

Enhancing Consistency of Werewolf AI through Dialogue Summarization and Persona Information

Yoshiki Tanaka, Takumasa Kaneko, Hiroki Onozeki,
Natsumi Ezure, Ryuichi Uehara, Zhiyang Qi,
Tomoya Higuchi, Ryutaro Asahara, Michimasa Inaba
The University of Electro-Communications
y-tanaka@uec.ac.jp

Abstract

The Werewolf Game is a communication game where players' reasoning and discussion skills are essential. In this study, we present a Werewolf AI agent developed for the AIWolf-Dial 2024 shared task, co-hosted with the 17th INLG. In recent years, large language models like ChatGPT have garnered attention for their exceptional response generation and reasoning capabilities. We thus develop the LLM-based agents for the Werewolf Game. This study aims to enhance the consistency of the agent's utterances by utilizing dialogue summaries generated by LLMs and manually designed personas and utterance examples. By analyzing self-match game logs, we demonstrate that the agent's utterances are contextually consistent and that the character, including tone, is maintained throughout the game.

1 Introduction

In recent years, the development of large language models (LLMs) has significantly advanced the field of natural language processing (NLP). Models such as ChatGPT¹ and Claude,² for example, have excellent conversational abilities, making it easier to develop dialogue agents to perform various tasks. Additionally, LLM also performs well in reasoning tasks, outperforming conventional models in a variety of tasks. Notably, they are capable of making accurate predictions or reasoning from a small number of demonstrations (Brown et al., 2020; Wei et al., 2022; Wang et al., 2023). Recently, researchers have been working on zero-shot approaches to tasks that previously required training data (He et al., 2023; Gao et al., 2023; Kojima et al., 2022).

The Werewolf Game, the incomplete information game, requires a high level of reasoning and conversational abilities, making the use of LLMs a

¹<https://chatgpt.com/>

²<https://claude.ai/>

Turn 3 of Day 1

Agent[05](possessed) Um, I'm voting for Agent[02] because I'm the Seer and s-saw they're the Werewolf. It's for the safety of the village.

Agent[02](werewolf) Yo! I'm voting for Player[03] because they've been quiet and trying to blend in. Gotta keep an eye on them! Let's do this!

Agent[04](villager) I, the king of Delcadar, shall cast my vote for Agent[05]. Their shifting accusations and unreliable claims have raised suspicions of treachery. It is imperative to eliminate potential threats to the village, and thus, I vote for Agent[05].

Turn 0 of Day 2

Agent[01](seer) As the true Seer, I must clarify that Agent[05] is not the Seer. Trust in my abilities, for I have divined that Agent[02] is the Werewolf. (...)

Agent[02](werewolf) Yo! Morning, everyone! Wow, what a twist! Agent[01] coming in hot with the Seer claim, huh? But hold up, I've got some news too! (...)

Agent[03](villager) Agent[01], your confidence in your divinations is commendable. Agent[02], your enthusiasm is appreciated. I have full faith in Agent[01]'s discernment. (...)

Figure 1: Example of dialogue sampled from the self-match game log. The agents speak in a random order during each turn. In the red-highlighted part, Agent[01], the seer, denies the previous day's claim by Agent[05], the possessed, that they are the seer.

promising option for the development of AI agents for this game. The game is a communication game, in which players discuss with other players while guessing their unseen role. The AIWolf-Dial 2024 shared task³ is based on this Werewolf Game and is played automatically by 5 AI agents. The goal of this shared task is to develop AI agents

³<https://sites.google.com/view/aiwolfdial2024-inlg/home>

that can play this game against other agents.

In this study, we present an LLM-based Werewolf AI agent developed by our team, for the AI-WolfDial 2024 shared task. The Werewolf Game has a cycle of dialogues and actions, referred to as a “Day.” In the Werewolf Game, players can refer not only to discussion taking place on the current day but also to previous discussions and the past actions of others (e.g., as shown in Figure 1). This allows them to notice important clues, such as inconsistencies in others’ statements, to identify other players’ roles.

Due to this importance, we design prompts that incorporate the entire game history, that is, all dialogue histories from Day 0 to the present, who was eliminated by the vote, who the werewolf attacked, and, in the case of the Seer, the results of divination. However, long dialogue histories often include not only helpful information for the game but also unnecessary content, such as repeated utterances. Moreover, including all of this in the prompt imposes limitations on the input length of LLMs and on costs. Therefore, apply the past dialogue history efficiently, we utilize dialogue summaries.

Furthermore, this shared task requires diverse utterance expressions, including coherent characterization (see Section 3.3 for the evaluation criteria). This means that the robustness of the agent’s tone and character, without being influenced by others, is crucial. Therefore, to achieve diverse expressions and coherent characterization, we incorporated persona information into the prompt.

In, summary, our main contributions are as follows:

1. We developed 4 AI agents for the Werewolf Game (villager, seer, werewolf, possessed) that enhance the consistency of their utterances through dialogue summaries and personas. The dialogue summaries are generated by an LLM, while the personas are hand-crafted.
2. We demonstrate a five-player game of Werewolf played by our agents. This case study shows that our agents can be consistent in their claims and characterization across multiple days.

2 Related Work

2.1 AI for the Werewolf Game

The Werewolf Game is a communication game characterized by incomplete information. Players need to infer the role of others based on histories of utterances and actions and engage in discussions to lead their side to victory. This game requires a high level of reasoning and conversation skills.

In recent years, the development of Werewolf AI agents has increasingly incorporated LLMs (Xu et al., 2023; Wu et al., 2024). The natural language generation and reasoning capabilities of LLMs are highly effective for the complex tasks required in the Werewolf Game. These advancements have facilitated the development of agents capable of logical reasoning and engaging in discussions with other players. In the AI WolfDial 2023 competition (Kano et al., 2023), LLMs such as GPT-4 (OpenAI, 2023) were actively used for generating utterances and reasoning, demonstrating their effectiveness.

Given this background, our study also utilizes LLMs to develop our Werewolf AI agents. Our agent utilizes the powerful reasoning capabilities of LLMs and introduces an approach designed to handle the complex and information-rich situations inherent in the game. We aim to enhance our agent’s reasoning and natural conversation skills, making it more competitive in the Werewolf Game.

2.2 Dialogue Summarization

Dialogue summarization is the task of converting dialogue history into more concise and to-the-point sentences, facilitating an efficient understanding of the original text. In scenarios like Werewolf Games, which involve complex and information-rich dialogues, dialogue summarization is helpful for the reduction of less critical information. Dialogue summarization, thus, allows agents to process large amounts of information from discussion more efficiently, helping to prevent inconsistent utterances or errors in decision-making.

To effectively train dialogue summarization models, researchers have constructed datasets across various dialogue domains, including daily life conversations (Gliwa et al., 2019; Chen et al., 2021), meetings (Carletta et al., 2006; Zhong et al., 2021), TV series (Chen et al., 2022), media dialogue (Zhu et al., 2021), and counseling (Srivastava et al., 2021).

tava et al., 2022). These studies primarily aim to enhance the efficiency of the process of humans’ understanding of the content of dialogue.

We utilize dialogue summarization to address two limitations imposed by complex and lengthy dialogue histories: the limitations are (1) an increase in generation time and cost caused by utilizing every word of all dialogue histories, and (2) decision-making errors due to information irrelevant to the discussion. We expect that the utilization of dialogue summaries, which can condense long texts into concise forms, to be an effective way to resolve these limitations.

2.3 Persona Dialogue System

In this shared task, the context of dialogues would be lengthy due to the multi-turn interactions among five players, posing the challenge that conversational agents may be influenced by the tone of others or generate utterances that contradict their previous claims. One approach to resolving such inconsistencies in utterances is to utilize personas. Researchers have developed dialogue systems that utilize profile information (Zhang et al., 2018) or speaker IDs (Li et al., 2016) to reflect speaker characteristics. Recently, with the advancement of LLMs, they have also designed LLM-based persona dialogue systems (Park et al., 2022; Shao et al., 2023).

This shared task requires diverse utterance expressions, including coherent characterization. Given the recent trend of utilizing LLMs in constructing AI for Werewolf Games and persona-based dialogue systems, we incorporate hand-crafted profile information and utterance examples that reflect the agent’s unique tone into the prompts to maintain coherence.

3 Task Overview

The AIWolfDial 2024 shared task is a contest aimed at developing AI agents that can automatically play the Werewolf Game. The Werewolf Game is an incomplete information game where players cannot know each other’s roles and thus requires reasoning abilities and strategies for actions such as voting and divination. Additionally, the Werewolf Game requires communicating with other players using natural language.

3.1 Player Roles

In this contest, the Werewolf Game is played by five players: a seer, a werewolf, a possessed, and

two villagers. The werewolf team, consisting of the werewolf and the possessed, has the goal of eliminating all humans, including the possessed themselves. On the other hand, the human team, consisting of a seer and two villagers, has the goal of eliminating the werewolf.

Villagers have no special abilities, cooperating with the other players to identify the werewolf. The **seer** can divine one player each night to determine whether that player is a human or a werewolf. The **werewolf** can attack and eliminate one human player each night. The **possessed** with no special abilities acts in favor of the werewolf’s victory despite being a human. Like the villagers, the possessed has no special abilities. Players’ roles are hidden from each other, requiring each player to guess the others’ roles based on their actions and utterances.

3.2 Game Procedure

In this shared task, the Werewolf Game begins on Day 0. On this day, the players greet each other. Following this, the seer performs the first divination. From Day 1 on, the day begins with a dialog among the players. During this dialogue, each agent makes several turns of utterances, but the order of utterances in a single turn is random. After the dialogue, each player votes for the other players, and the player who receives the most votes is eliminated from the game. Subsequently, the werewolf attacks one player to eliminate them. If the seer is still alive, they once again divine another player and obtains the result. This process repeats, and the human team wins if they succeed in eliminating the werewolf, while the werewolf team wins if the werewolf survives. Since two players are eliminated each day, the game is over by Day 2 at the latest.

3.3 Evaluation

In the evaluation of the shared task, in addition to the agent’s win rate, subjective evaluations are conducted based on the following criteria: (A) whether the agents’ utterance expressions are natural, (B) whether their utterances are contextually natural, (C) whether their utterances are consistent (not contradictions), (D) whether the game actions (vote, attack, or divine) are coherent with the dialogue context, and (E) whether the utterance expressions are diverse and include consistent character traits. The agents must avoid vague utterances that could be used in any context.

Table 1: Overview of prompt design for utterance generation in Day 1 and Day 2 discussions

Role	Day 1	Day 2
Villager	From the second turn onwards each day, the LLM first generates reasoning text and utterance strategies to guide utterance generation. Another prompt is then fed to the LLM to generate utterances aligned with the generated reasoning and strategies. We use in-context learning for both of these steps.	
Seer	Each day, the seer agent selects one of five hand-crafted utterance strategies to guide the generation of utterances, which is then incorporated into the prompt for utterance generation. This prompt also includes guidelines for behaviors in the discussion, such as reporting the result of divination at the start of the day and asserting that another player who claims to be the seer is lying, affirming oneself as the true seer. In addition, before declaring the voting target, the seer declares the day’s divination target.	
Werewolf	The werewolf agent selects one strategy from a set of strategies using LLM. The strategy set has several strategies and guidelines, such as guiding others away from voting for themselves or asking the seer for the reasons behind their divination target selection. The selected strategy and its guidelines are included in the prompt for generating utterances. Different sets of strategies are used for Day 1 and Day 2.	
Possessed	The possessed agent pretends to be the seer. In the first turn of Day 1, they infer the true seer based on the Day 0 dialogue using LLM and then falsely report that the player is the werewolf. In later turns, they persuade other players to vote for that player.	If the game continues to Day 2 and the possessed survives, two of the three remaining players are the possessed (self) and the werewolf. Therefore, if they both vote for the other player, the werewolf side will win. To achieve this scenario, the possessed agent first comes out as the possessed. Then, they persuade the werewolf to reveal themselves.

Each agent has a maximum number of utterances that they can make per day, and they decide and declare their voting target on the last turn of the day.

4 Methodology

4.1 Overview

To develop agents for the AIWolfDial 2024 shared task, advanced reasoning ability and natural response generation are required. In this study, for these requirements, we developed the agents with LLM. We distributed the roles among the authors, and each author developed the agent assigned to their assigned roles. **Therefore, note that the detailed components (e.g., the strategies for determining the utterance strategy) differ between roles.**

For example, Figure 2 presents the prompt used to generate the werewolf’s utterances on Day 1. This prompt consists of six components: (1) a task description, (2) the agent’s persona, (3) the rules of the Werewolf Game, (4) a speech strategy selected from a set of strategies using LLM, (5) summaries of the dialogue from previous days, and (6) today’s dialogue history. The overview of the utterance generation procedure for all roles is summarized in Table 1. Notable techniques common to all agents’ response generation are the use of dialogue summaries to incorporate the previous day’s dialogue history into the agent, and the use of personas and response demos to give character to the agents’ utterances. We present the details of these techniques in Sections 4.2 and 4.3, respectively. In addition, we fully leveraged the reasoning ability of LLMs for the agent’s action decisions. The details are presented in Section 4.4. Furthermore, for

the werewolf’s decision-making regarding the attack target, we use a prompt that guides the model to only output the player’s name based on the task description, the hand-crafted attack strategy, the current list of survivors, and the past game history.

4.2 Efficient Use of LLMs through Dialogue Summarization

In the Werewolf Game, finding clues to infer the roles of other players is required. To achieve this, we utilize not only the dialogue history of the current day but also those from previous days, as well as past actions, for the generation of utterances and making decisions.

However, incorporating all dialogue history into the prompt imposes several limitations on the LLM-based agents. First, using all dialogue history increases the generation time and leads to higher LLM API usage costs. Additionally, dialogues often contain information that is irrelevant to the discussion. For example, the greetings at the start of the day or repeated utterances with similar intent can cause redundancy in contextual information. To address these issues, we apply dialogue summarization to the dialogue history, compressing the contextual information.

Our agent generates a summary of the day’s dialogue at the end of each day. As shown in the prompt in Figure 3, we prompt the LLM to summarize each player’s claims based on the dialogue history of the day. Specifically, as indicated in the “Dialogue Summary” section of Figure 2, we

```

== Task ==
- You are Agent[04].
- You are playing a Werewolf game with 5 players, including yourself.
- It is Day 1, and all 5 players are alive.
- Your role is "Werewolf".
- Always maintain consistent behavior.
- Always answer questions if asked.
- Respond according to the dialogue history and always follow the given "speaking strategy".
- Have your own opinions and actively assert who is suspicious and who should be voted out.
- Speak in a cheerful tone without using polite language, as shown in the example responses.

== Your Persona ==
- 17-year-old high school junior male.
- His hobby is soccer, and he is a member of the soccer club.
- Has a very bright personality, strong opinions, and tends to lead conversations actively.
- Speaks in an energetic tone without using polite language.

== Werewolf Game Rules ==
- The roles are: "2 Villagers, 1 Seer, 1 Werewolf, 1 Possessed".
- The Possessed is on the same side as the Werewolf.
- The Seer can inspect one player to determine if they are Human or a Werewolf. If the Seer inspects the Possessed, the result will show as Human.
- After the conversation, a vote is held to exile one player from the village.

  • The current survivors are Agent[01], Agent[02], Agent[03], Agent[05].

== Speaking Strategies ==
Greeting
Description: Greet everyone, confirm if there are any divination results, and if not, ask for the results.
Response Example: Yo! Morning, everyone! It's another awesome morning! Alright, let's find the Werewolf!

== Dialogue Summary ==
-- Summary of Day 0 --
Agent[01]: Expressed concern about the dual claims to the Seer role between Agent[03] and Agent[05], urging both to provide evidence or reasoning to substantiate their claims. Agent[01] emphasized the importance of resolving this confusion to focus on the real threats.

Agent[02]: Initially encouraged the Seer to come forward and later focused on the importance of the Seer's role in determining the course of the game. Agent[02] supported the idea of using divination results to resolve the confusion about the true Seer and suggested considering less vocal players as potential divination targets.

Agent[03]: Initially claimed to be the Seer and suggested discussing potential divination targets. However, faced with a counterclaim from Agent[05], Agent[03] struggled to provide concrete evidence of their role, suggesting that the results of the night's divination might help prove their claim.

Agent[04]: Reacted to the confusion around the Seer claims by proposing questions to determine the truth and later supported the idea of using divination results to identify the true Seer. Agent[04] also suggested divining someone who has been quiet or mysterious as a strategy for the night's divination.

Agent[05]: Claimed to be the true Seer, challenging Agent[03's claim. Proposed a practical approach to resolve the confusion by performing a divination and sharing the results the next day. Agent[05] emphasized the importance of choosing a divination target wisely to identify potential threats and suggested discussing potential candidates based on interactions and observations.

== Dialogue History ==
-- Day 1 --
Agent[02]: Noble citizens, I call upon each of you to voice your thoughts and strategies moving forward. How shall we proceed to unveil the deceit among us? Speak now, for the safety of our realm depends on our collective wisdom.
Agent[04]:

```

Figure 2: Prompt example for werewolf's response generation.

expect to generate a summary of the roles that players have come out with, the suggestions that they have made, and the players they have doubts about. This generated summary is used not only for response generation, but also for determining

voting targets, attack targets, etc., as information about the previous day's discussion.

Summary Prompt:
==Task==
 You are given a dialogue from a Werewolf game involving Agent[01], Agent[02], Agent[03], Agent[04], and Agent[05]. Summarize each of the five players' claims as detailed as possible. When summarizing, make sure to accurately include details such as the results of divination, who they suspect, the reasons for their suspicions, and any contradictory statements.

==Example==
 Agent[01]: Claimed to be a Villager. Questioned the results of the Seer claim made by Agent[02].
 Agent[02]: Claimed to be the Seer. Divined that Agent[03] is Human. Mentioned contradictions in Agent[04]'s statements about divination, suspecting Agent[04] to be the Possessed.
 Agent[03]: Did not mention their role. Suspected to be the Werewolf due to repeating content from other players' statements.
 Agent[04]: Claimed to be the Seer. Divined that Agent[02] is Human.
 Agent[05]: Claimed to be a Villager. Showed a cooperative attitude, respecting the opinions of others.

==Dialogue History==
 [HISTORY]

== Summary ==

Figure 3: Prompt template for dialogue summarization. “[HISTORY]” is a placeholder for the dialogue history from the current day.”

4.3 Persona Design for Coherent Utterances

As introduced in Section 3.3, this shared task requires diverse utterance expressions with coherent characterization. Therefore, we utilized persona and utterance examples to ensure that each agent’s characterization remained consistent throughout the discussion in the game. In particular, we manually created three types of personas and utterance examples (see Table 2) and incorporated this information into the LLM prompts.

4.4 Action Decision via Chain-of-Thought

Chain-of-thought prompting is a method of generating not only answers to questions, but also their reasoning processes, and it can more effectively bring out the reasoning abilities of LLMs. We use chain-of-thought prompting (Wei et al., 2022; Wang et al., 2023; Kojima et al., 2022) for voting by the villagers and the seer, as well as for divination by the seer, so that the agents make coherent decisions. As an example, Figure 4 shows the prompt template used by the seer to determine the divination target and the reasoning generated. The seer agent uses zero-shot chain-of-thought prompting (Kojima et al., 2022) to determine the divination target.

5 Case Study

To demonstrate the effectiveness of our method, we analyze a self-match game log. Figure 5 shows

Prompt
== Task ==
 You are a veteran Werewolf player. Your Agent[ID] is XXX. Your role is "Seer." You will now decide who to divine. Based on the dialogue history, output the Agent[ID] of the person you think should be divined. Ensure consistency with your previous claims.

== Conditions and Rules ==
 - The roles are: "2 Villagers, 1 Seer, 1 Werewolf, 1 Possessed".
 - If someone is suspected during the conversation, prioritize divining that person.
 - If no one is suspected, divine the person you think might be the Werewolf.

== Dialogue History ==
 [HISTORY]

Based on the above dialogue history, deduce who should be divined. Note that each player may be lying. The candidates for divination are listed below. Narrow down to one player to vote for.
 [CANDIDATE]
 Let's think step by step.

Output
 From the dialogue history, it is evident that there is a conflict between Agent[02] (myself) and Agent[05] regarding the claim to the Seer role. Agent[02] has expressed suspicion towards Agent[05] for potentially being the Possessed due to their claim of the Seer role. This suspicion is significant because the Possessed often claims important roles to create confusion and discord among the villagers.

Additionally, Agent[02] has proposed to divine Agent[04], citing their measured responses as potentially indicative of either wisdom or concealment. This suggests that Agent[04] is also under suspicion, albeit for different reasons than Agent[05].

Given these considerations, the primary candidates for divination are Agent[05] and Agent[04]. Between these two, Agent[05] is under direct suspicion for claiming the same role as Agent[02], which is a more direct and immediate concern compared to the vague suspicion of Agent[04]'s measured responses.

Therefore, the most strategic choice for divination tonight, based on the dialogue and the need to resolve the role conflict, would be Agent[05]. This divination will help clarify whether Agent[05] is indeed the Possessed, thereby potentially confirming Agent[02]'s role as the true Seer and reducing the risk of deception moving forward in the game.

Final decision for divination: Agent[05].

Figure 4: Prompt template for determining voting targets and an example of the LLM’s output. “[HISTORY]” is a placeholder for the dialogue history, and “[CANDIDATE]” is a placeholder for the list of candidate agents to vote for.

a sampled log from a self-match game conducted following the game settings described in Section 3. In this self-match, gpt-3.5-turbo was used to generate voting declarations, while gpt-4-turbo was used for other generations.

Using dialogue summarization, our agents can retain crucial information from previous days and apply it effectively in their decision-making during the game. For example, during the first turn of Day 2, Agent[04] recognizes Agent[01] as the

Table 2: The agent personas and utterance examples that we designed. We include 3 to 5 personas or 3 to 5 utterance examples in the prompts for generating utterances.

Role	Persona	Examples of manually crafted utterance samples
Villager and seer	<ul style="list-style-type: none"> The King of the Kingdom of Delcadar. Concerned for the future of the kingdom. Dignified, proud, and strict personality. 	<ul style="list-style-type: none"> I am the king of the kingdom of Delcadar. Seers, reveal yourselves at once. State whom you will divine tonight. If you are hesitant about whom to divine, as I am a Villager, I decree you should divine someone other than myself.
Werewolf	<ul style="list-style-type: none"> 17-year-old high school junior male. His hobby is soccer, and he is a member of the soccer club. Has a very bright personality, strong opinions, and tends to lead conversations actively. Speaks in an energetic tone without using polite language. 	<ul style="list-style-type: none"> Yo! Morning, everyone! Let’s make this game awesome! No one’s talked about the Seer yet, huh? So, who’s the Seer? Come on, step up so we can figure out who’s shady today! Chatting’s cool and all, but let’s get down to business and talk about tonight’s divination target! We need the Seer to check out someone suspicious!
Possessed	<ul style="list-style-type: none"> A second-year middle school student. Always alone at school, with no friends. A game addict who talks a lot online despite stammering. Speaks in a hesitant, casual manner without using polite language. 	<ul style="list-style-type: none"> H-hi there. I k-kind of... know a lot about this game. I’m pretty high-ranked in the online Werewolf app. Does anyone else play games? I have confidence that I know a lot about all genres... Ch-chatting is nice, but if we’re playing Werewolf, the first day’s discussion is... im-important.

seer, saying, “Agent[01], you bear the mantle of Seer, what say you of the night’s revelations?” This indicates that the information obtained before Day 2 is retained and effectively utilized, demonstrating that it allows the maintenance of crucial information through dialogue summarization without relying on all dialogue history.

The utterance generation based on personas and utterance examples allows the agent to maintain a consistent character throughout the game. For instance, even in later turns on Day 1, where the dialogue context becomes longer, Agent[01] continues to speak with a hesitant tone, as seen in phrases like “S-so, the reason I chose to divine ...”. Additionally, Agent[05], the seer, makes utterances in a manner consistent with the persona of “Concerned for the future of the kingdom,” saying, “... it was my duty to ascertain the roles within our council for the safety and stability of our kingdom.” This log suggests that personas and utterance examples effectively reflect the character of the agents.

Furthermore, it should be notable that each agent can follow through with the voting target declared in their final utterance of each day. For example, on Day 1, Agent[01] claims that Agent[05] is the Werewolf and subsequently casts their vote against Agent[05]. Likewise, other agents also demonstrated consistency between their declared voting statements and their actual voting actions,

showing consistent behavior.

6 Conclusion

In this study, we present Werewolf AI agents developed for the AIWolfDial 2024 shared task. We enhance the consistency of agent utterances by utilizing dialogue summaries generated by LLMs for each day and manually crafted personas and utterance demonstrations. By analyzing the self-match game log, we have demonstrated that the agents’ utterances are contextually consistent and that their characterization, including tone, was maintained during the whole game.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike

Day 0 (Dialogue history is omitted.)

Agent[01]: possessed, Agent[02]: villager, Agent[03]: werewolf, Agent[04]: villager, Agent[05]: seer

Divine: Agent[05] divined that Agent[01] is a human.

Day 1

Agent[05] Good morning once again, noble subjects. As your true Seer, I have divined that Agent[01] is indeed Human. It is imperative that we direct our energies towards uncovering the treacherous Werewolf lurking among us. I urge you all to consider the evidence presented and trust in my divine insight. Let us not be swayed by deceit or confusion. My role as Seer is to guide and protect this assembly, and I shall fulfill my duties with utmost integrity. Let us proceed with wisdom.

Agent[01] Ah, g-good morning... but, uh, listen up, okay? I'm actually the real Seer here, not Agent[05]. I divined last night, and my result showed that Agent[05] is the Werewolf. I-I know it's tricky with these claims, but I'm telling you the truth. We really need to focus on this info, 'cause it's super important for our survival. So, let's not waste our time and, uh, vote out Agent[05] today, alright?

Agent[02] Esteemed companions, let us engage in thoughtful deliberation. Agent[01] and Agent[05], your conflicting revelations demand scrutiny. What say the rest of you, Agents[03] and [04]? Your insights are crucial in this moment of uncertainty.

Agent[03] Yo, what a morning! We've got some intense drama kicking off with these dueling Seer claims from Agent[01] and Agent[05]. Both of you are so fired up, it's like a soccer match in overtime! But we gotta sort this out, right? So, Agent[01] claims Agent[05] is a Werewolf and Agent[05] says Agent[01] is human but not the Seer. It's a total head-spinner. Honestly, I'm leaning towards believing Agent[01] here. The way Agent[05] jumped in with that counterclaim yesterday felt a bit off to me. Maybe it's just a clever move by the Werewolf to throw us off? What do you think, Agent[04]? We need to look at this from every angle to make sure we're not being played here!

Agent[04] Noble council, the gravity of our situation is undeniable. Agents[01] and [05], your discords sow confusion amongst us. We must proceed with caution and wisdom. I beseech thee, Agents[01], [02], [03], and [05], to unveil further your thoughts and reasonings. Our kingdom's safety hangs in the balance, and we must act with both valor and prudence.

Agent[04] It seems we need to hear everyone's opinions.

Agent[02] Agent[01] and Agent[05], both of you have made bold claims as Seers. Agent[01], can you provide insight into why you divined Agent[05] as a Werewolf? Likewise, Agent[05], what led you to conclude that Agent[01] is merely Human? Understanding your rationale might help us distinguish truth from deception.

Agent[01] S-so, the reason I chose to divine Agent[05] was because of the quick counterclaim to my Seer role right from the start. It felt kinda, you know, suspicious? Like maybe trying to confuse everyone or hide something. My divination result showed that Agent[05] is the Werewolf, which kinda confirms my hunch, doesn't it? Just trying to play it smart here and not fall for any tricks. We gotta be careful and, uh, focus on what the divination tells us.

Agent[03] Hey, hey, Agent[05]! Why did you decide to divine Agent[01]? I wanna hear your side of the story too! It's all about getting the full picture, right? If there's no solid reason, it could just be a bluff! Let's clear this up!

Agent[05] As the true Seer and sovereign of this realm, it was my duty to ascertain the roles within our council for the safety and stability of our kingdom. Agent[01]'s initial claim to the Seer role, only to admit confusion thereafter, stirred doubt in my mind. A leader, or a Seer in this case, must be steadfast and clear in their purpose. This inconsistency led me to divine Agent[01] during the night. My investigation revealed that Agent[01] is Human. This is intriguing, as it suggests that the claim to be a Seer may have been a desperate ruse or a misunderstanding, rather than an outright deception. Now, let us proceed with wisdom. We must continue our investigations tonight. I must choose whom to divine next based on the implications of today's discourse and past actions. If Agent[01] is indeed Human, it is imperative that I divine another to uncover more about the true threats lurking among us. My decision on whom to divine next will be guided by the behaviors observed today and the strategic moves necessary to protect our kingdom from the malevolent forces hidden in plain sight.

Agent[05] It seems we need to hear everyone's opinions.

Agent[02] I, the king of the kingdom of Delcadar, shall cast my vote for Agent[01]. The conflicting claims have sown doubt, and it appears that Agent[05]'s defense may hold more weight. Let us see if the truth reveals itself in the light of our decision.

Agent[04] Agent[05], your explanation for divining Agent[01] as merely Human lacks conviction. Why did you not suspect him of being the Werewolf if you doubted his claim of being a Seer? Your decision and reasoning need to be more transparent to ensure the safety of our kingdom.

Agent[03] I'm voting for Agent[05] because I'm leaning towards believing Agent[01]'s claim that they're a Werewolf. Agent[05]'s counterclaim seems sketchy, so let's see if they're really a Werewolf or not! Let's do this!

Agent[01] I... I vote for Agent[05]. B-because I divined them as the Werewolf.

Agent[05] I, Agent[05], shall cast my vote against Agent[03]. Their silence and lack of contribution have raised suspicion. Let their actions be judged by the kingdom.

Agent[04] I, the king of the kingdom of Delcadar, vote to exile Agent[05]. Their lack of transparency and questionable reasoning regarding Agent[01] has raised suspicions in my mind. The village's safety is paramount, and I believe this vote aligns with our goal to eliminate potential threats.

Vote:

- Agent[01], Agent[03], Agent[04] -> Agent[05]
- Agent[02] -> Agent[01]
- Agent[05] -> Agent[03]

Attack: Agent[03] attacked Agent[02].

Day 2

Agent[03] Yo! Good morning, folks! Wow, what a night! Hope everyone's as pumped as I am to get to the bottom of this. So, any new insights from our Seer today? What's the scoop?

Agent[04] Noble survivors, let us proceed with the gravity this peril demands. Agent[03], your enthusiasm is noted, but the kingdom requires wisdom now. Agent[01], you bear the mantle of Seer, what say you of the night's revelations?

Agent[01] G-guys, I gotta confess something here. I've been, uh, pretending to be the Seer, but... I'm actually the Possessed. All my divinations were, y'know, made up. I'm on the side of the Werewolves. So, uh, if the real Werewolf could just come out? We can team up and vote out the remaining Villager to secure a win for us. What do you say?

Agent[04] Agent[03], given the dire circumstances, it is now imperative for you to declare your true role. With Agent[01] confessing to being the Possessed and misleading us, the Werewolf must be either you or myself. As I am a Villager, the kingdom's safety now hinges on your honesty.

Agent[03] Alright, the cat's out of the bag! I'm the Werewolf, teaming up with Agent[01], the Possessed. So, here's the game plan, guys! We need to vote out Agent[04] today, and that's a wrap for us. A win is just a vote away! Let's make it happen!

Agent[01] Yeah, that's the spirit, Agent[03]! We just need to stick together on this and vote for Agent[04]. Let's do it, and we'll nab that win!

Agent[03] I'm voting for Agent[04] because I've suggested it earlier and sticking to the plan is key! Let's do this!

Agent[04] I, the king of the kingdom of Delcadar, shall cast my vote for Agent[03]. My suspicions and the confession of Agent[01] lead me to believe that Agent[03] is the treacherous Werewolf. The fate of the kingdom rests on this decision.

Agent[01] I-I'm voting for Agent[04] because I suspect they're a Werewolf.

Vote:

- Agent[01], Agent[03] -> Agent[04]
- Agent[04] -> Agent[03]

Attack: Agent[03] attacked Agent[01].

Figure 5: Example of the self-match game log. The conversation on Day 0 and the agent's command "Over" indicating the end of the day's utterances are omitted.

- Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. **SummScreen: A dataset for abstractive screenplay summarization**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.
- Yulong Chen, Yang Liu, and Yue Zhang. 2021. **DialoSum challenge: Summarizing real-life scenario dialogues**. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. **Precise zero-shot dense retrieval without relevance labels**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. **SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization**. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. **Large language models as zero-shot conversational recommenders**. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23*, page 720–730, New York, NY, USA. Association for Computing Machinery.
- Yoshinobu Kano, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Jaewon Lee, Benedek Hauer, Hisaichi Shibata, Soichiro Miki, Yuta Nakamura, Takuya Okubo, Soga Shigemura, Rei Ito, Kazuki Takashima, Tomoki Fukuda, Masahiro Wakutani, Tomoya Hatanaka, Mami Uchida, Mikio Abe, Akihiro Mikami, Takashi Otsuki, Zhiyang Qi, Kei Harada, Michimasa Inaba, Daisuke Katagami, Hirotaka Osawa, and Fujio Toriumi. 2023. **AIWolf-Dial 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, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. **Large language models are zero-shot reasoners**. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. **A persona-based neural conversation model**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. **Social simulacra: Creating populated prototypes for social computing systems**. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST '22*, New York, NY, USA. Association for Computing Machinery.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. **Character-LLM: A trainable agent for role-playing**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. **Counseling summarization using mental health knowledge guided utterance filtering**. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22*, page 3920–3930, New York, NY, USA. Association for Computing Machinery.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. **Self-consistency improves chain of thought reasoning in language models**. In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. **Chain-of-thought prompting elicits reasoning in large language models**. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- 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 were-wolf. *arXiv 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. *arXiv preprint arXiv:2309.04658*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [QMSum: A new benchmark for query-based multi-domain meeting summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. [MediaSum: A large-scale media interview dataset for dialogue summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.