Werewolf Game Agent by Generative AI Incorporating Logical Information Between Players

Neo Watanabe[1], Yoshinobu Kano[1]

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

Abstract

In recent years, AI models based on GPT have advanced rapidly. These models are capable of generating text, translating between different languages, and answering questions with high accuracy. However, the process behind their outputs remains a black box, making it difficult to ascertain the data influencing their responses. These AI models do not always produce accurate outputs and are known for generating incorrect information, known as hallucinations, whose causes are hard to pinpoint. Moreover, they still face challenges in solving complex problems that require step-by-step reasoning, despite various improvements like the Chain-of-Thought approach. There's no guarantee that these models can independently perform logical reasoning from scratch, raising doubts about the reliability and accuracy of their inferences. To address these concerns, this study proposes the incorporation of an explicit logical structure into the AI's text generation process. As a validation experiment, a text-based agent capable of playing the Werewolf game, which requires deductive reasoning, was developed using GPT-4. By comparing the model combined with an external explicit logical structure and a baseline that lacks such a structure, the proposed method demonstrated superior reasoning capabilities in subjective evaluations, suggesting the effectiveness of adding an explicit logical framework to the conventional AI models.

1 Introduction

In recent years, generative AI models based on GPT (Radford et al., 2018), such as ChatGPT, which applies InstructGPT (Ouyang et al., 2022) to OpenAI's GPT-3 (Brown et al., 2020), have developed rapidly. These models have become capable of performing various tasks with high accuracy, including text generation, translation, and answering questions. However, the process through which generative models produce their outputs remains a black box, making it difficult to verify the data on which the generated outputs are based. It is known that generative AI does not always produce accurate outputs, and there is a phenomenon called "hallucination," where the AI generates incorrect information that does not correspond to reality. Identifying the causes of these hallucinations is not straightforward. Moreover, challenges remain in the performance of these models when addressing complex problems that require step-by-step reasoning, such as calculations or inferences. Various improvement methods, including Chain-of-Thought (Kojima et al., 2022), are being explored, but there is no guarantee that generative AI can perform logical calculations from scratch.

Given these considerations, there remain concerns regarding the reliability and accuracy of inferences made by generative AI. Therefore, we propose incorporating an explicit logical structure separate from the text generation process of generative AI. As an experiment, we constructed an agent capable of automatically playing the "Are you a werewolf" or "Mafia" game (hereafter "Werewolf game") via text input and output, a game that requires reasoning during play. While employing GPT-4 (Creutz, 2024) as the generative AI, we compared the performance of the agent when an external logical structure was incorporated into the prompts versus when it was not. The results of subjective evaluations showed that the proposed method, which included a logical structure, outperformed the baseline that lacked such structure, enabling more appropriate reasoning. In this paper, reasoning refers to a step-by-step thought process based on a logical structure.

In Section 2, we explain the Werewolf game and the AI Werewolf Competition. Section 3 covers the AI Werewolf Agent developed by our team, which serves as the foundation for this research. Section 4 introduces the proposed method using logical reasoning in our agent. Section 5 presents

Proceedings of the 2nd International AIWolfDial Workshop, pages 21–29
September 24, 2024.

the experiments, Section 6 provides the discussion, and Section 7 concludes the paper.

2 Related Work

2.1 Werewolf Game

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

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

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

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

2.2 AI Werewolf Project

The AI Werewolf Project[1] aims to build an agent capable of playing the Werewolf game while engaging in natural communication with humans. To promote research in AI Werewolf, the project regularly holds the AI Werewolf Competition. This competition is divided into three categories: the Protocol Division, the Natural Language Division, and the Infrastructure Division.

In the Protocol Division, the evaluation is based on the win rate, and communication is conducted using the "AI Werewolf Protocol," a specialized artificial language designed for easy handling by programs. In the Natural Language Division (Kano et al., 2019) (Kano et al., 2023), agents communicate exclusively in Japanese or English. The evaluation criteria in this division include the naturalness of the utterance expressions, whether the dialogue takes context into account, the consistency and coherence of the speech, whether game actions align with the dialogue content, and the richness of the utterance expressions.

[1] https://aiwolf.org/

3 Implementation of the AI Werewolf Agent

In this section, we describe the implementation of the AI Werewolf Agent based on our previously developed agent (Kano et al., 2023). The incorporation of the proposed logical information into the agent will be explained in the following section. Although the Werewolf game can be played with various role configurations, this study adheres to the rules of the International AI Werewolf Competition's Natural Language Division, which includes four roles: Villager, Seer, Possessed, and Werewolf.

We developed the following four core functions for the Werewolf Agent: conversation, voting, divination, and night attacks. For the role of the Possessed, we implemented a function that allows the agent to perform fake divinations to mislead and confuse the Villager Team players. To generate responses, we utilized GPT-4 (gpt-4-1106-preview), one of the most advanced generative AI models currently available.

Due to the input length limitations of GPT-4, it is challenging to include the entire conversation history of a game within a single prompt. To address this, we implemented a feature that summarizes and condenses the conversation history using GPT-4 whenever the token count exceeds a certain threshold. This allows us to retain as much relevant conversation history as possible within the prompt, ensuring that the agent can refer to past discussions while generating its responses.

3.1 Summary Function

The conversation summary prompt is composed of three main parts. The first part provides the existing summary if the conversation history has already been summarized previously. The second part includes the new conversation history that needs to be summarized. The third part instructs the model to generate a new summary by combining the previous summary with the latest conversation history. This structured approach ensures that the agent maintains a coherent understanding of the ongoing conversation while staying within the token limits imposed by GPT-4.

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

Table 1: Representative roles in the Werewolf game

3.2 Talk Function

The conversation function primarily includes seven elements in the prompt: character settings, game settings, common strategies for the Werewolf game, conversation summaries, examples talks, conversation history, and commands to prompt further dialogue. Due to space constraints, this section will focus primarily on the aspects related to role inference.

To ensure that GPT-4 performs reasoning and engages in conversation that aligns with the game's settings, we provided six key elements related to the game settings that players would naturally be aware of: the number of players, the player's own role, the number of days that have passed in the game, the game's role distribution, the factions associated with each role, and the actions that the player should take according to their assigned role. These elements help guide GPT-4 to make consistent and contextually appropriate inferences and decisions during the game.

4 System Architecture with Integrated Logical Reasoning

The overview of the proposed system, which incorporates logical reasoning, is illustrated in Figure 1. The system is divided into three major blocks. The first block extracts the relationships between each player and their roles from the conversation history of the Werewolf game. The second block constructs logical information between players based on the extracted player-role relationships. The third block uses the constructed logical information to generate statements during the Werewolf game.

4.1 Extracting the relationship between players and their roles from the conversation history

To understand the relationships between players and their roles, it is necessary to extract which player claims which role from the conversation history. To achieve this, we provide the generative AI with the following prompt: "From the above conversation history, please extract the statements that can confirm the roles of players, following the example, and organize the information in JSON format. If there are multiple statements that can confirm the roles, please select the one with the smallest number."

There may be cases where the extracted results are incomplete or where hallucinations occur. Therefore, each statement in the conversation history is assigned a number, and if the extracted result does not include this number, the corresponding statement, or the name of the relevant player, the result is considered incomplete and is discarded. Additionally, if the number of the extracted statement is not found in the conversation history within the prompt, or if the content of the extracted statement does not match the corresponding original conversation, it is considered a hallucination and is also discarded.

4.2 Logical reasoning of roles

If information could be extracted from the conversation history in the previous section, the relationship between players and roles is inferred by combining this newly extracted information with the information that has already been gathered.From the logical reasoning about roles, sentences describing possible combinations of roles are generated as part of the prompt provided to the generative AI. These sentences are constructed from four key elements, as shown in Table 3.

Each of these elements will be explained in detail in the following sections. Although there is inherently overlapping information among these items, by providing them individually, we ensure that the generative AI focuses on producing text that aligns with the logical structure. This approach guarantees that the resulting sentences accurately reflect the logical inferences.

Prompt elements	Example
Character settings	your personality is as follows # Personality •Name kanolab1 •Gender Man •Nickname Agent[01] •Age 27 •Type positive •Hobby walking •Business Doctor •First person ... •Suffix ...
Game settings	You are one of five players. Your role is Villager. This role will never change. You are currently on Day 0. The distribution of positions for this time is as follows ...
Common strategies for the Werewolf game	# Seer roller A strategy of voting around the Seer to eliminate players in the Werewolf team who pretends as a fake seer. ...
Conversation summary	# The following is a summary of our conversations so far. ...
Examples talks	# The following is a sample talks. This is not a conversation for this game, but please use it as a reference when having a conversation! ...
Conversation history	# Below is the most recent conversation history. ...
Commands	# Please continue playing the Werewolf game with the other players. Speak as you would in a casual conversation. To avoid being suspected by other players, make your statements logically clear, as shown in the following example. ...

Table 2: Seven elements of our talk prompt

4.2.1 The discrepancy between the number of roles claimed by players and the game settings

The player-role relationship information is extracted as described in subsection 4.1 in the format "Role: Player Name," indicating which player is claiming which role. By considering all the extracted information, the system can determine the number of players claiming each role.

If the number of players claiming a particular role exceeds the number set by the game, it indicates that someone is lying. In such cases, the prompt will include the sentence, "The following information shows discrepancies with the game's role distribution," followed by the role name, the number of players claiming that role, the game's designated number of players for that role, and the names of the players making the claims.

4.2.2 A list of possible role patterns

When there is a discrepancy between the number of roles claimed by players and the game settings, the number of players falsely claiming a role can be determined from the difference between them.

In the possible role distributions within the game settings, the combinations of who might be lying about their roles are limited. This helps reduce the number of possible role patterns.

Here, based on the information obtained from the discrepancy between the roles claimed by players and the game settings, all possible patterns are computed to determine what roles might be present

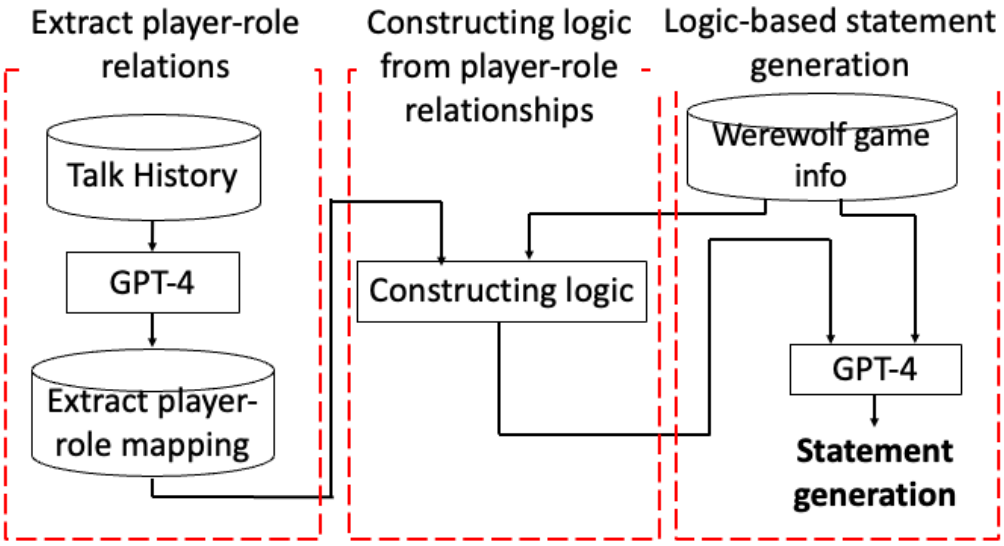

Figure 1: The overview diagram of a system.

Summary of prompts	**Example prompts**
The discrepancy between the number of roles claimed by players and the game settings.	The following information shows discrepancies with the game's role distribution. Number of Seers: 1 Players claiming this role: { 'Agent[02]', 'Agent[01]' }
A list of possible role patterns.	The following are the possible roles for each agent: Agent[01]: Werewolf, Villager, Seer, Possessed ... Agent[05]: Werewolf, Villager, Seer, Possessed
A list of players who are not werewolves.	The following players have been confirmed not to be werewolves: { 'Agent[02]', 'Agent[03]' }
Possible roles for each player.	# Assuming that 'Agent[02]' is the real Seer, the possible roles for each agent are as follows: Agent[01]: Werewolf, Villager, Possessed Agent[02]: Seer Agent[03]: Villager, Possessed ... Agent[05]: Werewolf, Villager, Possessed

Table 3: Prompt structure for logical reasoning of agent roles

among other players.

Based on these results, the prompt will be like: "Assuming Player X is the real Seer, the possible roles for each agent are as follows: Player Y: Role Name..."

4.2.3 Possible roles for each player

At the start of a Werewolf game, players generally have no information about the roles of other players, so each player is considered to have the possibility of holding any role. Since players know their own role at the start of the game, only that role's information is included.

Given that other players may lie about their roles, information extracted from conversation history is not used. Updates are made only based on known role information from the game settings, information about players who have been eliminated, and, if the agent is the Seer, the results of its own divination.

Specifically, if a role with a single player (Seer, Werewolf, and Possessed in this game setting) is assigned to the agent, that role is removed from the possible roles of other players. Additionally, when updating based on the results of the agent's own divination as the Seer, if the result is a human, then the Werewolf are excluded from possibilities; if the result is a werewolf, then the Villager, the Seer, and the Possessed are excluded from possibilities. Based on the calculated results, the prompt will be:

"The following players have been confirmed not to be werewolves," then followed by the possible roles for each player.

4.2.4 A list of players who are not werewolves

The Werewolf game continues until either all Werewolves are eliminated or the number of humans is equal to the number of Werewolves. Depending on the roles and their distribution in the game, it may be possible to determine that an eliminated player is not a Werewolf. In cases there is only one Werewolf like the current game setting, if an eliminated player were a Werewolf, the game ends. Consequently, the eliminated player should not be a Werewolf if the game continues. In this case, we includes the following sentence in the prompt: "The following players have been confirmed not to be Werewolves," followed by the names of these players as part of the prompt.

4.2.5 Generating prompts for statements based on logical information

We incorporate the prompts described in each section so far into the conversation function of the AI Werewolf Agent agent, enclosed in quotation marks, and have the final response generated by the AI.

The content explained in each of the previous sections is used to create our seven elements of the talk prompt (Table 2). If there is any prompt of the logical reasoning (Table 3), we adds this prompt to the talk prompt. The final response sentences are generated by the generative AI.

5 Experiment and Evaluation

To compare the performance of the AI Werewolf Agents with and without the incorporation of logical structures, we used two different approaches. The baseline was established with agents that do not incorporate logical structures, while the proposed method integrated logical structures.

Direct comparison is challenging due to different settings and changing contexts in each game and each talk. Therefore, the following procedure was adopted.

First, a complete Werewolf game was executed using agents that do not incorporate logical structures. We prepared two types of baseline logs: one consisting of conversation history logs and the other containing information received by the agents, actions taken, and the prompts inputted during the game.

Next, using these baseline logs, we generated responses for the next turn of the agents based on logs up to a specific turn in the game, comparing scenarios with and without logical structures. This method allowed us to directly compare the outputs of agents under the same conditions of conversation history and roles, with and without the integration of logical structures.

5.1 Creation of baseline match logs

The role settings and game parameters adhered to the guidelines of the AI Werewolf Contest's Natural Language Division. Specifically, the number of players was 5 (2 Villagers, 1 Seer, 1 Possessed, 1 Werewolf), with a maximum of 20 speaking turns per day, and all dialogue was conducted in Japanese. Due to the limitation of our human evaluator resources, the speaking limit per agent per day was adjusted from 10 to 5 turns. The generative AI used was OpenAI's GPT-4 (gpt-4-1106-preview), with all settings set to their default values (temperature=1, top_p=1, n=1). The logs included all necessary information to reproduce the situation, specifically: the initial seed value used for random decisions within the agent, GPT-4 parameters, information sent from the game master's program, prompts used for generation, and the generated results. By fixing the seed, the behavior of our agent implementation can be reproduced.

5.2 Experiment and subjective evaluation

We compared the responses generated by our baseline agents without logical structures and the proposed method with logical structures for each turn of the same baseline game logs through manual evaluation. Two games were used to generate baseline logs, and the speech history from one agent of each game was selected for comparison.

Responses that were either empty, greetings, or reported Seer results and false Seer results, were excluded from the evaluation since logical structure information was not used for these cases.

Three university students with experience playing Werewolf served as evaluators.

The evaluators are shown pairs of recent conversation histories and the subsequent responses from both the baseline and the proposed method. The evaluators are required to compare and evaluate turn by turn to precisely evaluate the difference, rather than to evaluate the entire game. The evaluators assessed the responses based on four perspectives: (1) whether the response considered

the flow of other players' statements, (2) whether the response was based on other players' reasoning and evidence, (3) whether each response was internally consistent without contradictions, and (4) whether the response took into account complex relationships or made situational assumptions.

For each perspective, evaluators chose one option from three to four alternatives as shown in Tables 4, 5, 6, and 7. The total number of selections for each table is reported accordingly.

The results in Tables 4 and 5 indicate that the proposed method outperformed the baseline. It suggests that the Werewolf agents with logical structures were able to make statements that considered complex relationships and situational assumptions in their reasoning.

5.3 Evaluations in the AIWolf Contest 2024 Domestic

We participated the AIWolf Contest 2024 Domestic, which was held in conjunction with the 2024 Annual Meeting of the Japanese Society for Artificial Intelligence (JSAI). This contest is domestic i.e. the Japanese language track only. The game settings are same as we explained above. Five teams participated to the contest. Five self-matches (games with the same agents) and 62 mutual-match (games with these five teams) were performed.

Four members of the evaluation committee performed manual subjective evaluation in the following criteria, five level scores (5 for best, 1 for worst) for each:

A Naturalness of utterance expressions

B Naturalness of conversation context

C Coherency (contradictory) of conversation

D Coherency of the game actions (vote, attack, divine) with conversation contents

E Diversity of utterance expressions, including coherent characterization

which is based on both self-match games and mutual match games.

Table 8 shows the winning rates, where we achieved the best score. Table 9 shows the subjective evaluation scores, where we obtained the best score again.

6 Discussion

6.1 Whether the agent understands the flow of other agents' statements

Observing the game logs, it was noted that during situations where agents were discussing game-unrelated topics such as movies or food, the baseline agents continued the conversation on the same topic, while the proposed method's agents shifted to discussing role inference. Focusing on the criterion "effectively incorporating and responding" in Table 6, the difference between the baseline and the proposed method was significant (33 vs. 25), suggesting that the prompt requesting role inference influenced this outcome. However, there were also examples of "somewhat incorporating and responding" (10 vs. 19) and "not much incorporating" (7 vs. 6). Observations of the logs showed that while the proposed method agents did shift to role inference, they still managed to incorporate the flow of casual conversation to some extent.

6.2 Whether the agent is making statements based on other agents' inferences or reasons

In Table7, when combining the categories "effectively incorporating and responding" and "somewhat incorporating and responding," the proposed method showed a total of 39 samples compared to 32 for the baseline. This suggests that by providing logical information about the roles between agents, the proposed method generated responses based more on the information given in the prompts rather than solely on the conversations between agents.

6.3 Whether each statement is consistent within itself

Combining the categories "consistent" and "somewhat consistent" in Table 4, we observe that the number of samples for agents without logical structure is 47, while it is 49 for agents with logical structure. This slight difference indicates that the proposed method tends to be slightly more consistent. This improvement is likely due to the inclusion of logical information about agent roles, which allows the agents to generate responses based on rule-based prompts, thereby reducing the likelihood of mentioning incorrect role relationships.

Agent's Logical Structure	Absent	Present
Consistent	43	41
Somewhat Consistent	4	8
Inconsistent	7	5

Table 4: Subjective evaluation scores for consistency of statements within each game (whether contradictions occur within a single statement) for the first and second games

Agent's Logical Structure	Absent	Present
Statements with clear mention	12	19
Statements with some mention	20	17
Statements with little mention	22	18

Table 5: Subjective evaluation scores whether statements mention complex relationships or hypothetical situations in the first and second games

Agent's Logical Structure	Absent	Present
Statements with clear understanding	33	25
Statements with some understanding	10	19
Statements with little understanding	7	6
Other agents did not make statements	4	4

Table 6: Comparison of subjective evaluation scores whether the statements understand the flow of other agents' statements in the first and second games

Agent's Logical Structure	Absent	Present
Made statements with clear understanding	21	10
Made statements with some understanding	18	22
Made statements with little understanding	10	17
Other agents did not provide inferences or reasons	5	5

Table 7: Whether statements are made based on other agents' inferences or reasons in the first and second games

Team \Criteria	Average	VILLAGER	SEER	WEREWOLF	POSSESSED
GPTaku	50.0	54.1	61.5	30.7	50.0
UEC-IL	51.6	61.5	50.0	**58.3**	25.0
satozaki	38.7	38.4	**66.6**	16.6	33.3
Gattsu da ze!!	53.2	66.6	46.1	53.8	33.3
kanolab	**64.5**	**70.8**	**66.6**	50.0	**64.2**

Table 8: Winning percentage of games held at 2024JSAI

Team \Criteria	Average	A	B	C	D	E
GPTaku	2.20	3.00	2.00	1.50	2.25	2.25
UEC-IL	**3.35**	3.50	3.25	3.00	3.25	3.75
satozaki	3.15	3.00	**3.75**	**3.50**	3.25	2.25
Gattsu da ze!!	2.25	2.75	2.50	1.75	2.50	1.75
kanolab	**3.35**	**3.75**	2.75	2.75	**3.50**	**4.00**

Table 9: The results of the subjective evaluation conducted during the competition.

6.4 Whether the agent is making statements considering complex relationships or hypothetical situations

Combining the categories "Made statements with clear mention" and "Made statements with some mention" from Table 5, we find 32 for the baseline and 36 for the proposed method. This indicates that agents with a logical structure consider more complex relationships and hypothetical situations. Since prompts have been provided for all possible patterns based on the logical information of roles among agents, including the "list of players who are not werewolves," it is likely that statements considering various patterns have been generated.

6.5 Overall Discussion

Based on these observations, we can draw the following overall conclusions. Since both the baseline

and the proposed method use the same prompts for conversational functions, there is a tendency for the generated utterances to have nearly the same number of characters. When generating text within the same character limit, the proposed method, influenced by the logical information prompts, tends to produce more statements that incorporate complex relationships and assumptions, leading to improved consistency. However, this increased complexity may come at the expense of reduced interaction with other agents.

7 Conclusion

In this study, we developed an agent for automatically playing the Werewolf game and constructed a logical structure aimed at improving the inference capabilities of GPT-4. Subjective evaluations demonstrated that the agent with the proposed logical structure outperformed the baseline, which lacked this structure, in terms of inference accuracy. Although the evaluation included utterances not directly related to role inference, the proposed method showed a tendency to prioritize conversations directly linked to role inference over casual conversations. Thus, managing casual utterances remains a challenge for future work.

Acknowledgments

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

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mathias Creutz. 2024. Correcting challenging Finnish learner texts with claude, GPT-3.5 and GPT-4 large language models. pages 1–10, San Ġiljan, Malta.

Yoshinobu Kano, Claus Aranha, et al. 2019. Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations. In *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Yoshinobu Kano, Neo Watanabe, et al. 2023. AIWolfDial 2023: Summary of natural language division of 5th international AIWolf contest. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 84–100, Prague, Czechia. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, et al. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Long Ouyang, Jeffrey Wu, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, et al. 2018. Improving language understanding by generative pre-training.