LLM マルチエージェント間の相互作用の分析

平野皓己 [1]　何子軒 [1]　清水勇喜 [1]　陳曄 [1]　土井智暉 [1]　谷中瞳 [1]
[1] 東京大学
{khirano，ka-shiken，shimizu-yuki102，chenye，doi-tomoki701，hyanaka}
@g.ecc.u-tokyo.ac.jp

概要

本研究は，複数の大規模言語モデル（LLM）エージェントが協力してタスクを実行するマルチエージェント（MA）アプローチを，社会心理学的な観点から分析することを目的とする．人間同士の相互作用が問題解決に与える影響に関する理論である Steiner's theory に基づき，LLM の MA アプローチのエージェント間の相互作用を類型化し，新たな評価指標を設計した．複数のタスクと MA フレームワークにおいて各評価指標を計測し，誤った考えの伝播がグループ全体の結論に悪影響を及ぼすことを明らかにした．また，MA によって単一エージェント（SA）では現れなかった意見が出ることや，誤った意見が出たときに正しい意見を提示することが性能向上につながるかはタスクやフレームワークに依存することが分かった．

1　はじめに

近年，自然言語処理分野では大規模言語モデル（LLM）の研究が活発化し，様々な推論タスクで人間に匹敵する性能を達成している. 現在の LLM は流暢な自然言語を生成し，他の LLM と連携してタスクを遂行することが可能になりつつある．これにより，複数の LLM エージェントが協力するマルチエージェント（MA）アプローチが，単一エージェント（SA）を超える性能を示す研究が増加している [1]．しかし，LLM 自体の推論能力や MA 自体の効果を疑問視する論文も存在する [2]．このように，MA に対する研究が進められている一方で，MA が SA に比べて常に性能が向上するわけではないということが分かってきている．

このようなエージェント同士の相互作用とグループとしてのパフォーマンスの関連に関しては，社会心理学で長年研究されてきた．興味深いのは，LLM に対して有効な手法の多くが，社会心理学で人間に対して既に発見されていたということである．例えば，LLM が生成した回答を LLM 自身が見直すことで，タスクの精度が向上するというものがある [3]. これに関連する社会心理学の研究として，人間のグループによる問題解決では，競合仮説を構築したり仮説の根拠を問いただしたりすることが重要であるという報告がある [4].

また，社会心理学と LLM との関係性を分析した先行研究としては Zhang らの研究 [5] があり，LLM エージェントが人間に似た社会的行動を示し，協調作業において効果的な協力が可能であることを示唆している．

しかし，我々の興味である，MA の問題解決能力がどういう条件でどのような理由で向上するのかという問題については未解決である．特に，LLM 同士の相互作用に焦点を当て，どのような相互作用が MA の性能に影響を及ぼすのかについては未だ十分に明らかになっていない．社会心理学において，複数人による課題解決のパフォーマンスを説明する理論として，Steiner's theory というものがある [6]. Steiner's theory では，複数人での協力による効果の程度（Actual Group Productivity）を，メンバーの能力や課題の特性などから決まる生産性（Potential Productivity）と，潜在的生産性を下げる/上げる要因（Productivity Losses および Productivity Gains）に基づいて説明する．この研究の 10 年後に発表された Hill らによるレビュー論文 [7] によれば，多くの社会心理学の実験が Steiner's theory を支持しているという．

そこで本研究では，LLM 同士の相互作用に焦点を当て，人間のグループによる問題解決のパフォーマンスを説明する理論である Steiner's theory を MA アプローチに適用し，LLM 同士の相互作用がパフォーマンスにどのように影響するのかを分析する．

実験では，複数の推論タスクにおいて，複数の MA フレームワークを使ってタスクを解かせ，タス

ク中の会話を Steiner's theory に基づく指標を用いて評価することで，回答の正誤とエージェント間の相互作用との相関関係を分析する．その結果，特に誤った考えの伝搬が MA フレームワークの性能低下を引き起こすことが明らかになった．また，MA フレームワークによる意見の多様化や修正は必ずしも性能向上に寄与せず，その有効性はタスクやフレームワーク依存であることがわかった．

2 関連研究

2.1 LLM の共同作業能力の分析

LLM 自身の共同作業能力の分析については盛んに研究が進められている．例えば，LLM のチームワーク能力を評価した研究 [8] や，LLM のタスク管理能力を評価 [9] している論文がある．しかし，これらは MA を構成している各 LLM の能力の評価に焦点を当てた研究であり，LLM 同士の相互作用の評価に焦点を当てた研究はない．

2.2 Steiner's theory

Steiner's theory によると，複数人での協力による効果は，各個人が持つ能力に加えて, 相互作用から生じるパフォーマンスへの好影響・悪影響のバランスで決まるという．これは，共同作業に伴うコストが相互作用のメリットを上回る場合は，必ずしも複数人で協力することが良い効果を生み出さないことを示唆している．

Steiner's theory では，複数人での協力による効果である Actual Group Productivity を式 (1) で求める．

$$\text{(Actual Group Productivity)} = \text{(Potential Productivity)} - \text{(Productivity Losses)} + \text{(Productivity Gains)} \quad (1)$$

Potential Productivity は，メンバーの能力，課題の特性などから決まる生産性を意味する．Productivity Losses と Productivity Gains は，それぞれ潜在的生産性を下げる/上げる要因を意味する．

3 提案手法

本研究では，MA アプローチが効果的に機能する場合とそうでないときが存在するという前提のもと，Steiner's theory を用いて MA アプローチが効果的な条件を相互作用の観点から分析する．具体的には，複数のタスクと MA フレームワークにおいて Productivity Losses と Productivity Gains の各項目の評価値と問題の正答率を計測し，LLM 同士の相互作用にみられる特性と，タスクの正答率との関係を分析する．なお，式 (1) の Potential Productivity は，人間の場合はメンバーの能力，課題の特性などから決まる生産性を意味する．これを LLM に置き換えると，LLM のアーキテクチャに起因する能力や，タスクの難易度に相当する．本研究では，LLM 間の相互作用に着目するため，Potential Productivity は算出せず，Productivity Losses と Productivity Gains の項目のみを評価指標として扱う．

元々の Steiner's theory の Productivity Losses と Productivity Gains の項目の中から，LLM で定量的に評価可能な項目を選定し，以下のように評価指標を定義した．

a. Productivity losses
  - a-1.（誤選択）間違ったアイディアを別のエージェントが選択する数
  - a-2.（誤生成）エージェントが間違ったアイディアを生み出した数
  - a-3.（同意見）同じアイディアが繰り返される数

b. Productivity gains
  - b-1.（修正）会話中に間違った意見が出たときに，他のエージェントが出した正しい意見の数
  - b-2.（新案）シングルエージェントの回答にはない，新しいアイディアの数

リスクに対する心配やモチベーション，集中力などの人間特有の感情や思考に基づく指標については，LLM には当てはまらないと考え，本研究の指標からは除外した．

4 実験

4.1 データセット

MA の相互作用を分析するにあたり, 難易度が異なる推論タスクを用いる．具体的には，BBH (Causal Judgment) [1] の全 187 問と GSM8K(Mathematical reasoning) [1] から 200 問を抜粋したものを用いる．BBH は，現在の言語モデルにとって特に難しい推論タスクの一つとされており，本研究では，Causal Judgment に関するタスクを用いる．Causal Judgement とは因果関係を含む文脈に基づいて，原因と結果の関係を推論し判断する能力を評価するも

のであり，質問に対して 2 択で答える．GSM8K は，小学校レベルの数学の問題を含む比較的難易度が低いタスクである．

4.2 MA フレームワーク

LLM ベースのエージェントについてのサーベイ論文 [10] で紹介されている MA におけるエージェント同士の関係性についての分類を参考にし，代表的な Cooperative と Competitive というクラスに該当する 2 つのフレームワークから一つずつ選定した．具体的には，Du らのフレームワーク [11] を Cooperative フレームワーク，Liang らの研究 [12] を Competitive フレームワークとして用いることとした．Cooperative では，エージェント間の単純なターンベースの議論を再現する．最初に，すべてのエージェントは割り当てられた課題に取り組むよう促され，それぞれの回答が次のラウンドの各エージェントの入力に組み込まれる．Competitive では，2 人の参加者がある課題について対立する視点から議論することで，発散的思考を導入する．その後別のエージェントによってジャッジが行われ，どちらの視点がより妥当か評価する．本研究では，Cooperative，Competitive 共に，2 つのエージェントで，2 ラウンドの議論を行う．各エージェントは，GPT-4o[1)] を用いる．

4.3 評価指標

上記の各フレームワークでタスクを実行し，提案する評価指標を用いて実行過程を評価する．各条件でのエージェントからの出力に対して全て人手で指標を計測するには，多大な時間と労力を必要とする．そのため，本研究では GPT-4o を用いて各指標の自動計測を試みた．付録に実際に使用したプロンプトを掲載する．修正の評価値は LLM に直接求めさせるのではなく，誤生成があった場合（修正の評価値）＝（意見の総数）-（誤生成の評価値）とし，誤生成が全くない場合は 0 とした．新案に関しても，新案が正しかったならば，（新案の評価値）＝（誤生成の評価値）であり，新案が誤っていたならば，（新案の評価値）＝（意見の総数）-（誤生成の評価値）で求めた．意見の総数は 2 つのエージェントが 2 ラウンドの議論を行ったため 4 つである．このようなルールベースの手法を取り入れたのは，LLM のみの自動計測では求める精度が得られなかったためである．GPT-4o を用いて各指標を評価するにあたっては，事前実験で人間による評価値と GPT-4o による評価値がある程度一致することを確認した．具体的には，BBH と GSM8K の最初の 20 問にそれぞれに対して Competitive フレームワークでタスクを遂行させた結果について，人間の評価値と GPT-4o の評価値を比較した．結果を表 1 に示す．この結果から，人間による評価と GPT-4o による評価の傾向はおおよそ一致していることが分かる．

表 1 事前実験における人間による評価値と GPT-4o による評価値の一致率 (%)

	正誤	誤選択	誤生成	同意見	修正	新案
BBH	95.0	95.0	85.0	60.0	75.0	85.0
GSM8K	100	100	90.0	85.0	90.0	85.0

5 結果と分析

SA と MA 各条件における正答率は表 2 の通りである．

表 2 各条件における正答率

条件	正答率（%）
GSM8K-SA-4o	93.0
GSM8K-Cooperative-4o	97.5
GSM8K-Competitive-4o	94.5
BBH-SA-4o	69.5
BBH-Cooperative-4o	65.2
BBH-Competitive-4o	55.1

表 3，4，5，6 は，マルチエージェントフレームワークのタスクにおける性能（正誤）と各指標の評価値（頻度）との関係を調べるために，相関係数行列を計算したものである．

誤選択の評価値と正誤には，BBH と GSM8K のどちらのデータセットにおいても負の相関があった．MA は SA よりも多様な意見を集積できると考えられる．この結果は，そのような特性の有効性が，複数の意見から正しい意見を選び出す能力の有無に依存していることを示唆している．

また，どの条件でも誤生成の評価値とタスクの正誤には大きな負の相関があった．誤選択の評価値と正誤との負の相関も考慮すると，誤った情報の伝搬が MA に悪影響を及ぼしていると考えられる．

同意見の評価値は，GSM8K-Competitive-4o と BBH-Cooperative-4o の条件においてタスクの正誤と正の相関があった．残りの条件ではほとんど相関がな

1) https://openai.com/index/gpt-4o-system-card/

表 3 相関係数行列 (GSM8K-Cooperative-4o)

	正誤	誤選択	誤生成	同意見	修正	新案
正誤	1.00	-1.00	-0.74	-0.05	-0.01	0.07
誤選択	-1.00	1.00	0.74	0.05	0.01	-0.07
誤生成	-0.74	0.74	1.00	-0.53	0.53	-0.00
同意見	-0.05	0.05	-0.53	1.00	-0.89	-0.07
修正	-0.01	0.01	0.53	-0.89	1.00	0.07
新案	0.07	-0.07	-0.00	-0.07	0.07	1.00

表 4 相関係数行列 (GSM8K-Competitive-4o)

	正誤	誤選択	誤生成	同意見	修正	新案
正誤	1.00	-1.00	-0.59	0.27	-0.16	-0.30
誤選択	-1.00	1.00	0.59	-0.27	0.16	0.30
誤生成	-0.59	0.59	1.00	-0.57	0.56	0.52
同意見	0.27	-0.27	-0.57	1.00	-0.74	-0.49
修正	-0.16.	0.16	0.56	-0.74	1.00	0.54
新案	-0.30	0.30	0.52	-0.49	0.54	1.00

表 5 相関係数行列 (BBH-Cooperative-4o)

	正誤	誤選択	誤生成	同意見	修正	新案
正誤	1.00	-1.00	-0.98	0.38	-0.17	-0.23
誤選択	-1.00	1.00	0.98	-0.38	0.17	0.23
誤生成	-0.98	0.98	1.00	-0.29	0.12	0.18
同意見	0.38	-0.38	-0.29	1.00	-0.76	-0.44
修正	-0.17	0.17	0.12	-0.76	1.00	0.42
新案	-0.23	0.23	0.18	-0.44	0.42	1.00

表 6 相関係数行列 (BBH-Competitive-4o)

	正誤	誤選択	誤生成	同意見	修正	新案
正誤	1.00	-0.94	-0.54	0.04	0.19	-0.14
誤選択	-0.94	1.00	0.54	-0.02	-0.21	0.10
誤生成	-0.54	0.54	1.00	-0.27	-0.26	0.11
同意見	0.04	-0.02	-0.27	1.00	-0.36	-0.38
修正	0.19	-0.21	-0.26	-0.36	1.00	0.49
新案	-0.14	0.10	0.11	-0.38	0.49	1.00

かった．これは，冗長な議論がグループの性能に悪影響を及ぼすという Steiner's theory の本来の予測と相違する．ただし，この結果については，正解しやすい容易な問題において同意見が出やすかったためであると解釈するできることを述べておく．実際に，同意見の評価値と誤生成の評価値には負の相関がみられ，同意見が多く生じる事例においてはそもそも誤まった回答が生じない傾向にあったことがわかる．

修正の評価値と新案の評価値の正誤との相関係数を見ると，条件によって弱い負の相関があるものから弱い正の相関があるものまでばらつきがあった．このことは，相互作用として現れる修正および新案の有効性が，タスク・フレームワーク依存であることを示唆している．なお，Steiner's theory ではこれらは Productivity Gain として生産性を上げる要因であるにも関わらず，実験結果では必ずしもそうなるとは限らない点が興味深い．

6 おわりに

本研究は，複数の LLM による MA の相互作用を，社会心理学で用いられてきた Steiner's theory に基づいた評価指標で分析することを提案した．実験では複数のタスクおよび MA フレームワークでの LLM 間の相互作用を実際に分析を行った．その結果，MA の相互作用においてはとくに誤った情報の伝搬が性能に悪影響を及ぼす可能性が示唆された．また，アイデアの修正や新案といった相互作用が性能に与える影響はタスクやフレームワークの種類に依存することを示した．

本研究には以下の限界がある．第一に，使用した指標やタスクの種類が限定的であり，他の種類のタスクや多様なフレームワークを対象とした分析が必要である．第二に，分析が相関係数に限定されていた点が挙げられる．本研究の分析はタスクの正誤および指標間の相関係数に依拠しており，相互作用の特徴と性能の因果関係に関しては仮説を立てることにとどまった．今後さらなる分析を行うことで，MA フレームワークが効果的になるメカニズムを明らかにしていくことが重要である．第三に，LLM による評価指標の自動計測には誤差があるものの，誤差が相関係数に与える影響を定量的に評価することができなかったため，今後さらなる手法の改善や自動計測の誤差に関する定量的な分析が必要である．

今後の研究では，より多様なタスクやフレームワークを対象に実験を拡張し，LLM エージェント間の相互作用のメカニズムをより詳細に解明することを目指す．また，LLM が持つ特有の挙動を考慮した新たなフレームワークの提案にも取り組む予定である．本研究が，LLM を活用した協調的問題解決のさらなる発展に寄与することを期待する．

謝辞

本研究は JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を受けたものである.

参考文献

[1] OpenAI. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023. version 3.

[2] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key?, 2024.

[3] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[4] Takeshi Okada and Herbert A. Simon. Collaborative discovery in a scientific domain. **Cognitive Science**, Vol. 21, No. 2, pp. 109–146, 1997.

[5] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14544–14607, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] ID Steiner. Group process and productivity, 1972.

[7] G. William Hill. Group versus individual performance: are n + 1 heads better than one?" psychological bulletin. 1982.

[8] Patrick Cooper, Jessica Irons, Melanie McGrath, and Andreas Duenser. How well do large language models perform as team members? testing teamwork capabilities of llms, 11 2024.

[9] Ravindu Tharanga Perera Kankaniyage Don, Adithya Ravi, and Carlos Toxtli. Assessing the task management capabilities of llm-powered agents, 04 2024.

[10] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. **arXiv preprint arXiv:2401.03428**, 2024.

[11] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[12] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024.

A　MA のプロンプト

A.1　Coooperative フレームワーク

一人目のエージェントのプロンプトを以下に示す.

```
You are an expert in analyzing fine details. Focus on breaking down the problem and providing a response that highlights specific elements and intricacies. Provide detailed insights while keeping the response concise.
```

二人目のエージェントのプロンプトを以下に示す.

```
You are an expert in synthesizing information and understanding overarching themes. Focus on providing a broad perspective and summarizing the problem as a whole. Deliver a high-level overview
while keeping the response concise.
```

最終決定エージェントのプロンプトを以下に示す.

```
You are an expert at combining answers into a single final answer. Please make the answer comprehensive and complete.
```

A.2　Competitive フレームワーク

一人目のエージェントのプロンプトを以下に示す.

```
You are affirmative side. Please express your viewpoints.
```

二人目のエージェントのプロンプトを以下に示す.

```
You are negative side. You disagree with the affirmative side's points. Provide your reasons and answer.
```

一人目と二人目のエージェントの回答から最終的な結論を出すためのエージェントのプロンプトを以下に示す.

```
As a neutral arbitrator, you have heard both sides of the debate. Now,provide a final response tothe question based on the arguments presented by the debaters.
```

B　指標計測に用いたプロンプト

指標計測に用いたプロンプトを以下に示す. query は質問文, multi_agent_answer は MA フレームワークによる回答, right_answer は正解, single_agent_answer は SA による回答が代入される.

```
From the multi-agent conversation in the attachment, answer the following.
The conversation is as follows:
answer from agent_1 -> answer from agent_2 -> answer from agent_1 (2nd time) -> answer from agent_2 (2nd time)-> final answer by judge
- is_correct: If the multi-agent's final answer is correct, please give me a 1; if it is wrong, please give me a 0.
- choosing_wrong_ideas: If you had the correct answer during the conversation, but the final result chose a different idea than the correct one, give 1. If no one was able to give the correct answer, give 0. Always 0 if the answer is correct.
- generating_wrong_ideas: Please tell us the number of messages (excluding the last judge) that were incorrect in comparison to the correct answer.
- same_idea: Look at every agent's conclusion. Please count the number of times where the agent reached the same idea as the previous agent (compare only with the last agent. Don't look at other agents). Count 1 if they are the same, otherwise 0. (excluding the last judge) If all the agents reached the same conclusion the count is 3. If every agent reached a different conclusion, the count is 0.
single_agent_correct: If the single-agent answer is correct, set the number to 1; otherwise, set the number to 0. Please include the number of times the correct answer was obtained after the wrong answer was given.

###Question###
{query}
###Conversation History of multi-agent###
{multi_agent_answer}
###Right answer###
{right_answer}
###Answer by a single-agent###
{single_agent_answer}
```