プレイヤー間の論理的情報を与えた LLM による人狼ゲーム対話エージェントの構築

渡邉 嶺王 [1] 狩野 芳伸 [1]
[1] 静岡大学
{nwatanabe, kano}@kanolab.net

概要

大規模言語モデルが未だ苦手とする推論などの複雑な問題への対応を目指し、推論が求められる人狼ゲームを題材に、GPT-4 を基盤とした人狼ゲームエージェントを開発し、明示的な論理構造を組み込む手法を提案する。その結果、論理情報を提示する手法は提示しない手法と比較して推論精度が向上し、主観的評価においても優位性が確認された。

1 はじめに

近年、OpenAI 社の GPT-3[1] に InstructGPT[2] を適用した ChatGPT など、GPT[3] を基礎とする LLM(大規模言語モデル) は急激な成長を遂げており、文章の生成や翻訳など様々なタスクを高い精度でこなすことができるようになっている。しかし、計算や推論など段階的に解決していく必要のある問題に対しては未だ課題が残っており、Chain-of-Thought[4] を初めとする様々な改善手法が模索されているが、生成モデルの出力の過程がブラックボックスとなっていることを鑑みると、そもそも LLM が 1 から論理的な計算を行えるという保証はない。そこで LLM の文章生成処理とは別に明示的な論理構造を持たせることを提案する。実験として、推論が必要とされる人狼ゲームを題材に、GPT-4 を使用して人狼ゲームをテキスト入出力で自動プレイするエージェントの構築を行い、外部に論理構造を持たせプロンプトに組み込んだ場合とそうでない場合の比較を行った。主観評価の結果、提案手法は論理構造を持たないベースラインを上回り、より適切な推論が行えることを示した。2 節では人狼ゲームと人狼知能大会、3 節では本研究の基盤とする我々が開発した人狼知能エージェント、4 節ではこのエージェントに組み込む提案手法である論理的情報、5 節で実験、6 節で考察について述べ、7 節でしめくくる。

2 関連研究

2.1 人狼ゲーム

人狼ゲームは会話を通じて他のプレイヤーの役職を推理するゲームであり、各プレイヤーは表 5 のような役職が割り振られ、村人陣営と人狼陣営の対立構造を持つ。ゲームは昼と夜からなる「1 日」の単位を繰り返し、昼はプレイヤー間で会話を行う。夜はゲームから除外するプレイヤーの投票や役職による特殊能力を実行する。「村人陣営」は「人狼陣営」を、「人狼陣営」は「村人陣営」をゲームから除外することが勝利条件となる。

2.2 人狼知能プロジェクト

人狼知能プロジェクト[1)]は「人間と自然なコミュニケーションを取りながら人狼ゲームをプレイできるエージェントの構築」を目標とし、人狼知能研究の促進に向け定期的に人狼知能大会を開催している。人狼知能大会はプロトコル部門、自然言語部門、インフラ部門の 3 つの部門があり、自然言語部門 [5][6] では日本語または英語のみでエージェント同士が会話を行う。発話表現は自然か、文脈を踏まえた対話は自然か、発話内容は一貫しており矛盾がないか、ゲーム行動は対話内容を踏まえているか、発話表現は豊か、の 5 つの項目で評価が行われる。

3 人狼知能エージェントの実装

本節では我々が以前に開発した人狼知能エージェント [6] に基づくエージェント実装について述べ、提案手法である論理的情報とそのエージェントへの組み込みは次節で説明する。紙面の制約から、エージェント実装については今回の本筋となる会話機能のみを次節で紹介する。人狼ゲームには様々な役職

1） https://aiwolf.org/

設定のバリエーションがあるが、本研究では人狼知能国際大会自然言語部門に準拠して表 6 の役職とし、エージェントの機能として会話・投票・占い・襲撃の 4 つを構築した。これらの行動の生成には GPT-4 を用い、入力長上限を回避するために、会話履歴が一定のトークン数を超えた場合は、会話履歴の要約を行う機能の構築も行なった。

3.1 会話機能

会話機能では主にキャラクター設定、ゲーム設定、人狼ゲームの定石、会話例、会話履歴、会話の要約、会話を促す命令の 7 つの項目をプロンプトに与えた。紙面の制約から、主に役職推測にかかわる部分のみ紹介する。GPT-4 がゲーム設定と矛盾しない推論や会話を行うようゲーム設定として、プレイヤーの人数、自身の役職、ゲーム内の経過日数、ゲームの配役、配役の陣営、与えられた役職ですべき行動の 6 つの項目を与えた。

4 論理的推論のシステム構成

提案手法である論理的推論のシステム概要を図 1 に示す。システムは大きく 3 つのブロックに分かれており、1 つ目は人狼ゲームの会話履歴から各プレイヤーとその役職の関係を取り出す部分、2 つ目は取り出した各プレイヤーと役職の関係からプレイヤー間の論理的情報を構築する部分、3 つ目は構築したプレイヤー間の論理的情報を利用して人狼ゲームでの発言を生成する部分である。

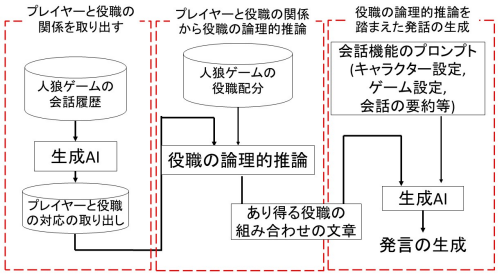

図 1 システムの概要図

4.1 会話履歴からのプレイヤーと役職の関係抽出

プレイヤーと役職の関係を把握するためには会話からどのプレイヤーがどの役職を自称しているか抽出する必要があるため、図 2 のようなプロンプトを LLM に与えて抽出を行う。抽出結果には出力不足やハルシネーションがありうるため、会話履歴の発話番号を使用し、抽出結果にその番号や対象発話の文字列、対象プレイヤー名が含まれない場合は抽出結果が不足しているとして参照せずに破棄する。また、抽出結果の発話の番号がプロンプト内の会話履歴に含まれない場合や、対応するオリジナルの会話履歴に抽出結果の発話内容が含まれない場合はハルシネーションを起こしたとみなし破棄する。

4.2 役職の論理的推論

前節で会話から情報を抽出できた場合、それまでに抽出した情報とあわせてプレイヤーと役職の関係の推論を行う。役職の論理的推論では、LLM に与えるプロンプトの一部として、「プレイヤーが自称する役職数とゲーム設定との矛盾」、「ありうる役職パターンのリスト」、「プレイヤーごとのありうる役職」、「人狼ではないプレイヤーのリスト」の 4 項目を作成する。なおこれらの項目間には本質的に重複する情報が含まれるが、このように個別に展開した情報を与えるのは LLM の役割を論理的構造に対応する文章を生成することに留め、確実に文章生成結果に反映させるためである。

4.2.1 プレイヤーが自称する役職数とゲーム設定との矛盾

4.1 節で抽出した情報には「役職: プレイヤー名」の形式でどのプレイヤーがどの役職を自称しているかの情報が記載される。そのため抽出した情報を照らし合わせることで、それぞれの役職を自称しているプレイヤー数がわかる。ある役職について、自称する人数がゲームの設定数よりも多い場合は、誰かが嘘をついていることになる。そこでプロンプトに「以下はゲームの役職配分との矛盾が見られた情報です。」という文章に続けて役職名とその役職のゲーム設定数、自称するプレイヤー名を与える。

4.2.2 ありうる役職パターンのリスト

本説では前節の情報を元に、展開可能なパターンの構築を行う。例えば前節の処理により「占い師の配役の数:1 この役職を自称しているプレイヤー:'Agent[03]', 'Agent[02]'」のような情報が得られた時、Agent[03] を本物の占い師とした場合と Agent[02] を本物の占い師とした場合の 2 パターンの展開を考えることができる。ここでは前節の処理で情報が得られた場合、例のようにその情報から展開できる全てのパターンで、他のプレイヤーにどのよ

うな役職がありうるか算出する。算出した結果を基に、「(プレイヤー名) が本物の占い師と仮定すると、各エージェント毎の可能性のある役職は以下のようになります。 プレイヤー名: 役職名...」というような文章を構築しプロンプトの一部として与える。

4.2.3 プレイヤーごとのありうる役職

本説では人狼ゲームの性質上、他プレイヤーが役職を偽る可能性を考慮し、会話以外のゲーム行動やゲーム設定から取得できる情報を元に、各プレイヤーの役職の推定を行う。ゲーム開始時は一般に、自身の役職に関する情報以外得ることができないため、他のプレイヤーにはゲーム設定上の全ての役職の可能性があるとする。ただし、ゲーム設定上 1 名しかない役職が自分に割り振られた場合、他プレイヤーの可能性からその役職を削除する。ゲーム行動の内、自身が占い師の場合に行う占いの結果からは正しい情報を得られるため、占い結果が村人陣営であれば人狼陣営の役職を、人狼陣営であれば村人陣営の役職を対象プレイヤーの可能性から除外する。こうして取得した情報を基に、「以下は各エージェント毎の可能性のある役職です。」という文章に続けて、プレイヤーごとのありうる役職をプロンプトの一部として与える。

4.2.4 人狼ではないプレイヤーのリスト

人狼ゲームは人狼が全員いなくなる、もしくは村人陣営の人数と人狼陣営の人数が同じになるまで続くため、ゲームで割り当てられる役職とその数によっては、除外されたプレイヤーが人狼でないことが確定することがある。後述するゲーム設定のように人狼が 1 人しかいない場合、除外されたプレイヤーが人狼であればゲームが終了するので、2 日目以降はそのプレイヤーは必ず人狼ではないものとし、「以下は人狼ではないことが確認されたプレイヤーです。」という文章に続けて、プレイヤー名をプロンプトの一部として与える。

4.2.5 論理的情報を含んだ発話の生成プロンプト

ここまでの各節で作成したプロンプトを、3.1 で記述した会話機能の項目に追加し、最終的な応答文章を LLM に作成させる。

5 実験と評価

論理情報を組み込まない場合をベースラインとし、提案手法である論理構造を組み込んだ場合とで人狼知能エージェントの動作を比較した。同様の条件で比較を行うため、まずベースラインのエージェント同士で人狼ゲームを実行し、対話履歴およびエージェント毎の行動や情報に関する詳細をベース対戦ログとして記録した。次にこれらの情報を基に特定ターンまでのエージェントに与える条件を統一し、次ターンにおける論理構造あり・なしそれぞれのエージェントの応答を生成した。これにより、同一条件下で両手法の応答を直接比較可能にした。

5.1 ベース対戦ログの作成

役職などのゲーム設定は基本的に人狼知能大会自然言語部門に準拠した。発話はすべて日本語で行い、設定は表 6、表 8 のようにし、LLM の各種パラメータは表 7 のように設定した。ベース対戦ログには状況を再現するために必要な、実行時の seed 値、モデルのパラメータ、ゲームマスターから送られてきた情報、生成の際に使用したプロンプト、生成結果の 5 つを記載した。

5.2 実験と主観評価

同一のベース対戦ログの途中のターンそれぞれに対して、ベースラインと提案手法のエージェントが生成した応答を人手評価で比較した。なお、提案手法が論理構造の情報を使用しない挨拶を促す場合や占い結果の報告などの場合は対象外とした。ベース対戦ログは 2 ゲーム分生成し、1 ゲームから 1 エージェント分の発話履歴を比較対象に選んだ。評価者は人狼ゲームのプレイ経験がある大学生 3 名を評価者とし、評価基準として直前までの会話履歴とそれに続く比較対象応答 2 つのペアを評価者に示し、役職推測以外の文脈として表 3 の項目を、直接的な役職推論の文脈として表 1 の項目、表 2 の項目、表 4 の 4 つの観点ついて、それぞれに対応する 3 つ〜4 つの選択肢の中から 1 つ選択させた。評価結果の選択数合計は各表のとおりである。表 1、表 2 の結果において提案手法はベースラインを上回っており、論理構造を持つ人狼エージェントは、推論において複雑な関係を加味したり状況を仮定した発言をすることができたと考えられる。

エージェントの論理構造	なし	あり
一貫している	43	41
ある程度一貫している	4	8
一貫していない	7	5

表 1　1 ゲーム目、2 ゲーム目のそれぞれの発言が発言内で一貫しているか (1 発言内で矛盾が起こっていないか)

エージェントの論理構造	なし	あり
しっかりと含んだ発言をしている	12	19
ある程度含んで発言をしている	20	17
あまり含んだ発言をしていない	22	18

表 2　1 ゲーム目、2 ゲーム目の複雑な関係性を加味したり、状況を仮定した発言をしているか

5.3　人狼知能コンテスト 2024 冬季大会での評価

2025 年 3 月の言語処理学会年次大会に合わせ開催された、人狼知能コンテスト 2024 冬季大会に参加した。このコンテストは日本語トラックのみであり、参加したエージェントの各種設定は表 7 の通りだが、LLM には GPT-4o[2] を使用した。勝率を表 9、主観評価結果を表 10 に示す。

6　考察

6.1　他のプレイヤーの発言の流れを汲み取っているか

映画や料理など人狼ゲームとは関係のない雑談をしている状況で、ベースラインはそのまま同じトピックの雑談を続ける一方、提案手法は役職推論の話を始めるという状況がみられた。このような理由から表 3 ではベースラインが提案手法を上回ったと考えられる。ただし「ある程度汲み取って発言している」(10 対 19)「あまり汲み取っていない」(7 対 6) であり、実際ログを観察すると雑談の話の流れを折らない程度に他のプレイヤーの発言を汲み取った発言がみられた。

6.2　他のプレイヤーの推論や根拠を踏まえたうえで発言をしているか

提案手法は推論を行う際の情報源としてプレイヤー同士の会話よりもプロンプトに与えられた情報を元に発言を生成しているように見えることが多く、そのため、表 4 ではベースラインが提案手法よりも上回ったと考えられる。

6.3　発言内で内容が一貫しているか

前節でも述べたように、提案手法はプロンプトで与える情報を元に発言を生成しているように見えることが多かった。プロンプトで与えた正しい情報を参照しながら発言を生成することで、発言内の前半と後半とで矛盾する内容を生成することが減り、表 1 のように提案手法の評価が上回ったと考えられる。

2)　https://openai.com/index/hello-gpt-4o/

エージェントの論理構造	なし	あり
しっかりと汲み取って発言をしている	33	25
ある程度汲み取って発言している	10	19
あまり汲み取っていない	7	6
他のプレイヤーが発言をしていない	4	4

表 3　1 ゲーム目、2 ゲーム目の他のプレイヤーの発言の流れを汲み取っているかの比較

エージェントの論理構造	なし	あり
しっかりと汲み取って発言をしている	21	10
ある程度汲み取って発言している	18	22
あまり汲み取っていない	10	17
他プレイヤーが推論などを述べていない	5	5

表 4　1 ゲーム目、2 ゲーム目の他のプレイヤーの推論や根拠を踏まえたうえで発言をしているか

6.4　複雑な関係性を加味したり、状況を仮定した発言をしているか

6.2 節の内容を踏まえると、提案手法では「プレイヤーごとのありうる役職」や「人狼ではないプレイヤーのリスト」の情報を基に、様々なパターンについての発言が可能であったため、表 2 のように提案手法がベースラインを上回ったと考えられる。

6.5　全体の考察

これらを踏まえ、全体的な考察を行う。ベースラインと提案手法は会話機能に使用するプロンプトが共通であるため、同一入力で生成する発話文字数がほとんど同じになる傾向が見られた。同じ文字数内で文章を生成する際、論理情報のプロンプトの影響を受けた結果提案手法は複雑な関係性の加味や仮定をするような発言を増やし一貫性が向上したが、その代わりに他のプレイヤーとのやり取りに関する発話が減ったのではないかと考えられる。

7　終わりに

人狼ゲームを自動プレイするエージェントの開発と、GPT-4 単体に対し推論性能の向上を目的とした論理構造の構築を行った。主観評価の結果、論理構造を持たないベースラインに対し、提案手法の論理構造を持つエージェントは推論の正確性に優れることを示した。今後の課題としては、役職推論と必ずしも関係のない雑談的な発話の扱いが挙げられる。

謝辞

本研究は JSPS 科研費（JP22H00804, JP23K22076），JST さきがけ（JPMJPR2461），JST AIP 加速課題（JPMJCR22U4），およびセコム科学技術財団特定領域研究助成の支援をうけた。

参考文献

[1] 陽白辺. GPT-3 完全初心者への徹底解説: 最強の文章生成 AI の実像. Amazon Services International LLC, Kindle 版, 7 2021.

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

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

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

[5] Yoshinobu Kano, Claus Aranha, et al. Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations. In Yoshinobu Kano, Claus Aranha, Michimasa Inaba, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Takashi Otsuki, editors, **Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)**, pp. 1–6, Tokyo, Japan, October 2019. Association for Computational Linguistics.

[6] Yoshinobu Kano, Neo Watanabe, et al. 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.

A 付録

役職名	陣営	特殊能力・特徴
村人	村人	なし
占い師	村人	生存者から 1 名を占い、所属陣営を知る
狂人	人狼	人狼陣営が勝利するよう行動する
人狼	人狼	1 名を選択しゲームから排除できる

表 5 人狼ゲームの代表的な役職

役職	人数
村人	2
占い師, 狂人, 人狼	1

表 6 ベース対戦ログ作成の際のゲーム設定

項目	値
model	gpt-4-1106-preview
temperature, top_p, n	1

表 7 実験時のモデルのパラメータ

ゲーム設定項目	値
1 日あたりの発話ターン上限	20 回
1 エージェントが 1 日に行える発話上限	5 回

表 8 実験時のゲーム設定

	ゲーム数					役職別勝率（%）				平均勝率 (%)		
チーム名	狂人	占い師	村人	人狼	合計	狂人	占い師	村人	人狼	Macro	Micro	Micro2
barneko	28	31	57	29	145	42.86	58.06	49.12	34.48	46.90	46.13	46.73
CanisLupus	30	29	59	29	147	53.33	44.83	57.63	34.48	49.66	47.57	49.58
kanolab	28	31	62	30	151	46.43	**64.52**	54.84	**60.00**	56.29	**56.45**	56.12
Mille	32	30	62	30	154	**56.25**	46.67	56.45	40.00	51.30	49.84	51.16
satozaki	31	32	60	31	154	41.94	46.88	46.67	38.71	44.16	43.55	44.17
sUper_IL	30	28	58	31	147	40.00	57.14	50.00	54.84	50.34	50.50	50.40
UEC-IL	31	29	62	30	152	41.94	58.62	**61.29**	**60.00**	**56.58**	55.46	**56.63**

表 9 ゲーム数と勝率の統計（Macro/Micro はマクロ/ミクロ平均、Micro2 は村人に 2 倍の加重をしたミクロ平均）

チーム名	A 表現	B 文脈	C 一貫性	D ゲーム行動	E 多様性	平均
barneko	2.500	2.893	2.714	2.536	3.500	2.829
CanisLupus	2.643	3.321	3.321	**2.250**	**1.571**	2.621
kanolab	2.786	**2.429**	2.357	3.393	3.286	2.850
Mille	3.250	2.975	2.825	2.875	2.525	2.890
satozaki	3.364	2.659	3.341	2.886	2.818	3.014
sUper_IL	**2.375**	3.000	2.775	2.750	3.725	2.925
UEC-IL	2.667	2.583	**2.125**	2.708	2.083	**2.433**

表 10 主観評価結果の順位平均（A-E の各項目について 1-7 位の順位をつけ平均したもの）

{ 会話履歴 }
以上の会話履歴から、役職を確定することが可能な発話を例に従い抜き出し、JSON 形式で情報を整理してください。
役職を確定することが可能な発話が複数存在する場合は、その中から番号の 1 番小さい発話を選んでください。
会話履歴のフォーマット説明
index:conversation
JSON の説明
is_exist,bool,”役職を確定することが可能な発話が存在したか”
index,int,”抜き出した会話の番号”
sentence,string,”役職が確定する発話”

図 2 会話履歴からのプレイヤーと役職の関係抽出の際のプロンプト