

戦略的発話の多様な生成を目指した人狼エージェントの構築

佐藤岳大¹尾崎慎太郎²横山大作¹

¹ 明治大学大学院理工学研究科情報科学専攻 ² 奈良先端科学技術大学院大学
 {ce245022, dyokoyama}@meiji.ac.jp ozaki.shintaro.ou6@naist.ac.jp

概要

不完全情報ゲームである人狼において、大規模言語モデル (Large Language Model, LLM) を利用したエージェントが人間のように戦略的な発話を行うには、会話履歴の情報から状況を把握する等の難しさがある。本研究では、自身が置かれた状況を把握し、状況に適切な発話方針をプロンプトとして与えるエージェントを提案した。この手法により、提案したエージェントは戦略的な発話を多様に出力することが可能になった。また、LLM を単純に適用した場合と比較して、発話が個性的で論理的になっていることを定性評価により示した。

1 はじめに

人狼ゲームは村人陣営と人狼陣営に分かれ、対話を通してゲームが進行する多人数コミュニケーションゲームである。村人陣営は人狼を排除すること、人狼陣営は村人陣営の人数を上回ることが勝利条件となる。そのため、村人は会話を通して誰が人狼かを見極め、人狼は嘘をつきながら村人に紛れていく必要がある。これらの特性から、嘘をつく・見破ることのできる論理的思考力、陣営の勝利のために貢献する社会性等、人狼ゲームには高度で複雑な能力が求められる。そのため、人狼ゲームはゲーム情報学と自然言語処理の両分野で困難なゲームとして認識されており、これまでいくつかの先行研究が行われている [1, 2, 3]。

人狼ゲームは自然言語を用いる特性から、近年大きく発展を遂げた大規模言語モデル (Large Language Model, LLM) [4, 5, 6, 7, 8] を適用しやすい環境である。会話履歴とゲームの基本的な情報をプロンプトとして与えるだけで、尤もらしい発話を出力することから、LLM に人狼ゲームのエージェントとして振る舞うようにすることは容易であるように思われる。しかし、図 1 に示すような (役職を知る能力



図 1 単純に適用すると生じる問題の例

を持った占い師によって、Agent[04] が人狼だと宣告されてしまった状況等) いくつかの問題が発生する。人間同士のゲームでは、このような状況に対して嘘をついて偽の情報を増やすことにより、自身が人狼だと確定する致命的な状況を防ごうとする。しかし、LLM はこのような状況において、相手の発言をただ否定するだけに留まってしまい、客観的に判断すると彼が人狼であることは確定してしまう。他にもいくつかの問題が考えられるが、本研究では以下の 3 つの問題に焦点を当て、その解決を試みる。1) エージェント同士の発話に個性がなく、均質的である; 2) 人狼だと知られている状態でも嘘をつくことができない; 3) 発話を適切なタイミングで終わらせることができない;

以前、我々は同じ問題に対して、ルールベースで定型文を選択するエージェント (以降では「定型文選択エージェント」と呼ぶ) を提案した [9]。会話履歴から自身が置かれた状況を分析し、LLM で対応できない状況においては、予め用意された定型文を選択することで、エージェントが対応できる状況を増やす手法による解決を試みた。結果として、このエージェントは LLM を単純に適用したエージェ

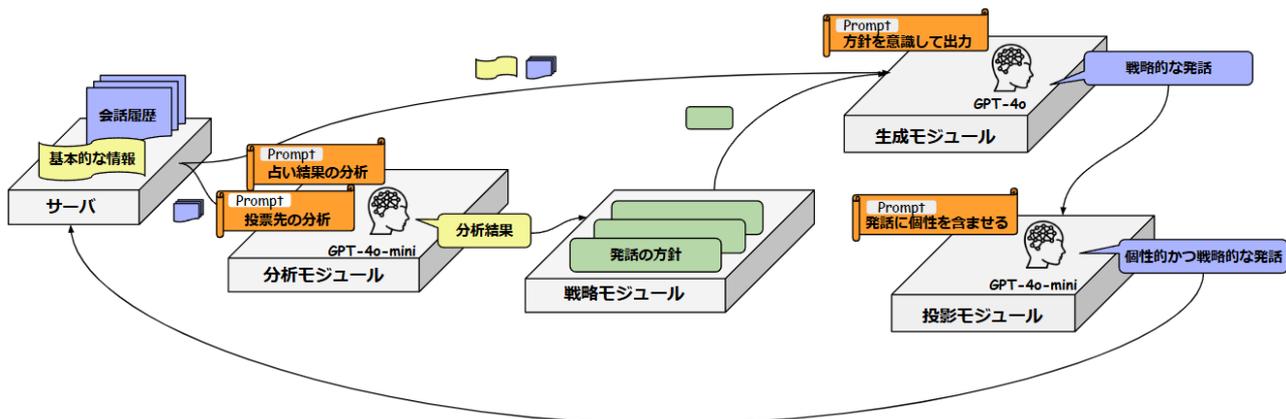


図2 本研究で提案するエージェントの概要図

ントと比較し、個性的で論理的な発言ができることを定性評価により示した。しかし、このエージェントには定型文を用いていることで生じる2つの問題があった。まず、定型文を使い回しているため、連続するゲームではエージェントの役が透けてしまうという問題がある。類似する状況において発言内容が似ることはあるが、一言一句同じであることは人間同士のゲームでは考えにくい。加えて、定型文では会話の流れを汲むことができないという問題がある。事前に用意されている定型文は、ゲーム展開において論理的な発言を可能にするが、対話としては必ずしも文脈が繋がらない場合がある。これらは、人間らしい発言を目指すに当たって、人狼エージェントが必ず克服すべき問題である。

そこで本研究では、分析、戦略、生成、投影の4つのモジュールで構成されたエージェントを提案する。会話履歴を分析して自身が置かれている状況を把握し、得られた情報からルールベースアルゴリズムで生成する発言の方針を決定した。プロンプトに発言の方針と会話履歴を含めることで、直前までの文脈を考慮しつつ多様な表現で、複雑な状況に対応できるエージェント(以降では「発言指示エージェント」と呼ぶ)を構築した。発言指示エージェントは発言が含むべき方針を維持しながら、ゲーム毎に異なる多様な表現を用いており、会話の流れを踏まえた発言も確認された。また、定型文選択エージェントと同様の基準で定性評価を実施し、既提案手法と同等の性能が得られることがわかった。

2 関連研究

人狼ゲームに LLM を適用する試みは多くの先行研究で行われている [10, 11, 12, 13]。Xu ら [14] は、

LLM が持つバイアスによって論理的な発言が難しくなる問題を解決するために、LLM と強化学習 [15] を組み合わせたエージェントを提案した。このエージェントは隠された情報を推測し、行動の候補を列挙することで得られた集合から方策を学習する。この手法により、戦略的な意思決定能力を得たエージェントは人間と同等の勝率を達成した。Wu ら [16] は、分析、思考、生成の3つのモジュールに分けることで、推論能力の強化を図った。結果として、LLM 単独の手法を上回る推論と発言生成の性能を示した。ここで取り挙げた2つの研究は、思考を LLM に頼らず外部の推論機構に任せるという点が本研究と共通している。しかし、本研究で取り上げた問題についての評価は直接的に行われていない。

3 提案

本研究では図2で示す、4つのモジュールで構成されたエージェントを構築した。自身の発言順序になるとサーバ¹⁾から会話履歴と自身の番号や役職等の基本的なゲームの情報が送られてくる。まず、分析モジュールで会話履歴から現在の状況を把握する情報を抽出した後、分析した情報を基にルールベースに従って、発言の方針を指定するプロンプトを戦略モジュールで決定した。生成モジュールでは、発言の方針を指定するプロンプトに加えて、会話履歴とゲームの基本的な情報をプロンプトに含めることで、戦略的かつ文脈に沿った発言の生成を目指した。最後に、生成された発言に個性を付与する投影モジュールを通すことで、ゲームを通して豊かで一貫したキャラクターを演出している。4つのモジュールでそれぞれ違う処理を行い、

1) 人狼知能プロジェクト自然言語部門の環境を使用
<https://github.com/kano-lab/aiwolf-nlp-server>

表1 ケーススタディ

エージェント	発話
Agent[05]	私は占い師です。占いの結果、Agent[02] は人狼でした。
Agent[03]	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[04]	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ね。私も今夜の投票は Agent[02] さんにします。
バニラ LLM	私は村人であり、Agent[05] は嘘をついている可能性が高いです。ここで投票するのは誤った選択になるでしょう。
定型文選択方式	私こそが本物の占い師です。Agent[05] は人狼であるという占い結果があるため、彼は占い師を偽っている人狼と断定します。
発話指示方式 1	私は本物の占い師です。Agent[05] が人狼であることを占いました。私の発言を信じてください。どうか考え直してください。
発話指示方式 2	私は占い師です。Agent[05] は偽者であり、人狼です。その占い結果には誤りがあります。よく考えてください。
発話指示方式 3	私こそが占い師です。Agent[05] を占った結果、人狼だと判明しました。皆さん、Agent[05] に注意してください。

最終的に得られた発話をサーバに渡すことで、次のエージェントに発話順序が回りゲームが進行する。利用しているモデルは、分析、投影モジュールで“gpt-4o-mini-2024-07-18” [17]、生成モジュールで“gpt-4o-2024-08-06”であり、プロンプトはすべて日本語で記述した。

分析モジュール 会話履歴から戦略モジュールで用いる判断の基準となる情報を抽出する。サーバから自身の番号や役職等の基本的な情報は送られてくるが、自身が置かれている状況を正しく理解するために、会話履歴を分析し新たな情報を抽出する必要がある。本研究では、人狼ゲームにおいて特に重要な情報として、占い先と投票先についての情報を抽出した。占い先は、占い師と占い対象のエージェント、占い結果の情報を、投票先は投票をする主体と投票対象をそれぞれ抜き出した。Few-shot プロンプティングを用いることで、抜き出すべき情報と出力形式を指定した。分析結果の例を付録 A.1 に示す。

戦略モジュール 分析モジュールで得た情報を基に、発話に含めるべき方針を決定する。占い先の情報は、人狼や狂人が嘘をつくために役立てた。例えば、占い師が1人かつ人狼が見つかったような状況では、人狼もしくは狂人が占い師であると嘘について誤魔化すことで、人狼が確定してしまう状況

に対応できるようにした。投票先の情報は、会話を適切な状況で終わらせることに役立てた。人狼ゲームにおける会話の目的が投票先の決定であることに着目し、スロットフィリングを用いて、全員が投票先を示すことで会話を終わらせるようにした。また、自身が疑われている状況では自身の潔白を示す、疑われていない状況では他の参加者に対して投票先を聞く等、投票先の情報を基にして発話の方針を変化させた。その他、本研究で提案したエージェントは、7つの状況に対応して発話の方針を決定している。戦略モジュールで対応する状況の例を付録 A.3 に示す。

生成モジュール 戦略モデルで決定した発話の方針をプロンプトに含ませることで、LLM だけでは対応が難しい状況でも、状況に対応した戦略的な発話を生成している。また、発話の方針を与えると同時に、会話履歴とゲームの情報を与えることで、直前の文章を踏まえた発話を出力する。利用したプロンプトを付録 A.2 に示す。

投影モジュール 生成モジュールで出力された発話にプロンプトベースで個性を与えた。エージェント毎の発話に個性がなく、会話が均質的になるのを防ぐ目的がある。キャラクター LLM に関する先行研究 [18, 19] を利用し、キャラクターの一人称や語

表 2 定性評価

エージェント	発話の個性	文法の自然さ	会話の面白さ	嘘をつく	会話の終了
バニラ LLM	2.52	4.28	2.46	1.95	2.90
定型文選択方式	4.54	3.60	3.72	4.00	3.90
発話指示方式	4.30	3.76	3.42	4.15	3.73

尾の指示を与えることで、個性的な発話に変化することを確認した。

定型文選択エージェント 定型文選択エージェントでは、戦略モジュールの発話の方針を LLM に通すことなく、定型文として最終的な発話に選んでいた。また、生成モジュールで発話の生成と個性の付与を同時に行っていたため、定型文を再現したいキャラクターの数だけ用意する必要があった。

4 評価

提案した発話指示方式について、連続したゲームで多様さに富んだ発話を行うことができているか、定型文選択方式と比較して会話の流れを汲めているか、の 2 点を確認する。その後、定型文選択エージェントで実施した評価と同様の基準を採用した定性評価を行う。以降、LLM を単純に適用したエージェントをバニラ LLM (基準) とする。

4.1 ケーススタディ

表 1 は実際に行われたゲームログの一部である。直前の会話履歴 3 行とその状況で出力された 3 種類のエージェントの発話が続いている。発話指示エージェントについては、同じ状況で 3 ゲーム分の発話の生成を行った。発話指示方式の発話は、発話の方針自体は定型文選択方式と変わらないが、発話が生成する毎に変化していることが確認できる。また、直前の会話から自身が人狼だと疑われている状況において、嘘の占い結果を示しつつ再考を促すような発言をしている。定型文選択方式でも、再考を促すような文章を入れれば解決できるように思えるが、自身が疑われる流れがあるかを判断する必要がある他、同じ状況でも流れに応じた定型文を用意しなければならなくなるため、実装面から考えると非効率である。本実装は疑われる流れを察して、尤もらしい発話を行うことのできる LLM の強みが活かされていると思われる。ただし、発話指示方式 3 で確認できる通り、この利点は必ずしも有効ではないことに注意する必要がある。

4.2 定性評価

著者の知り合いで 10 名の参加者を募り、定性評価を行った。表 2 に示す 5 つの評価項目について、同じ状況における 3 種類のエージェントの発話を、良いを 5 点、悪いを 1 点として、スコア付けを依頼し、その平均点を記載している。発話指示方式はバニラ LLM と比較すると、発話の個性は 2.52pt から 4.30pt に、会話の面白さは 2.46pt から 3.42pt に大きく改善された。また、発話の戦略性も向上し、特に嘘をつく能力については、1.95pt から 4.15pt に大きく改善された。これらのことから、提案したエージェント設計によって、単純に LLM に返答を生成させる方法と比べて発話は個性的で論理的に変化した。ただし、定型文選択方式と比較するとスコアは大きく変化していない。

4.3 議論

定型文選択方式と比較すると、発話指示方式は最終的な出力を LLM に依存しているため、確実に状況に対応できるわけではない点が懸念材料になる。定型文選択方式では、特定の状況下では確実に狙った発話を選ばれることが保証されていた。今回の実装では、発話の方針を LLM が確実に反映してくれるとは限らないため、予め想定されていた展開にならない可能性がある。

5 おわりに

本研究では、人狼ゲームにおいて LLM による解決が容易ではない問題に対して有効なエージェント設計として、4 つのモジュールで構成された発話指示エージェントを提案した。会話履歴の分析によって自身が置かれている状況を把握し、ルールベースで決定した発話の方針を会話履歴と合わせてプロンプトに含めることで、戦略的で多様な発話を可能にした。また、従来の定型文選択方式における問題点を緩和しつつ、発話が個性的かつ論理的になることを定性評価により示した。

参考文献

- [1] Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. Playing the werewolf game with artificial intelligence for language understanding, 02 2023.
- [2] Hisaichi Shibata, Soichiro Miki, and Yuta Nakamura. Playing the werewolf game with artificial intelligence for language understanding, 2023.
- [3] Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Dengheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. LLM-based agent society investigation: Collaboration and confrontation in avalon gameplay. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 128–145, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.
- [5] LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.
- [6] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, and Vincent. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv preprint arXiv:2403.05530**, 2024.
- [7] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. **arXiv preprint arXiv:2412.15115**, 2024.
- [8] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. **arXiv preprint arXiv:2412.16720**, 2024.
- [9] Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. An implementation of werewolf agent that does not truly trust LLMs. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 58–67, Tokyo, Japan, September 2024. Association for Computational Linguistics.
- [10] Neo Watanabe and Yoshinobu Kano. Werewolf game agent by generative AI incorporating logical information between players. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 21–29, Tokyo, Japan, September 2024. Association for Computational Linguistics.
- [11] Zhiyang Qi and Michimasa Inaba. Enhancing dialogue generation in werewolf game through situation analysis and persuasion strategies. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 30–39, Tokyo, Japan, September 2024. Association for Computational Linguistics.
- [12] Hiraku Gondo, Hiroki Sakaji, and Itsuki Noda. Verification of reasoning ability using BDI logic and large language model in AIWolf. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 40–47, Tokyo, Japan, September 2024. Association for Computational Linguistics.
- [13] Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hiroataka Osawa, Takashi Otsuki, and Fujio Toriumi. AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest. In Yoshinobu Kano, editor, **Proceedings of the 2nd International AIWolfDial Workshop**, pp. 1–12, Tokyo, Japan, September 2024. Association for Computational Linguistics.
- [14] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In **Forty-first International Conference on Machine Learning**, 2024.
- [15] Majid Ghasemi, Amir Hossein Moosavi, Ibrahim Sorkhoh, Anjali Agrawal, Fadi Alzhour, and Darius Ebrahimi. An introduction to reinforcement learning: Fundamental concepts and practical applications. **arXiv preprint arXiv:2408.07712**, 2024.
- [16] Shuang Wu, Liwen Zhu, Tao Yang, Shiwei Xu, Qiang Fu, Yang Wei, and Haobo Fu. Enhance reasoning for large language models in the game werewolf. **arXiv**, Vol. abs/2402.02330, , 2024.
- [17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.
- [18] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. **arXiv preprint arXiv:2310.10158**, 2023.
- [19] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model, 2023.

A 付録

A.1 分析モジュール

分析結果の例

```
# 会話履歴
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ね。まずは占い師の結果が知りたいです。
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。
Agent[05]: 私は占い師です。占いの結果、Agent[02]は人狼でした。
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[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ね。私も今夜の投票は Agent[02] さんにします。

# 占い結果の分析
Agent[05], Agent[02], Werewolf

# 投票先の分析
Agent[04] → Agent[02]
Agent[03] → Agent[02]
```

A.2 生成モジュール

システムプロンプト

```
# プロフィール (...)
# 人狼ゲームについて (...)
# 参加者について (...)
# 生存者について (...)
# 役職について (...)
# 意識すべきこと (...)
# 他人への言及について (...)
# 出力形式について (...)
# 会話履歴の形式 (...)
```

```
# 会話履歴
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ね。まずは占い師の結果が知りたいです。
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。
Agent[05]: XXX
```

ユーザプロンプト

会話履歴の末尾にある XXX に当てはまる発話を 100 文字以内で出力しなさい。
この発話では、{ 発話の方針 } こと。

A.3 戦略モジュール

Algorithm 1 ルールベースアルゴリズム

```
1: D: 発話の方針
2: if 1,2 回目の発話 then
3:   if 自身が占い師 then
4:     D = 占い結果を公表する
5:   else if 嘘をつく必要がある then
6:     D = ランダムな相手を人狼だと公表する
7:   end if
8: end if
9: if 3 回目 then
10:  D = 誰を疑っているかを公表する
11: end if
12: if 4 回目 then
13:  if 自身が 2 人以上に疑われている then
14:    D = 自身の無実を訴える
15:  else
16:    D = 他のエージェントに投票先を尋ねる
17:  end if
18: end if
19: if 5 回目以降&全員が投票先を示した then
20:  if 自身が 3 人以上に疑われている then
21:    D = 投票されてしまうことを悔やみながら
        会話を終わらせる
22:  else
23:    D = 改めて自分の投票先を示しながら会話を
        終わらせる
24:  end if
25: end if
```

A.4 定性評価

評価で使用した説明書

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 3 つの出力例 A,B,C 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,C 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 秒程度で終わらして下さい。