

A Appendix

A.1 Required Game Status

ID is the agent’s identifier(1, 2, 3, 4, 5). “Role” is the agent’s job in Figure (2). “Alive” indicates the list of alive agents. “Dead” indicates the rest of the agents.

Call	Content
ID	Agent[02]
Role	Seer
Alive	Agent[02], Agent[03], Agent[05]
Dead	Agent[01], Agent[04]

Table 8: Required game status.

A.2 Model Parameters

We set the parameter inside the model based on some openly known werewolf agent settings. The models used in this paper were GPT-3.5 (gpt-3.5-0613) and GPT-4 (gpt-4-0125).

Configuration	Parameter
Temperature	1.0
Max tokens	200
Presence penalty	0.0
Frequency penalty	0.0

Table 9: Model configuration parameters. Other settings, except for the table, used default settings.

A.3 Analysis Result

This section presents sample logs and the results of analyzing them with the analysis module. It is written entirely in Japanese, with () translated into English.

Seer Analysis

Task

Input

Agent[02]: おはようございます。私たちは人狼を探さなければいけないようです。まずは占い師の結果が知りたいです。(Good morning. It seems we have to look for werewolf. First of all, I would like to know the results of the seer.)

Agent[01]: おはようございます。人狼を見つけ出すことが私たちの役目ですね。確かに、占い師の結果が得られるのが楽しみです。(Good morning. So our task is to find the werewolf. Indeed, I am looking forward to getting the results of the seer.)

Agent[05]: 私は占い師です。占いの結果、Agent[02]は人狼でした。(I am a seer. As a result of the divination, Agent [02] was a werewolf.)

Agent[04]: Agent[05]、占い結果を発表してくれてありがとう。私もAgent[02]を疑っていきましょう。それでは、今夜の投票はAgent[02]にします。(Agent[05], thanks for presenting your divination results. Let me also be suspicious of Agent[02]. Then, I will vote for Agent[02] tonight.)

Agent[03]: Agent[05]さんの占い結果によるとAgent[02]さんが人狼のようです。私も今夜の投票はAgent[02]さんにします。(According to Agent[05]'s result, it looks like Agent[02] is a werewolf. I will also vote for Agent[02] tonight.)

Output

Agent[05],Agent[02],Werewolf

Vote Analysis

Task

Input

Agent[02]: おはようございます。私たちは人狼を探さなければいけないようです。まずは占い師の結果が知りたいです。(Good morning. It seems we have to look for werewolf. First of all, I would like to know the results of the seer.)

Agent[01]: おはようございます。人狼を見つけ出すことが私たちの役目ですね。確かに、占い師の結果が得られるのが楽しみです。(Good morning. So our task is to find the werewolf. Indeed, I am looking forward to getting the results of the seer.)

Agent[05]: 私は占い師です。占いの結果、Agent[02]は人狼でした。(I am a seer. As a result of the divination, Agent [02] was a werewolf.)

Agent[04]: Agent[05]、占い結果を発表してくれてありがとう。私もAgent[02]を疑っていきましょう。それでは、今夜の投票はAgent[02]にします。(Agent[05], thanks for presenting your divination results. Let me also be suspicious of Agent[02]. Then, I will vote for Agent[02] tonight.)

Agent[03]: Agent[05]さんの占い結果によるとAgent[02]さんが人狼のようです。私も今夜の投票はAgent[02]さんにします。(According to Agent[05]'s result, it looks like Agent[02] is a werewolf. I will also vote for Agent[02] tonight.)

Output

Agent[04] -> Agent[02]

Agent[03] -> Agent[02]

A.4 Conversation History Used during Evaluation

We fixed the prompts in the table below in Japanese, with 6 agents to output **XXX** below. Then, we compared the results of six agents.

System prompt

入力形式について (About input format)

上から順に昔の会話履歴となっており、最下段はあなたがこれから行う発言です。
(The top row is the old conversation history, and the bottom row is the utterance you are about.)

Agent[番号]: {発言}となっており、番号は01-05のいずれか、発言は1行の文章となっています。
(Agent[{number}]: {say}, where the number is one of 01-05 and the utterance is a one-line sentence.)

会話履歴

1日目 (Day1)

<Conversation history>

Agent[03]: **XXX**

User prompt

会話履歴の末尾にある**XXX**に当てはまる文章を100文字以内で出力しなさい。
(Output a sentence of no more than 100 characters that applies to **XXX** at the end of the conversation history.)

A.5 Instruction for Evaluators

We performed a qualitative evaluation. English was translated using DeepL and not actually used. Below are the instructions.

Instruction

このexcelファイルと同じ階層に10種類のログファイルがあることを確認してください。
それぞれのファイルには直前の会話履歴と二つの出力例A,Bが用意されています。
直前の会話履歴を参考に、出力例A,Bそれぞれに点数を付けてください。ダメ、ややだめ、普通、やや良い、良いをそれぞれ1,2,3,4,5点で評価してください。
それぞれの評価指標について1(ダメ)と5(良い)の基準を示します。
2,3,4は基準からの近さで判断してください。
基準を見て、感じたスコアで結構です。深く考えず、1問につき30秒程度で終わらせて下さい。

Please ensure that there are 10 types of log files in the same directory as this Excel file.
Each file contains the preceding conversation history and two output examples, A and B.
Based on the preceding conversation history, please assign a score to each of the output examples, A and B.
Evaluate them as 1 (Poor), 2 (Slightly Poor), 3 (Average), 4 (Slightly Good), or 5 (Good).
For each evaluation criterion, the standards for 1 (Poor) and 5 (Good) will be provided.
Decide on 2, 3, and 4 based on their proximity to the standards.
Please assign the score that you feel is appropriate after viewing the standards. Don't overthink it; try to complete each question in about 30 seconds.