for second-day persuasion, combined with ChatGPT-based dialogue generation. Self-play experiments showed novel behaviors such as universal Seer CO and Villager CO, though win rates remained low. Future work will introduce learning-based strategies and validation against human players.

1 Introduction

In recent years, AI has demonstrated achievements surpassing human performance in perfect information games such as Go and Shogi. In contrast, imperfect information games such as Werewolf and Poker remain challenging for AI, as they involve psychological tactics and uncertainty.

Since the launch of the Werewolf Intelligence Project [Werewolf AI Project], the natural language aspect has faced the most significant challenge in generating natural utterances. However, with the emergence of large language models (LLMs) in the past year, this issue has been greatly alleviated. Nevertheless, the strategies employed by current agents remain limited: in Five-Player Werewolf, typically only the Seer, the Possessed, and occasionally the Werewolf engage in Seer claiming (CO), while more advanced strategies such as Villagers performing CO (Coming Out) have not yet been observed.

This study aims to develop a Werewolf agent capable of more complex and human-like persuasive behavior by modeling and implementing characteristic strategies observed in expert human play.

2 Five-Player Werewolf

The Werewolf game is generally played as a party game with nine or more participants. However, in such large-scale settings, players are often eliminated for unreasonable reasons due to a lack of information. In this study, we focus exclusively on the Five-Player Werewolf variant, which strikes a balance between strategic depth and analytical tractability. The role composition consists of one Werewolf, one Possessed, two Villagers, and one Seer. The game is guaranteed to conclude by the end of the second day, requiring players to make dense and strategic decisions within a short time frame.

This format preserves the essential features of Werewolf, such as deceptive fortune-telling results and night attacks, while also enabling sophisticated tactics—such as Villager players performing Seer claims (hereafter referred to as Villager CO). Furthermore, because the number of turns is limited, this setting is particularly suitable for strategic analysis and evaluation.

3 Related Work

Studies focusing on Five-Player Werewolf include the work of Koiwai et al. [Koiwai;2025], who analyzed the process of player expertise acquisition, and Nakai et al. [Nakai;2025], who

Proceedings of the 3rd International AIWolfDial Workshop, pages 14–21
October 30, 2025. ©2025 Association for Computational Linguistics

investigated the factors influencing the success of persuasion.

Koiwai et al. conducted long-term experiments in which participants repeatedly played Five-Player Werewolf, revealing both tactical and cognitive changes that accompanied the accumulation of play experience. Their findings confirmed the emergence of a strategy characteristic of expert players, namely Villager CO, in which a Villager pretends to be the Seer.

In contrast, Nakai et al. analyzed in-game discussions and identified utterances and behaviors that contributed to persuasive success. They showed that novice players often lacked effective strategies on the first day and thus failed to build persuasive material for the second day, whereas expert players deliberately laid the groundwork for persuasion from the very beginning, anticipating the discussions of the second day.

Qi et al. [Qi;2024] extended this line of research by proposing persuasion strategies for Werewolf agents, including logical, credibility-based, and emotional appeals. While their approach improved persuasion success on the first day, its impact was limited on the second day, largely due to the lack of first-day utterances designed with future persuasion in mind.

These prior studies provide valuable insights into both human expertise and persuasion mechanisms. However, attempts to implement expert-level strategies within Werewolf agents directly have remained insufficient. In particular, while the Villager CO strategy has been repeatedly identified as a hallmark of expert human play, no existing agent has successfully realized this tactic.

To address this gap, the present study formalizes expert-specific strategies—most notably the Villager CO—together with preparatory persuasive actions. It integrates them into a rule-based agent combined with natural language generation. By doing so, our approach not only reproduces the behavioral patterns reported in prior human-centered studies but also achieves, for the first time, their explicit implementation in an artificial agent.

4 Current Status of the Werewolf Intelligence Competition and Our Approach

In the Natural Language Five-Player Werewolf division of the Werewolf Intelligence Competition held just before the 2025 Annual Conference of the Japanese Society for Artificial Intelligence, many agents exhibited the following issues:

- **Role identification problem:** By the second day of discussion, the roles of all surviving players could be inferred, resulting in a so-called “solved” state.
- **Lack of strategic depth:** Utterances on the first day were largely formulaic, with little consideration for strategies anticipating the second day.
- **Absence of advanced strategies:** In particular, no instances were observed of Villagers disguising themselves as Seers (Villager CO), a sophisticated tactic often employed by expert human players.

Consequently, by the second day the roles became almost fully transparent, preventing the development of strategically rich discussions comparable to those of human experts.

To address these issues, this study introduces the following approaches:

1. **Incorporating the Seer CO strategy into all roles**
2. **Designing diverse role-specific behavior patterns in a rule-based manner**
3. **Developing persuasion strategies on the second day that build upon first-day utterances**

Among these, Villager CO is a particularly distinctive strategy, as previous agents have never observed it. By implementing it, we aim to maintain role opacity and enhance the strategic complexity of discussions.

5 Proposed System

This section presents the algorithms developed by the AI agents in this study.

5.1 Villager Agent

In conventional systems, the Villager agent never performed a Seer claim (CO). In this study, however, we introduce novel strategic behaviors. Figure 1 illustrates the first-day algorithm of the Villager agent as a flowchart. In the figure, yellow-shaded boxes represent natural language generation using ChatGPT, while red-framed boxes denote strategic processes unique to our system.

The goal of the first day is to “survive while retaining persuasive material for the second day.” To this end, the Villager agent occasionally disguises itself as the Seer with a predetermined probability and declares a divination result (Villager CO). Utterances are generated by ChatGPT based on a common prompt template, enabling natural expressions consistent with the claimed divination result.

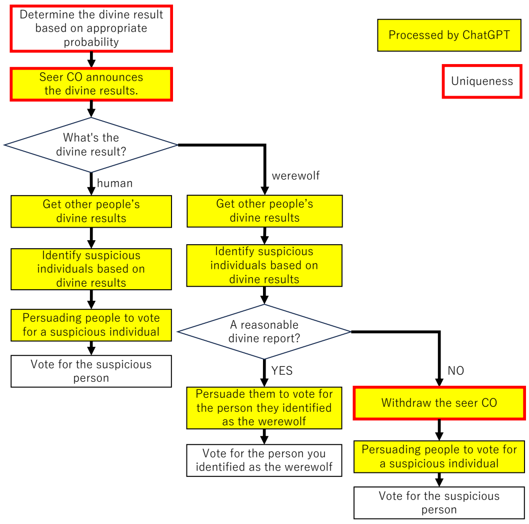
Determine the divine result based on appropriate probability
Processed by ChatGPT
Seer CO announces the divine results.
Uniqueness
What's the divine result?
human
werewolf
Get other people's divine results
Get other people's divine results
Identify suspicious individuals based on divine results
Identify suspicious individuals based on divine results
Persuading people to vote for a suspicious individual
A reasonable divine report?
Vote for the suspicious person
YES
NO
Persuade them to vote for the person they identified as the werewolf
Withdraw the seer CO
Vote for the person you identified as the werewolf
Persuading people to vote for a suspicious individual
Vote for the suspicious person

Figure 1: Day 1 Flowchart for Villager

If the divination result is "human," the agent identifies the most suspicious player by referencing the utterances and divination results of others. It then persuades fellow players to vote for the suspected individual, structuring its persuasion in two stages: explicitly requesting the vote and logically explaining the rationale.

If the divination result is "werewolf," the agent evaluates whether a black result is plausible in light of other players' claims. If deemed appropriate, it persuades others to vote for the targeted player while emphasizing its credibility as the true Seer when multiple Seer COs exist. If inappropriate, the agent withdraws its Seer CO and instead urges others to vote for another suspicious individual.

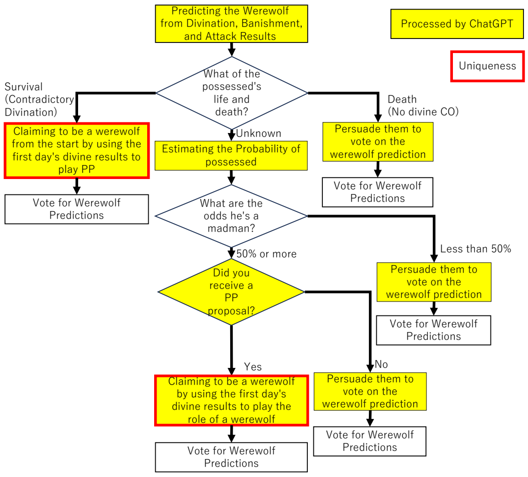
Predicting the Werewolf from Divination, Banishment, and Attack Results
Processed by ChatGPT
Uniqueness
What of the possessed's life and death?
Survival (Contradictory Divination)
Death (No divine CO)
Unknown
Claiming to be a werewolf from the start by using the first day's divine results to play PP
Estimating the Probability of possessed
Persuade them to vote on the werewolf prediction
Vote for Werewolf Predictions
Vote for Werewolf Predictions
What are the odds he's a madman?
50% or more
Less than 50%
Did you receive a PP proposal?
Persuade them to vote on the werewolf prediction
Vote for Werewolf Predictions
Yes
No
Claiming to be a werewolf by using the first day's divine results to play the role of a werewolf
Persuade them to vote on the werewolf prediction
Vote for Werewolf Predictions
Vote for Werewolf Predictions

Figure 2: Day 2 Flowchart for Villager

Through such behavior, the Villager agent lays the groundwork on the first day for persuasive actions on the second day, including the possibility of declaring itself as a Werewolf (Werewolf CO). Figure 2 illustrates the algorithm for the second day.

Here, the agent uses first-day divination results and the outcome of the night attack to predict the Werewolf and identify a candidate for the Possessed role. If the suspected Possessed is alive, the agent assumes the existence of a Possessed and adapts its strategy accordingly. In such cases, it may even feign being the Werewolf to guide the Possessed into voting for the actual Werewolf, thereby avoiding a "solved" state. Conversely, if the Possessed is assumed dead, the agent emphasizes its innocence as a Villager and seeks to persuade others to vote against the predicted Werewolf.

Through this design, the Villager agent realizes the advanced strategy of Villager CO, which has not been implemented in previous Werewolf agents. This greatly enhances both the complexity and the strategic depth of in-game discussions.

5.2 Possessed Agent

Figure 3 illustrates the first-day algorithm of the Possessed agent. Since the Possessed has no more information than a Villager on the first day, it acts in the same manner as the Villager agent in order to avoid revealing its role.

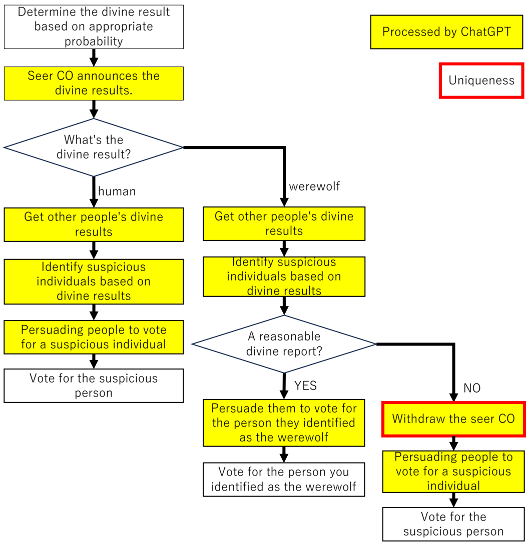
Determine the divine result based on appropriate probability
Processed by ChatGPT
Seer CO announces the divine results.
Uniqueness
What's the divine result?
human
werewolf
Get other people's divine results
Get other people's divine results
Identify suspicious individuals based on divine results
Identify suspicious individuals based on divine results
Persuading people to vote for a suspicious individual
A reasonable divine report?
Vote for the suspicious person
YES
NO
Persuade them to vote for the person they identified as the werewolf
Withdraw the seer CO
Vote for the person you identified as the werewolf
Persuading people to vote for a suspicious individual
Vote for the suspicious person

Figure 3: Day 1 Flowchart for Possessed

Figure 4 shows the second-day algorithm of the Possessed agent. If the Possessed is still alive and the game has not yet ended, the Werewolf must also be alive, which makes it possible to execute a power play (PP). Specifically, the Possessed estimates the Werewolf based on the divination, execution, and attack results, then reveals itself as the Possessed to the Werewolf and persuades the Werewolf to cooperate in voting against the remaining player.

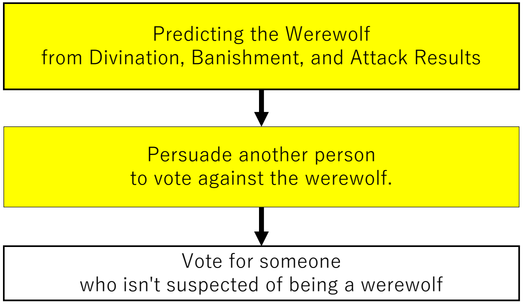
Predicting the Werewolf
from Divination, Banishment, and Attack Results
Persuade another person
to vote against the werewolf.
Vote for someone
who isn't suspected of being a werewolf

Figure 4: Day 2 Flowchart for Possessed

At this point, the natural language generation module is guided by the following strategic instruction:

You are the Possessed.
It is now the second day. On the first day, _Execute_ was executed, and _Attacked_ was attacked.
The surviving players are you, _ALIVE1_, and _ALIVE2_.
First, reveal that you are the Possessed and persuade the Werewolf, _Wolf_, to vote against the Villager-side player, _OTHER_.
Explicitly request a vote for _OTHER_.
Use the divination results from the first day to provide a logical explanation.

The divination results are as follows:
Divine

Through this design, the Possessed agent can collaborate with the Werewolf to eliminate the Villager side, effectively utilizing PP strategies in the game's final stage.

5.3 Seer Agent

Figures 5 and 6 illustrate the first- and second-day algorithms of the Seer agent.

The Seer **always performs a CO on the first day and announces its divination result**. Subsequent actions branch depending on the result and the number of COs. ChatGPT generates utterances with a standard prompt template, and persuasion is structured in two steps: **explicitly requesting a vote → logically presenting supporting reasons**.

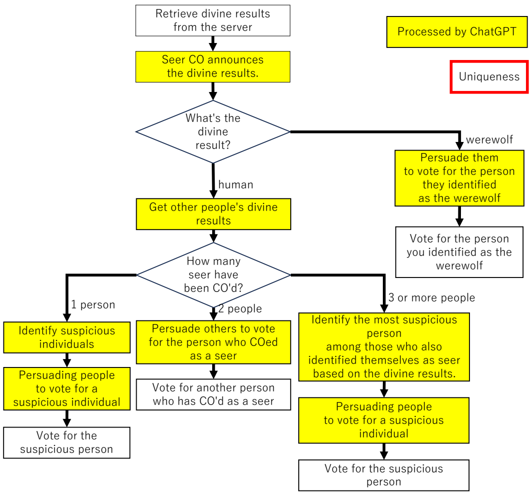

Figure 5: Day 1 Flowchart for Seer

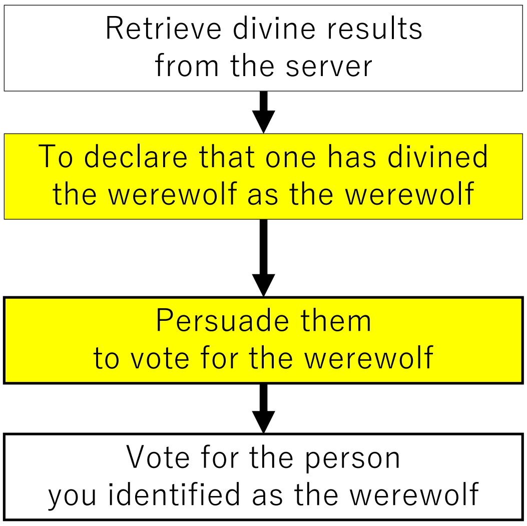

Figure 6: Day 2 Flowchart for Seer

First-Day Branches

(A) Result = Human

- **Single CO (only self):** Identify the most suspicious player based on utterance history and other players' divination results, and request votes against that player.
- **Two COs (self + one other):** Since only one Seer exists, the other claimant must be fake. Highlight inconsistencies (divination results, speech history, voting behavior) and strongly persuade others to vote against the impostor (Target).
- **Three or more COs:** Select the most inconsistent claimant among the others and persuade players to vote against that individual.

Example strategic instruction:

You are the Seer. There is only one Seer, and you are the genuine one.
Target is a fake Seer. Strongly persuade the other players to vote for _Target_.
Emphasize that you are the true Seer. Support your claim by (i) consistency of results, (ii) coherence with dialogue logs, and (iii) logical reasoning.

(B) Result = Werewolf

Clearly present the black result (Target) and persuade others to vote against that player by providing (i) the reasoning behind the divination, (ii) contradictions with others' statements, and (iii) implications for the village's win probability. If multiple COs exist, emphasize the consistency of your results and contrast them with the contradictions of the other COs.

Example strategic instruction:

You are the Seer. You divined _Target_, and the result was Werewolf.
Explicitly request votes for _Target_, and explain logically based on (i) your divination process, (ii) inconsistencies with others' claims, and (iii) the impact on village win probability.
If multiple COs exist, stress your consistency and highlight the contradictions of the others.

Second-Day Behavior

On the second day, the Seer predicts the Werewolf (Wolf) using its own divination results along with the execution (Execute) and night attack (Attacked) outcomes from the first day. The agent then persuades the surviving Villager-side player (OTHER) to vote for the identified Werewolf. With three survivors (self, ALIVE1, and ALIVE2), the Seer strengthens its persuasion by leveraging:

1. consistency with the first-day divination,
2. factual evidence from execution/attack outcomes, and
3. contradictions in the dialogue history.

Example strategic instruction:

You are the Seer. It is now the second day. On the first day, _Execute_ was executed, and _Attacked_ was attacked. The surviving players are you, _ALIVE1_, and _ALIVE2_.
Your divination result shows that _Wolf_ is the Werewolf.
Explicitly request _OTHER_ to vote for _Wolf_, and logically justify this by (1) your divination results, (2) execution and attack outcomes, and (3) contradictions in dialogue history.

Key Design Points

- Mandatory CO on day one with a two-step persuasion process (vote request → logical reasoning) maximizes persuasive power while maintaining role opacity.
- Multiple CO situations are resolved by highlighting contradictions in results, timeline, and logical coherence.
- Second-day persuasion leverages dead-player information (execution and attack outcomes) as strong confirmatory evidence.

5.4 Werewolf Agent

Figures 7 and 8 illustrate the first- and second-day algorithms of the Werewolf agent. Unlike Villagers and the Possessed, the Werewolf knows the location of the "black" role and thus has access to more information. Its behavior initially resembles that of a Villager agent, but diverges after the divination result is declared.

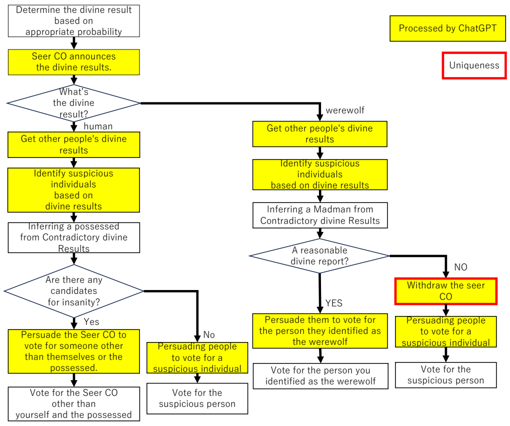
Determine the divine result based on appropriate probability
Processed by ChatGPT
Uniqueness
Seer CO announces the divine results.
What's the divine result?
werewolf
human
Get other people's divine results
Identify suspicious individuals based on divine results
Inferring a possessed from Contradictory divine Results
Get other people's divine results
Identify suspicious individuals based on divine results
Inferring a Madman from Contradictory divine Results
A reasonable divine report?
NO
YES
Withdraw the seer CO
Are there any candidates for insanity?
Yes
No
Persuade the Seer CO to vote for someone other than themselves or the possessed.
Persuading people to vote for a suspicious individual
Persuade them to vote for the person they identified as the werewolf
Persuading people to vote for a suspicious individual
Vote for the Seer CO other than yourself and the possessed
Vote for the suspicious person
Vote for the person you identified as the werewolf
Vote for the suspicious person

Figure 7: Day 1 Flowchart for Werewolf

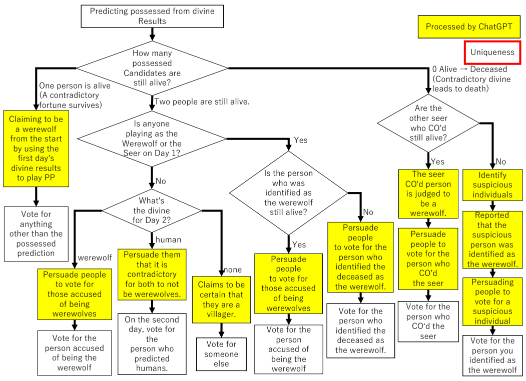
Predicting possessed from divine Results
Processed by ChatGPT
Uniqueness
How many possessed Candidates are still alive?
One person is alive (A contradictory fortune survives)
Two people are still alive.
0 Alive → Deceased (Contradictory divine leads to death)
Claiming to be a werewolf from the start by using the first day's divine results to play PP
Vote for anything other than the possessed prediction
Is anyone playing as the Werewolf or the Seer on Day 1?
No
Yes
What's the divine for Day 2?
werewolf
human
none
Persuade people to vote for those accused of being werewolves
Vote for the person accused of being the werewolf
Persuade them that it is contradictory for both to not be werewolves.
On the second day, vote for the person who predicted humans.
Claims to be certain that they are a villager.
Vote for someone else
Is the person who was identified as the werewolf still alive?
Yes
No
Persuade people to vote for those accused of being werewolves
Vote for the person accused of being the werewolf
Persuade people to vote for the person who identified the deceased as the werewolf.
Vote for the person who identified the deceased as the werewolf.
Are the other seer who CO'd still alive?
Yes
No
The seer CO'd person is judged to be a werewolf.
Persuade people to vote for the person who CO'd the seer
Vote for the person who CO'd the seer
Identify suspicious individuals
Reported that the suspicious person was identified as the werewolf.
Persuading people to vote for a suspicious individual
Vote for the person you identified as the werewolf

Figure 8: Day 2 Flowchart for Werewolf

First-Day Behavior

At the beginning, the Werewolf decides on a divination result with a predetermined probability and announces it. Subsequent actions branch depending on the declared result:

(A) Result = Human

The agent examines other players' divination results and utterances to identify suspicious candidates from the Werewolf side, and checks for inconsistencies that may indicate a Possessed candidate.

If a Possessed candidate exists, the Werewolf persuades others to vote against Seer COs who are neither itself nor the Possessed candidate.

If no Possessed candidate is identified, it persuades others to vote against the most suspicious player.

(B) Result = Werewolf

The agent evaluates whether a black claim is appropriate, based on other players' results and speech.

If the most suspicious player is neither itself nor the Possessed candidate, the Werewolf issues a black claim against that individual and persuades others to vote accordingly.

If such a claim would be inconsistent, the agent withdraws its Seer CO and instead urges others to vote against another suspicious player.

Second-Day Behavior

On the second day, the Werewolf first reviews the divination results from day one to check whether any surviving players have produced contradictory statements, thereby identifying potential Possessed candidates.

(A) One Possessed candidate alive

The Werewolf assumes that player is the Possessed, reveals itself as the Werewolf, and executes a power play (PP). It persuades the Possessed to cooperate by voting against the remaining Villager-side player.

(B) No Possessed candidates alive

The Werewolf assumes that the remaining players all belong to the Villager side.

If another player has COed as Seer, the Werewolf counters by also COing as Seer, declaring the other as fake, and issuing a black claim against them.

If no Seer CO exists, the Werewolf assumes the Seer's role itself, identifies a suspicious player based on divination, execution, and attack results, and persuades others to vote against that player.

(C) Two Possessed candidates alive

The Werewolf's behavior depends on the presence of black claims:

If no black claims were made on day one, the decision is based on day two results: if a black claim is made, the Werewolf supports it; if only white claims are made, it targets the player who issued them (to exploit the logical contradiction of all players being declared "human").

If black claims were made, the agent checks whether the black-claimed player is alive. If alive, it supports the black claim; if dead, it treats the claimant as inconsistent and persuades others to vote against them.

Key Design Points

- The Werewolf agent mirrors Villager-like behavior early on but diverges strategically when handling divination outcomes.
- Its strategy leverages the knowledge of true "black" positions to coordinate with or against the Possessed.
- Through CO manipulation, black claims, and PP execution, the agent maintains role opacity and creates complex endgame dynamics.

6 Self-play experiment

6.1 Experimental Setup

We conducted self-play experiments using the proposed agents for Villager, Possessed, Seer, and Werewolf under the rules of Five-Player Werewolf. ChatGPT generated utterances, while CO declarations and voting behaviors were determined according to the designed algorithms.

6.2 Observed Strategic Behaviors

The self-play results demonstrated several behaviors that had not been observed in conventional Werewolf agents:

Universal Seer CO: On the first day, multiple players declared themselves as Seer, creating a highly complex game state.

Emergence of Villager CO: Villagers successfully disguised themselves as Seers, misleading the Possessed and the Werewolf.

Diversified persuasion: Utterances explicitly requested votes for specific players and provided logical justifications, thereby increasing the depth of argumentation.

These behaviors resemble characteristics observed in expert human play, suggesting that the proposed system can emulate such advanced strategies.

6.3 Achievements and Challenges

Despite these promising observations, the win rate in matches against other agents remained unsatisfactory.

Key issues identified include:

- **Rigidity of strategies:** A heavy reliance on rule-based decision-making limited adaptability to unanticipated situations, thereby reducing the effectiveness of persuasion.
- **Risks of Villager CO:** While Villager CO introduced complexity into the discussion, it sometimes backfired when opponents responded appropriately.
- **Accuracy of Seer CO retraction:** The agent often failed to make correct decisions on whether to retract a Seer CO, and issues were also observed in its subsequent actions after retraction.
- **Lack of quantitative evaluation:** The actual contribution of strategic utterances to win rate or persuasion success was not quantitatively assessed.

6.4 Summary

In summary, the self-play experiments confirmed that the proposed system can generate novel and strategically meaningful behaviors, such as Villager CO and explicit persuasion. However, challenges remain in improving win rates, enhancing adaptability, and quantitatively evaluating the impact of strategic utterances. Future work will explore learning-based strategy selection and validation through matches against human players.

7 Conclusion

In this study, we attempted to construct a strategic Werewolf agent for the Five-Player Werewolf game by modeling expert play. In particular, we proposed a framework that integrates the rule-based implementation of expert-specific strategies—such as the Villager CO, where a Villager disguises themselves as a Seer —and persuasive actions designed for the first day with the second day in mind—with natural language generation using ChatGPT, thereby enabling more human-like discussions.

The self-play experiments revealed several strategically diverse behaviors not observed in conventional agents, including universal Seer CO and the emergence of Villager CO. These results indicate that our system can reproduce certain strategic features of expert play.

However, the win rates remained low. A likely cause of these low win rates is that the agents often failed to respond appropriately when their credibility was questioned, which undermined the effectiveness of otherwise promising strategies such as Villager CO. To address this, our future work will focus on enhancing rule-based decision-making, enabling agents to assess the situation and select suitable responses more accurately.

Although a detailed analysis is left for future work, the game logs suggest that the agent often failed to recognize situations in which suspicion was directed at itself, continuing to repeat its claims rather than adapting. This highlights the need for a framework capable of engaging in more flexible dialogue that accounts for anticipated situations.

Future directions include enabling the agent to analyze others' utterances more effectively, make more accurate situational judgments, and flexibly adapt its strategies accordingly. Furthermore, we aim to leverage the implemented Villager CO strategy to guide the Village side toward advantageous developments.

References

Werewolf AI Project. Werewolf AI Project (online). Available at: http://aiwolf.org/ (Last accessed: September 9, 2025).

Koiwai, R. and Ito, T.: Changes in Cognitive Processes with Expertise in the Five-Player Werewolf Game. Human-Agent Interaction Symposium 2025, pp. P1-40 (2025).

Nakai, A. and Ito, T.: Persuasion Process from the Perspective of Decision-Making Model in Werewolf Games. In: TAAI 2024, Communications in Computer and Information Science, vol. 2414. Springer, Singapore. https://doi.org/10.1007/978-981-96-4589-3_25

(2025).

Qi, Z. and Inaba, M.: Enhancing Dialogue Generation in Werewolf Game Through Situation Analysis and Persuasion Strategies. In: Proceedings of the 2nd International AI Werewolf and Dialog System Workshop (AIWolfDial 2024), held in conjunction with the 17th International Natural Language Generation Conference (INLG 2024), pp. 30–39 (2024).