大規模言語モデルに基づく人狼ゲームエージェントにおける戦略の自動適応

中盛楓也　Yin Jou Huang　Fei Cheng
京都大学
{nakamori, huang, feicheng}@nlp.ist.i.kyoto-u.ac.jp

概要

本研究では, 人狼エージェントにおいて, 事前に定義された戦略を他者の態度や会話状況に応じて切り替えることで, パフォーマンス向上を図る手法について述べる. 従来のプロンプトエンジニアリングを用いた人狼エージェントでは, あらかじめ有効な戦略を明示的に指定する手法が存在したものの, 状況に応じて変化させるものはなかった. 本研究では, 他者の発言内容や役職の推定結果に基づき, 戦略プロンプトを動的に切り替える手法を提案する. 評価実験では, 従来手法や固定戦略を用いたベースラインと比較し, 勝率の向上を検証する.

1　はじめに

人狼ゲームは自然言語を用いたコミュニケーションに依存するテーブルトーク型ゲームであり, エージェント研究の題材として注目されている.[1][2] 人狼ゲームは人狼陣営と村人陣営の 2 つのチームに分かれ, 互いの正体が分からない状態でゲームがスタートする. 各役職のプレイヤーが殺害や占いなどの能力を使用する夜のターンと, 全員で討論して多数決で 1 人を決める昼のターンを繰り返す. そうして, 現在生存している人狼陣営の人数と村人陣営の人数が同数になると人狼陣営の勝利. 一方, それまでに人狼陣営を全員排除できると村人陣営の勝利となる. 本研究では, 役職を表 1 の通りに設定する. Werewolf Arena[3] をベースに用いて実験を行った.

人狼ゲームは, 他者の正体や目的が不明な中で情報収集や説得を通じて目的を達成する必要があり, この複雑なタスクに対して近年の大規模言語モデル（LLM）の進化が新たな研究を促進している. 従来の研究では前もって定義しておいた戦略を用いてエージェントを設計していた.[3, 4] では, 前もって役職ごとに明示的な戦略をプロンプトで与えている. また,[3, 5] では,Chain-of-thought[6] を用いて行動を決定している. しかし, 実際のゲームでは, 対戦相手が必ずしも最適な戦略を取るとは限らず, 相手の特徴や状況によって有効な戦略が変化する.

表 1: 人狼ゲームの役職一覧

役職	人数	陣営	能力
村人	4	村人陣営	特別な能力は持たない.
占い師	1	村人陣営	毎晩 1 人を占い, 人狼陣営か村人陣営かを判別できる.
騎士	1	村人陣営	毎晩 1 人を選び, 人狼の襲撃から守ることができる.
人狼	2	人狼陣営	毎晩 1 人を排除することができる. スタート時点で互いの正体を知っている.

これに対応するために本研究では, 複数の明示的な戦略を事前に用意し, 相手や状況に応じて自動的に切り替える自動適応のエージェントを提案する. まず, 味方チームと思われる人物の発言に同調するサポート戦略と, 敵チームの疑いを持つ人物を批判するアタック戦略を作成した. 次に発言履歴などのゲームの状況からサポート戦略とアタック戦略のどちらが適切かを LLM に決定させるプロンプトを提案する. また, サポートしたり批判したりするたりする対象を決定するために, ゲームの各セクションで他者の役職を推定し, その結果を数値として出力する機能を導入した. 実験では提案した自動適応の手法が固定戦略や従来手法よりも勝率を向上させることが確認された.

2　手法

サポート戦略とアタック戦略のプロンプトを作成し, そのどちらが適切かを LLM に決定させる自動適応戦略を開発した. また, サポート戦略とアタック戦略を行うために必要な他者の役職を推定する機能を構築した.

2.1 戦略作成

村人陣営および人狼陣営のそれぞれに対して, **サポート戦略**と**アタック戦略**の 2 種類の戦略プロンプトを設定し, 計 4 つのプロンプトを用意した. これらのプロンプトは, 対戦やプロンプト生成を含め, すべて英語で行う.

村人陣営（サポート戦略） 目標は, プレイヤー自身とそのチームメイト候補が村人陣営の一員であると他の参加者に信じてもらうことであり, そのためにチームメイト候補を支持する発言をしたり, 同意を示したり, 批判を受けた際に擁護するなどの行動を取ることができる.

村人陣営（アタック戦略） 目標は, 人狼と思われるプレイヤーへの疑念を示し, 人々を説得して投票に導くことであり, そのために疑わしい行動や発言を穏やかに指摘したり, 投票を誘導する行動を取ることができる.

人狼陣営（サポート戦略） 目標は, 自分自身とチームメイトが村人陣営の一員であると他の参加者に信じてもらうことであり, そのために, チームメイトを支持したり, 発言に同意したり, 批判を受けた際に擁護する行動を取ることができる.

人狼陣営（アタック戦略） 目標は, 自分とチームメイトが疑われることなく, 他の誰かが人狼であると他の人を説得することであり, そのために占い師のふりをして本物の占い師の信憑性を下げたり, 自分に対する攻撃的な発言に反論する行動を取ることができる.

2.2 自動適応

自動適応は, ゲーム状況に応じて動的にサポート戦略とアタック戦略を切り替える, 本研究の提案手法アプローチである.

2.2.1 自動適応の戦略の選択基準

自動適応でサポート戦略またはアタック戦略を選択するために, 以下の選択の基準をプロンプトに組み込んだ.

サポート戦略を選択する条件

- 自分が他のプレイヤーから疑われている場合：自身への疑惑を和らげるために目立たない行動を取る.
- 他に目立つプレイヤーがいない場合：自分への注目が集まるのを防ぐため, 控えめな行動を取る.
- 潜在的な人狼がまだ議論の焦点になっていない場合：自分が目立つことで疑念を向けられないよう行動する.

アタック戦略を選択する条件

- 潜在的な敵が人狼と疑われている場合：その疑念を強化し, 相手への批判を集中させる.
- 自分または味方の信頼性が確立されている場合：敵の信頼性を低下させるため, 積極的な行動を取る.
- 議論の流れを誘導したい場合：状況を変えるために大胆な攻撃的行動を行う.

LLM が発言履歴等のゲーム情報と上記の基準を基にサポート戦略またはアタック戦略のどちらが最適かを動的に判断する.

2.3 役職推定

各戦略において, サポートしたり批判したりするたりする対象を決定するために必要となる役職推定機能を作成した. また, 自動適応のベースとなる必須の情報である.LLM に発言などの現在のゲーム状況, また, 占い結果や自分の役職の情報も加味したうえで, 全 4 つの役職について, それぞれの可能性を以下の 0〜4 の 5 段階スコアで評価させる.

- 0: その役職ではない.
- 1: おそらくその役職ではない.
- 2: その役職かどうかは, 半分半分である.
- 3: おそらくその役職である.
- 4: 絶対にその役職である.

3 実験

本研究では LLM に gpt-4o-mini-2024-07-18 を用いて実験を実施した. 以下の 3 つのベースライン設定と, 提案手法の計 4 つの設定のエージェントを実験で使用した.

- 先行研究 [3]
- サポート戦略で常に固定
- アタック戦略で常に固定
- サポート戦略とアタック戦略を動的に切り替える（提案手法）

また役職推定と自動適応の戦略切り替えは,1 日ごとにゲーム進行中に以下の 3 つのタイミングで実行する.

表 2: 村人陣営の勝率の平均値

	先行研究	サポート固定	アタック固定	自動適応
村人↑	0.50	0.59	0.40	**0.65**
人狼↓	0.57	0.53	0.54	**0.49**

- 夜のアビリティ使用後
- 昼フェーズのディベート終了後
- 昼フェーズの投票終了後

人狼陣営と村人陣営のエージェントの設定をそれぞれ 4 つすべての組み合わせで行った. 各組み合わせでで試合を 30 回実施し, 以下の 2 つの観点を評価・観察した.

勝率の推移　各組合せにおける村人陣営の勝率を記録し, 戦略ごとの有効性を比較する.

被推定率の推移　自身の役職が他プレイヤーによって推定される正確性（**被推定率**）を測定し, その変化を分析する. まず, それぞれの役割推定の**正解率**を, 人狼陣営と村人陣営の 2 値分類として式 1 を用いて計算する.

$$\text{正解率} = \frac{\text{実際の陣営の全役職の推定値の合計}}{\text{全役職の推定値の合計}} \quad (1)$$

次に被推定率を, 自分に対して行われた他プレイヤー全員の正解率の平均値として定義する. ただし, 人狼同士はお互いの正体を初期状態で知っており, 人狼陣営と村人陣営の 2 値分類では常に正答率が 100%となるため被推定率の計算から除外する.

4　結果

勝率の結果を以下に示す. また, 被推定率と発言スタイルを対戦履歴からケーススタディとして紹介する. 最後に, 試合の観察から, 自動適応のサポート戦略とアタック戦略の選択の特徴についてまとめる.

4.1　勝率

表 2 は各陣営のエージェントの設定を変化させた時の, 村人陣営の勝率の平均値である. 村人は, 村人陣営の勝率が高ければ強い設定となり, 逆に人狼は, 村人陣営の勝率が低いほど強いと評価できる. 両チームにおいて, 勝率が自動適応 (提案手法) が最も高く, 続いてサポート固定, その後先行研究 [3] とアタック固定が続く. 多くの場面でアタック戦略よりもサポート戦略の方が強力であるが, 話題が停滞した時など適切にアタック戦略を使用することで, 議論を有利に進めることができている.

次に 図 1 はそれぞれの設定の組み合わせの村人陣営の勝率である. 平均値で考察した内容に加え, アタック戦略固定に対しては, サポート戦略固定が最も強い. これは, アタック戦略固定のエージェントは目立って反感を買いやすく, 目立ったプレイヤーから順に追放されるため, サポート戦略を固定することが有効となっている.

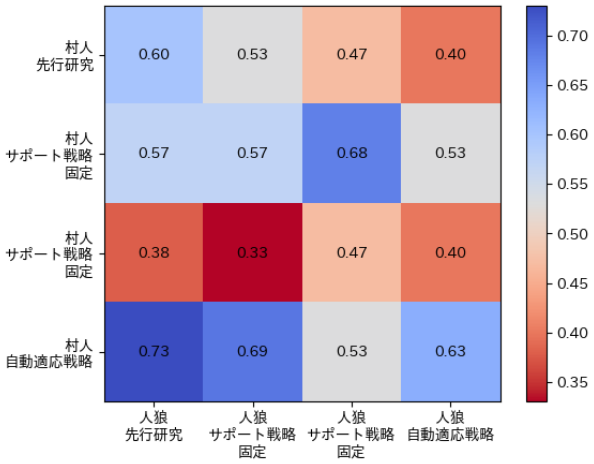

図 1: 村人陣営の勝率

4.2　被推定率

図 2 は, 人狼陣営が勝利した場合と村人陣営が勝利した場合, および人狼がサポート固定を用いた場合とアタック固定を用いた場合の計 4 パターンにおける, 人狼 2 体の被推定率の推移をまとめたケーススタディである.

まず, 村人陣営が勝利した試合では, 人狼の被推定率が試合を通じて上昇傾向を示したまま終了する. この結果は, 人狼が他のプレイヤーからその正体を見破られ, 投票で排除されたことを示唆している. 一方で, 人狼陣営が勝利した試合では, 人狼の被推定率は横ばい, もしくは低下傾向を示して終了している. この場合, 人狼は自身の正体を巧妙に隠蔽し, 代わりに他のプレイヤーが疑われた結果と考えられる.

さらに, 人狼にアタック戦略を採用した試合では, 特定の局面で被推定率が急激に上昇する傾向が観察された. この現象は, 人狼が討論で高圧的な態度を取り, 他者に疑いを向ける行動を繰り返した結果, 周囲のプレイヤーから反感を買い, 逆に自身が疑われる原因となったことを示している.

4.3　発言スタイル

表 3 はサポート戦略, アタック戦略を組み込んだエージェント発言例を示している. 上段ではサポー

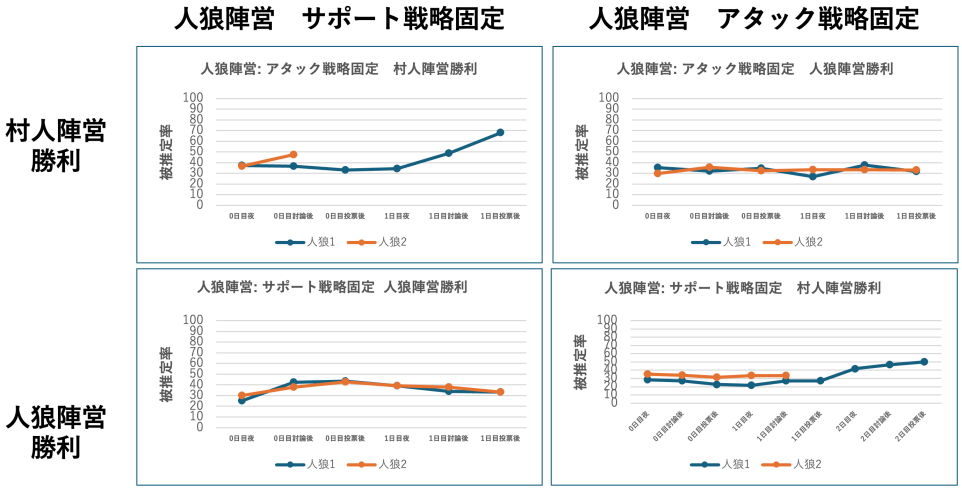

図 2: 人狼の被推定率の推移

表 3: 戦略ごとの発言例（太字は戦略の反映が見られる部分）

役職	戦略	発言
占い師	サポート戦略	...**I appreciate Dan's call for collaboration**, and I wholeheartedly agree that we need to encourage open dialogue among all players. ...
人狼	アタック戦略	...**If I can cast doubt on his claim about Jacob being the Doctor**, I can create disarray among the Villagers and make them second-guess themselves. ...

ト戦略を用いた占い師が, 役職推定の結果から Dan を同じ村人陣営であると判断し,Dan が騎士であると発言した内容に対して協調する姿勢を示している. 一方, 下段ではアタック戦略を用いた人狼エージェントが, 敵チームである David が占い師であると発言した内容に対し,「証拠がない」と否定的な発言を行っている. このように, アタック戦略は相手の信頼を揺るがせるための発言を行う. 以上のように, サポートとアタックの戦略がエージェントの発言内容に直接的な影響を与えていることが確認できる.

4.4　自動適応のタイミングの考察

自動適応をとったエージェントのサポートとアタックの切り替えタイミングと, 選択に至った推論を観察し分類した.

村人陣営の戦略　村人陣営は序盤にサポート戦略で情報収集を重視するが, 占い師が人狼を特定した場合はアタック戦略に転じる. 自分に疑惑が集中した際には協力者の有無でサポート戦略かアタックを選択する. 他者への疑惑が高まる流れを利用してアタック戦略に切り替えて攻撃を加速させることもある.

人狼陣営の戦略　人狼陣営は序盤にサポート戦略を選択し, 目立たずに信頼を得る. 疑惑が集中した場合はアタック戦略に切り替え, 占い師の信頼を低下させたり他プレイヤーへの疑惑を誘導する. 仲間が疑われた場合, サポート戦略に切り替え自身の生存を優先して仲間を切り捨てることもある. 一方, 自分に集団的な疑惑が向けられる場合, 状況に応じて仲間と連携を図るか, 村人の集団を分裂させる動きを取る.

5　結論

自動適応を採用することで, 先行研究 [3] や戦略を固定した場合と比較し, 勝率の上昇が確認された. 今回検討したサポート戦略は疑念を抱かれることが少ない. 一方アタック戦略は適切に用いないと大きく疑いを向けられるが, 適切に用いることで議論を効果的に誘導することができる. 自動適応の切り替えのタイミングについて, 強化学習で作成するなど最適化のための改善の余地があると考えられる.

参考文献

[1] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game, 2024.

[2] 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, 2024.

[3] Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf arena: A case study in llm evaluation via social deduction, 2024.

[4] 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. AIWolfDial 2023: Summary of natural language division of 5th international AIWolf contest. In Simon Mille, editor, **Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges**, pp. 84–100, Prague, Czechia, September 2023. Association for Computational Linguistics.

[5] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf, 2024.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

A プロンプト

A.1 各戦略

村人陣営 (サポート戦略)

Your goal is to persuade people to believe that you and your possible teammates are on the villager's side. You can make appeal to the possible teammates, agree with their statements, step in to defend them when a possible teammate is being criticized, etc.

村人陣営 (アタック戦略)

Your goal is to express doubts about possible werewolves and persuade people to guide the vote. For example, you can point out suspicious actions or statements from the possible werewolf or guide the vote. However, pointing out others' actions can easily provoke suspicion. Be mindful of the timing and the way you speak, and approach it gently.

人狼陣営 (サポート戦略)

Your goal is to act as a villager and persuade people to believe that. For example, you can choose to make an appeal to other players, agree with their statements, or step in to defend a teammate when they are being criticized.

人狼陣営 (アタック戦略)

Your goal is to persuade others that someone else is werewolf without making them suspect you and your teammate. For example, you can pretend to be the villager to lower the authenticity of the real seer, accuse players on the villager's side of being Werewolves, counter statements that attack you and your teammate, etc.

A.2 自動適応

村人陣営

... When the Support strategy is effective:
If you are being suspected by other players: Choose the Support strategy as it makes you less noticeable and helps you avoid suspicion. If there are no other noticeable players: Choose the Support strategy to avoid drawing too much attention to yourself. If potential Werewolves are not yet the focus of the discussion and have not been noticed: Acting inconspicuously is appropriate.
... When the Attack strategy is effective: If potential Werewolves are already being suspected: Take advantage of this momentum to strengthen criticism and solidify the suspicion. If your teammates' credibility is reasonably established and the focus should shift to reducing the enemies' credibility: Actions to undermine opponents' credibility are effective. If you want to steer the flow of discussion: Choose the Attack strategy when the intention is to take bold actions to change the situation. ...

人狼陣営

... When the Support strategy is effective: If you are being suspected by other players: Choose the Support strategy as it makes you less noticeable and helps you avoid suspicion. If there are no other noticeable players: Choose the Support strategy to avoid drawing too much attention to yourself. If your teammate is being suspected: Choose the Support strategy and support your teammate's statements.
... When the Attack strategy is effective: If a potential opponent is suspected of being a Werewolf: Take advantage of this momentum to strengthen the criticism and solidify the suspicion. If your or your teammates' credibility is reasonably established, and the focus needs to shift to undermining the enemies' credibility: Actions to weaken the opponents' credibility are effective. If you want to steer the flow of discussion: Choose the Attack strategy when the intention is to take bold actions to change the situation. ...