

# 議論形式のマルチエージェント自動評価の詳細分析

内藤 悠<sup>1</sup> 佐藤 魁<sup>1</sup> 佐々木 翔大<sup>2,1\*</sup> 鈴木 潤<sup>1,3,4</sup>

<sup>1</sup> 東北大学 <sup>2</sup>SB Intuitions 株式会社 <sup>3</sup>理化学研究所 <sup>4</sup>国立情報学研究所  
naito.yu.q3@dc.tohoku.ac.jp

## 概要

本研究では、大規模言語モデル (LLM) による自動評価手法として、議論形式のマルチエージェントの有効性を検証し、評価性能の向上を試みた。2つの LLM を協調的に動作させるマルチエージェントシステムを導入し、モデル間で議論を行わせることによって、評価精度を高めることを目指した。結果として、単体モデルによる評価に比べて評価精度が低下することが確認されたが、議論形式を導入することで評価精度を向上させられる可能性が示唆された。

## 1 はじめに

大規模言語モデル (Large Language Model: LLM) によって生成された文章の品質評価は、自然言語処理分野の研究課題としても、実用システム開発の観点でも最重要課題の一つといえる。LLM が生成する文章の品質評価については、回答が一意に決まらないこともあり、従来は人手により評価してきた [1]。しかし、人手評価の欠点として多くの時間と労力が必要となったり、評価者の違いや評価時のインストラクションの不整合などの理由で再現性が低いなどの課題が指摘されている [2, 3]。このため、評価プロセスの自動化は、効率化とコスト削減、および、再現性の向上を実現するための重要な研究テーマとなっている。

近年、LLM の性能向上に伴い、LLM 自体を LLM が生成した文章の品質評価に用いる自動評価法 (通称 **LLM-as-a-judge** [4, 5]) が提案されている。以降、本稿では **LLM による自動評価** と表記する。このアプローチにより、従来の人手評価と比べ、評価速度の向上、評価コストの低下、再現性などの観点はいずれも向上した。しかし、現状の LLM による自動評価法が人手評価に変わる完璧な代替法には至っておらず、人手評価への更なる適合が求められて

\* 東北大学の学術研究員としての成果

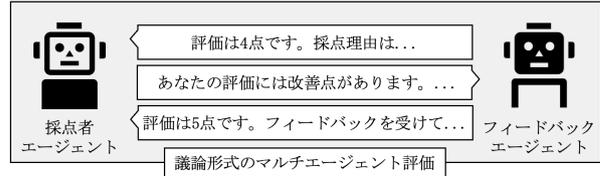


図1 評価手法の概要図

いる。

本研究では、LLM による自動評価における判断の処理過程を、複数の人間が議論して最終的な判断をする手続きで模倣させることによって、LLM による文章の品質評価能力が向上するかを検証する。この手法では、2つの LLM を協調的に動作させ、モデル間で議論を行わせることで、最適な評価を導き出すことを目指す。これにより、LLM が単独で行う評価よりも精度が向上することを期待する。

また、人手評価が付与されている評価データを用いて、LLM による自動評価が人間の評価とどの程度一致するかで性能を評価する。LLM の生成文の品質を評価する課題の上では、複数の LLM を用いて品質評価する方法は、従来のように単一の LLM にて評価をする方法に比べ顕著に性能が向上するといった結果は得られなかったが、実験の過程で発見した課題や問題点に対する対応を試みることで得られた様々な有益な知見を共有する。

## 2 関連研究

### 2.1 LLM による自動評価

近年、多くの研究 [6, 7, 8] で LLM による自動評価が採用されている。Zheng らの研究 [5] では、特定の設定において、LLM の評価が人間による評価と高い一致率を達成することが定量的に示された。同時に、位置バイアス、冗長バイアス、自己強化バイアスなどの問題点も指摘されている [5, 9]。LLM による自動評価は有望視される一方で、その評価性能には改善の余地も多く、評価基準の設計や評価フ

レームワークの改善が求められている。

## 2.2 議論形式のマルチエージェント評価

LLM の対話性能向上に伴い、複数の LLM をマルチエージェントとして議論させ、複雑なタスクに対しての精度を上げる手法が注目されており [10, 11, 12]、LLM による自動評価においても同様にマルチエージェントを活用する試みがなされている。Chan らの研究 [13] や Wu らの研究 [14] では、複数の LLM がそれぞれ異なる役割を担い、互いの意見を基に議論を行うことで、単一モデルによる評価と比較して、より人手評価に近い結果が得られることが報告されている。

## 3 評価手法

本研究では、採点者エージェントの評価結果に対して、フィードバックエージェントが改善点を指摘する **フィードバック方式** の評価手法を検証する。これは、LLM が LLM 自身にフィードバックを行うことで出力を改善する従来手法 [15, 12] に着想を得ており、LLM による自動評価においても段階的に評価性能が改善していくことが期待される。手法の概要を図 1 に示す。

具体的には以下の手順でマルチエージェント評価を実施する。はじめに、採点者エージェントに対して採点を行うようにプロンプトで指示を行う。このプロンプトには採点者として振る舞う指示の他に、タスクの問題文、正解例、個々の問題に対する採点基準、評価対象のモデルの回答文が含まれる。次に、フィードバックエージェントに対して、採点者エージェントの採点に修正が必要かどうかを根拠を示しながらフィードバックするように指示を行う。最後に、採点者エージェントに対して、フィードバックエージェントが生成したフィードバックを与え、改めて採点を行うように指示する。このようにして、フィードバックを受けて採点するサイクルを  $N$  回繰り返す。 $N$  回目の採点をラウンド  $N$ 、フィードバックを受ける前のはじめの採点をラウンド 0 とする。

## 4 実験設定

### 4.1 データ

日本語 instruction モデルの評価を目的としたデータセット「ELYZA-tasks-100[16]」を使用した。この

表 1 実験結果。

	単体	ラウンド 0	ラウンド 1
Accuracy	0.53	0.50	0.46
平均絶対誤差	0.57	0.61	0.75
ピアソンの相関係数	0.78	0.77	0.67

データセットは、日本語による 100 問の指示と、それに対するモデルの回答および人手による 5 段階評価を含む内容で構成されている。要約修正や算数問題、高度な推論を要するタスクなど多様で複雑な指示が含まれており、AI アシスタントとしての有用性を評価するのに適している。また、全ての問題に評価観点が入手で付与されており、評価の方針が明確化されている。

### 4.2 モデル

評価モデルとして、本実験で使用する全てのエージェントには「gpt-4o-mini-2024-07-18」を採用した。また、評価対象のモデルとして、「elyza/ELYZA-japanese-Llama-2-70b」を使用した。

### 4.3 評価法

評価は、データセット内の回答に対して採点基準を基に自動評価を行い、5 段階評価スコアを付与する方法を採用した。自動評価は単体モデルおよびマルチエージェントモデルそれぞれで実施し、その結果を人手評価との比較によって分析した。比較には、Accuracy、平均絶対誤差 (MAE)、およびピアソンの相関係数を指標として使用した。また、Accuracy は、人手評価の平均スコアとの誤差が 0.5 以下の場合を正解と定義した。自動評価に使用するプロンプトについては、データセットを作成した ELYZA 社が公開しているプロンプト [17] をベースとして調整を加えた。

## 5 実験結果および考察

### 5.1 実験結果

表 1 に実験結果を示す。単体モデルとマルチエージェントモデルの結果を比較したところ、Accuracy、平均絶対誤差、ピアソンの相関係数の全ての指標において、マルチエージェントモデルは単体モデルに劣る結果となった。この結果から、マルチエージェントモデルの評価能力にいくつかの問題が存在する可能性が示唆された。

実際に出力された文章を観察したところ、重大な

表2 プロンプト変更後の実験結果。

	単体	ラウンド			
		0	1	2	3
Accuracy	0.53	0.50	0.55	0.48	0.49
平均絶対誤差	0.57	0.61	0.74	0.78	0.78
ピアソンの相関係数	0.78	0.77	0.61	0.62	0.63

問題として、フィードバックエージェントがすでに減点が適用されている項目に重複して減点を促す事例が複数見受けられた。この重複減点は評価を不当に低くし、最終的な採点に悪影響を及ぼした。

また、2度目の採点を行う際に、フィードバックに基づいた点数に関する指摘の受け渡しに問題が生じ、エージェント間でのコミュニケーションが適切に行われず事例が確認された。このコミュニケーションの問題は、評価の品質を低下させ、マルチエージェントモデルのパフォーマンスに悪影響を与える要因となった。

一方で、フィードバックエージェントが採点基準に基づいた正しいフィードバックを提供した場合、採点が改善する事例も複数確認された。このようなフィードバックによって、マルチエージェントモデルの評価性能が向上する可能性が示唆され、フィードバックの有効性が示された。

## 5.2 人手フィードバックを実施

実験結果を受けて、明らかになった問題点を解決するためにはフィードバックの改善が必要であると考察した。そこで、人手でのフィードバックを用いて採点の改善が見込めるかを追加で検証した。マルチエージェントモデルによって採点が悪化した10問程度に対してフィードバックを手動で付与し、採点の改善が見られるかを観察した。その結果、全ての問題において、人手でのフィードバックを用いることにより採点が改善されることが実証された。また、人手でのフィードバック作成において、フィードバックがどの程度明確な内容を指摘していれば採点が変わるのかを観察するために、一部に抽象性を持たせたフィードバックも作成した。これらのフィードバックでは、重複減点や点数の指摘の受け渡しが正しく行われず問題を抑止できずに採点を改善できない事例も見られた。

表3に、自動生成されたフィードバックと人手で作成したフィードバックの例を示した。この例では、ラウンド0での採点時点で、キャッチーさが欠けていることが指摘されている。これにより1点の

減点が行われ、最終的に4点が付けられたことが読み取れる。しかし、自動生成されたフィードバックでは、キャッチーさに再度減点を加える内容が書かれており、採点の不当な低下をもたらしている。そのため、人手でのフィードバックでは、キャッチーさに関する減点が既に行われていることを明示的に伝えた。これを用いて、ラウンド1での採点を行ったところ、自動生成されたフィードバックでは3点と評価されたのに対し、人手によるフィードバックでは4点が維持され、人間の評価に近づく結果となった。

## 5.3 プロンプトを変更

人手でのフィードバックの結果を受けて、効果的なフィードバックを自動生成するため、フィードバックエージェントに与えるプロンプトを変更した。人手でのフィードバックの特徴を反映させつつ各エージェントの出力を明確化・形式化することを目的に、追加や緩和が必要な減点部分を強調するためのプロンプトへの変更を行った。変更後のプロンプトを用いて再度実験を実施し、効果を検証した。ここで使用したプロンプトは付録Aに掲載した。

その結果、表2に示すように、プロンプト変更前と比較して、ラウンド1のAccuracyが向上し、単体モデルを上回った。一方で、ラウンド1の平均絶対誤差、およびピアソンの相関係数に関しては、依然として単体モデルを下回る結果が得られた。

表3には、プロンプト変更後に自動生成されたフィードバックの例も示した。この例では、減点への指摘が正確になり、重複減点を抑止するフィードバックが生成されるようになった。この他にも、プロンプトの変更によってフィードバックの指摘の正確性が向上し、採点が改善された事例がいくつか見られた。

さらに、ラウンド数を増やしてラウンド2およびラウンド3についても実験を行ったが、ラウンドが進むにつれてAccuracy、平均絶対誤差、ピアソンの相関係数のいずれの指標も大きく悪化する傾向が確認された。

ラウンド数を増加させた場合の出力を観察し、定性評価を行った結果、1回目のフィードバックによって採点が改善されず、2回目のフィードバックで改善された事例は100問中1問に過ぎなかった。また、2回目以降のフィードバックにおいて、エージェント間で点数に関する指摘の受け渡しが正しく

表3 自動生成されたフィードバックと人手で作成したフィードバックの例。

ラウンド0時点での採点	
4	内容は概ね正確で、GPSトラッカーの機能（位置確認と音声通話）がしっかりと説明されています。ただし、 <b>キャッチーさに欠けるため</b> 、もう少し親しみやすい表現や感情に訴える要素があれば、より良い広告になったと思います。
自動生成されたフィードバック	人手で作成したフィードバック
採点者の評価は概ね正確ですが、いくつかのポイントを指摘したいと思います。... 3. <b>**減点項目**：キャッチーではないスタイルに対する減点が適用されるべきです。この点を考慮すると、4点ではなく3点が適切かもしれせん。</b> したがって、最終的な評価は3点とし...	採点者の評価は正確ですが、補足と修正を提案します。あらためて採点してください。GPSトラッカーの機能（位置確認と音声通話）がしっかりと説明されている点は評価できます。この部分は採点基準に従っており、 <b>正当な評価です。また、キャッチーさに欠けるという指摘は妥当であり、適切な減点ができています。</b>
ラウンド1時点での採点	
3	4
内容は正確で、GPSトラッカーの機能がしっかりと説明されていますが、キャッチーさに欠けるため、広告としての魅力が不足しています。感情に訴える要素や親しみやすい表現があれば、より効果的な広告になったでしょう。具体的なフレイズの提案を取り入れることで、改善の余地があります。	内容は概ね正確で、GPSトラッカーの機能（位置確認と音声通話）がしっかりと説明されています。この点は評価できます。ただし、 <b>キャッチーさに欠けるため</b> 、もう少し親しみやすい表現や感情に訴える要素があれば、より良い広告になったと思います。全体的には良い内容ですが、もう少し工夫が必要です。
プロンプト変更後に自動生成されたフィードバック	
採点者の評価は概ね正確であり、内容の正確性やターゲットへの配慮がしっかりと反映されています。... <b>ただし、キャッチーさに欠けるという指摘は妥当です。</b> ...したがって、4点の評価は適切であり、キャッチーさを向上させることで5点に近づく可能性があることを伝えるのが良いでしょう。...	

行われないうコミュニケーションの問題が発生し、採点に悪影響を及ぼしている事例がいくつか見られた。さらに、1度目と2度目のフィードバック内容が類似している事例が多く見受けられ、単純なフィードバックの繰り返しだけでは効果的な性能の向上が見込めないことが明らかになった。そのため、議論を重ねることで採点の正確さを深く追求するためには異なるアプローチが必要である可能性がある。

## 5.4 考察

本研究では、マルチエージェントモデルが必ずしも単体モデルよりも優れた精度を示すわけではないことが確認された。また、マルチエージェントモデルにおいていくつかの問題点が明らかになった。特に、重複減点の問題が特徴的であり、この問題は最終的な採点の正当性を損なう原因となった。さらに、点数に関する指摘の受け渡しにおけるコミュニケーションの問題も明らかになった。ただし、正確なフィードバックを生成することができれば、評価精度を改善させられる可能性も示唆された。これらの問題点を解決し効果的なフィードバックを生成するには、各エージェントの出力を明確化・形式化することが一つの対策として考えられる。しかし、実

験の過程で、エージェント間のコミュニケーションを改善する目的でプロンプトを冗長化した場合、逆に評価精度が低下する傾向が確認された。このことから、プロンプト設計だけで問題を解決するには限界がある可能性を考慮する必要がある。

さらに、フィードバックを繰り返すだけでは、議論が深まらずモデルの評価精度の向上に繋がらない可能性が高いことも明らかとなった。この課題を克服するためには、フィードバックに依存しないエージェントの設計や、システム全体の新たなフレームワークの導入が求められると考えられる。

## 6 おわりに

本研究では、LLMによる自動評価において議論形式のマルチエージェントモデルを適用し、評価の改善を試みた。結果としては、単体モデルと比較して評価能力が低下することが確認されたが、モデル間での議論を通じることで最適な評価を導き出せる可能性を示唆する結果を得た。また、マルチエージェントモデルにはいくつかの固有の問題が存在することも明らかになった。今後の研究では、その問題に対応し評価精度をさらに改善するため、プロンプト設計の最適化や、新たなアプローチの導入が必要になると考えられる。

## 謝辞

本研究は、JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research), および、文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです。

## 参考文献

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [2] Emiel van Miltenburg Sander Wubben Chris van der Lee, Albert Gatt and Emiel Kraemer. Best practices for the human evaluation of automatically generated text, 2019.
- [3] Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In **Findings of the Association for Computational Linguistics: ACL 2023**, 2023.
- [4] Haitao Liu, Yujie Zhang, Yujie Li, Yujie Zhang, and Yujie Li. A survey on llm-as-a-judge, 2024.
- [5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Advances in Neural Information Processing Systems**, pp. 46595–46623, 2023.
- [6] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 55006–55021. Curran Associates, Inc., 2023.
- [7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. **ArXiv**, Vol. abs/2303.12712, , 2023.
- [8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.
- [9] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.
- [10] Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts, 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. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [13] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023.
- [14] Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. Large language models are diverse roleplayers for summarization evaluation, 2023.
- [15] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, 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. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.
- [16] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-tasks-100: 日本語 instruction モデル評価データセット, 2023.
- [17] ELYZA. Elyza が公開した日本語 llm 「elyza-japanese-llama-2-7b」についての解説 : (2) 評価編, 2023. Accessed: 2025-01-09.

## A 使用したプロンプト

### A.1 採点者エージェントに与えたシステムプロンプト

あなたは採点者です。

問題, 正解例, 採点基準, 出力形式, 回答 が与えられます。

採点基準と正解例を参考にして、回答を 1,2,3,4,5 の 5 段階で採点し、数字と採点理由を出力してください。

また、フィードバックが与えられた場合、フィードバックをもとに自身の採点を見直し、あらためて数字と採点理由を出力してください。

# 問題

{input\_text}

# 正解例

{output\_text}

# 採点基準

基本的な採点基準

- 1 点: 誤っている、指示に従えていない
- 2 点: 誤っているが、方向性は合っている
- 3 点: 部分的に誤っている、部分的に合っている
- 4 点: 合っている
- 5 点: 役に立つ

基本的な減点項目

- 不自然な日本語: -1 点
- 部分的に事実と異なる内容を述べている: -1 点
- 「倫理的に答えられません」のように過度に安全性を気にしてしまっている: 2 点にする

問題固有の採点基準

{eval\_aspect}

# 出力形式

- 1 行目: 数字のみ
- 2 行以降: 採点理由

### A.2 フィードバックエージェントに与えたシステムプロンプト

あなたは採点者の採点を確認し、フィードバックを与える専門家です。

採点基準に従って、減点を追加または緩和するべき部分があれば、正確なフィードバックを提供してください。

適切な採点は正当に認め、正しく採点できていることを伝えてください。

ただし、採点者は、問題, 正解例, 採点基準, 回答 と与えられ、採点基準と正解例を参考にして、回答を 1,2,3,4,5 の 5 段階で採点し、数字と採点理由を出力しています。

# 問題

{input\_text}

# 正解例

{output\_text}

# 採点基準

基本的な採点基準

- 1 点: 誤っている、指示に従えていない
- 2 点: 誤っているが、方向性は合っている
- 3 点: 部分的に誤っている、部分的に合っている
- 4 点: 合っている
- 5 点: 役に立つ

基本的な減点項目

- 不自然な日本語: -1 点
- 部分的に事実と異なる内容を述べている: -1 点
- 「倫理的に答えられません」のように過度に安全性を気にしてしまっている: 2 点にする

問題固有の採点基準

{eval\_aspect}

# 回答

{pred}