評価結果の予測確率を用いた LLM による LLM の相対評価

森田隼功 [1]　大林弘明 [2]　田村晃裕 [1]　濱田充男 [2]　加藤恒夫 [1]
[1] 同志社大学大学院　[2] トランスコスモス株式会社
[1]{ctwj0130@mail4,aktamura@mail,tsukato@mail}.doshisha.ac.jp
[2]{oobayashi.hiroaki,hamada.mitsuo}@trans-cosmos.co.jp

概要

本論文では，評価結果の予測確率を用いる大規模言語モデル（LLM）による LLM の新たな相対評価手法を提案する．従来の相対評価では，評価対象 LLM の提示順を入れ替えた際に評価結果が変動する位置バイアスが生じた際，最終的な評価結果を適切に定める手法が確立されていない．そこで，本研究では，位置バイアスが生じた際に，提示順入れ替え前後の評価結果の予測確率の平均が最も高い結果を最終的な評価結果とする手法を提案する．そして，本提案手法を文章作成等のビジネスタスクを対象に評価し，有効性を確認した．

1　はじめに

近年，大規模言語モデル（LLM）の研究が盛んに行われており，その中で，LLM の性能評価に関する研究も多数行われている．LLM の文章生成タスクに対する性能評価として，高性能な LLM（評価役 LLM）を用いて LLM の性能評価を行う手法が提案されている [1, 2, 3]．Zheng ら [1] は，評価役 LLM として GPT-4[4] を用いた場合，文章生成タスクにおいて，人手評価と自動評価の相関は人手評価同士の相関と同等レベルであり，人手評価を LLM で代替できる可能性を示唆している．

既存の LLM による LLM の評価手法は，評価対象の各 LLM の生成文を独立に採点する絶対評価と，2 つの LLM の生成文同士を比較して優劣を評価する相対評価に大別される．本研究では，Chen ら [5] のように相対評価に焦点を当てる．相対評価の問題点として，評価対象の 2 つの LLM の提示順を入れ替えて評価すると，評価結果が変動するという点（位置バイアス）が挙げられる [1]．森田ら [6] は，質問と参照回答から自動生成した評価観点を用いることで位置バイアスの発生を低減させたが，依然として位置バイアスは発生している．位置バイアスが発生した際，Zheng ら [1] は，最終評価結果を「引き分け」としている．また，森田ら [6] は，その自動評価は「適切でない」として評価手法を評価する際には誤りとしている．このように，これまでは，位置バイアス発生時の最終的な評価結果はヒューリスティックに決められている．

そこで，本研究では，評価役 LLM が予測した評価結果の予測確率を用いて最終的な評価結果を定める手法を提案する．具体的には，提示順の入れ替え前後で，評価結果を示すアルファベット token（A：LLM A の勝ち，B：LLM B の勝ち，C：引き分け) の予測確率（すなわち，その評価に対する確信度）を取得し，提示順入れ替え前後の予測確率の平均が最も高いアルファベットを最終的な評価結果として採用する．この提案手法により，位置バイアスが生じた場合でも，出力された評価結果の予測確率を考慮した上で評価結果を一意に定めることができる．

提案手法の有効性を検証するため，文章作成，校正，要約タスクで構成される合計 69 問の質問からなるビジネスタスクを対象にした評価実験を行った．その結果，提案手法は，従来のヒューリスティックに最終評価結果を定めるベースライン手法よりも人手評価に近い結果を得ることができた．

2　関連研究

本節では，LLM による LLM の性能評価の関連研究を概説する．LLM による LLM の性能評価の代表的な英語のベンチマークとしては，MT-Bench[1] や AlpacaEval[3] 等が挙げられる．また，日本語のベンチマークとしては，Japanese Vicuna QA Benchmark[2] や Rakuda Benchmark[1)] 等がある．MT-Bench は，文章

1）https://github.com/yuzu-ai/japanese-llm-ranking/tree/main

作成やロールプレイ等の 8 タスクで構成される合計 80 問のマルチターンの質問からなるデータセットを用いて評価を行い，AlpacaEval は，文章作成をはじめとする多様なタスクで構成される 805 問のデータセットを用いて評価を行う．また，Japanese Vicuna QA Benchmark は，常識や数学等の 8 タスクで構成される合計 80 問の質問からなるデータセットを用いて評価を行い，Rakuda Benchmark は，歴史，社会，政治，地理の 4 タスクで構成される合計 40 問のデータセットを用いて評価を行う．

LLM を用いた評価手法は，評価対象の各 LLM の生成文を独立に採点する絶対評価と，2 つの LLM の生成文同士を比較して優劣を評価する相対評価に分けられる．本研究では相対評価に焦点を当てるため，以降では，相対評価の関連研究について説明する．なお，従来の相対評価では，参照回答を使用するものと使用しないものが存在するが，本研究では，森田ら [6] に倣い，参照回答を使用した相対評価を対象とする．

2.1 LLM による LLM の相対評価

MT-Bench 及び Japanese Vicuna QA Benchmark の相対評価では，評価役の LLM を用いて，比較対象の 2 つの LLM（LLM A と LLM B）の生成文について，どちらが優れているか，もしくは引き分けであるかを評価理由とともに出力する．具体的には，質問，参照回答，LLM A の回答，LLM B の回答を評価役 LLM に入力として与え，相対評価用のプロンプトによって評価を行う．この相対評価において，2 つの LLM の回答の提示順を入れ替えて評価を行うと，入れ替え前後で評価結果が変動する位置バイアスが存在することが知られている．Zheng ら [1] では，評価役 LLM として GPT-4 を用いて相対評価を行った際，35%の割合で入れ替え前後で評価結果が変わったと報告されている．なお，Zheng ら [1] は，位置バイアスが発生した際，最終的な評価結果を「引き分け」としている．

森田ら [6] は，LLM を用いて質問と参照回答を基に評価観点を自動で作成し，自動作成した評価観点を指定して評価役 LLM で評価対象 LLM を相対評価する手法を提案した．この手法により，提示順を入れ替えて相対評価する際，統一した視点での評価が可能になり，位置バイアスの低減及び人手評価との一致率向上を実現した．なお，森田ら [6] は，位置バイアスが発生した場合，その評価を「適切でない」と判断し，人手評価との一致率を算出する際，人手評価と一致していないという計算を行なっている．

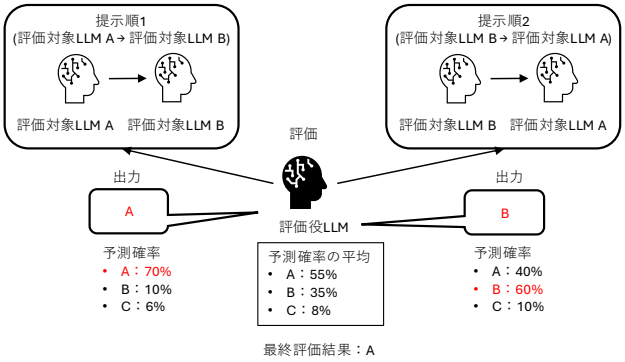

図 1　提案手法の概要

このように，従来の LLM による LLM の相対評価では，位置バイアスが発生した場合に，最終的な自動評価結果をヒューリスティックに定めている．そこで，位置バイアスが発生した際に，最終評価結果をより適切に定める手法が必要であると考える．

3 提案手法

本節では，2 節で指摘した問題点を解消するため，評価結果を示すアルファベットの予測確率をその評価に対する確信度とみなし，提示順入れ替え前後の予測確率の平均が最も高いアルファベットを最終評価結果とする手法を提案する．予測確率の取得には，OpenAI 社の API で各 token の予測確率を取得する logprobs 機能を活用し，得られた logprobs の値に指数関数を適用した値（exp(logprobs)）を予測確率として用いる．なお，logprobs は各位置において上位 20 個の token に対する予測確率しか取得できないため，評価結果を示す位置において予測確率上位 20 個の token に A，B，C が含まれない場合，そのアルファベットの予測確率は 0 とする．

提案手法の概要を図 1 に示す．従来手法では，提示順の入れ替え前後に出力されるアルファベット（図 1 では，提示順 1 が A，提示順 2 が B）に基づき最終評価結果を定める．一方，提案手法では以下の手順で最終評価結果を決定する．

1. 2 つの評価対象 LLM の提示順を入れ替えて評価役 LLM に評価させる．その際，評価結果を表す token（A，B，C）の予測確率を取得する．
2. 取得した予測確率をアルファベットごとに提示順入れ替え前後で平均をとる．
3. 平均値が最も高いアルファベットを最終評価結果とする．

4 実験

4.1 実験設定

実験データ 本実験では，企業で実際に使用された Azure OpenAI Service[2] に与えられた日本語の質問 69 問（文章作成：23 問，校正：21 問，要約：25 問）において提案手法の有効性を検証する．本実験で使用した質問例は付録 C を示す．各質問に対する参照回答は人手によって作成した．

本実験では，Azure OpenAI Service[3] のモデル ID:gpt-35-turbo (0613), ELYZA-japanese-Llama-2-7b-fast-instruct[4]，shisa-gamma-7b-v1[5] の 3 つの LLM の性能を評価する．また，評価観点の作成及び評価役の LLM には OpenAI 社の gpt-4o-2024-11-20 を使用する．なお，ハイパーパラメータ temperature は 0 とし，seed 値を 1 に固定した．

評価プロンプト 本実験では，森田ら [6] の相対評価プロンプトを一部改変したプロンプトを評価役 LLM に与えて評価を行う．具体的には，森田ら [6] の評価プロンプトでは，評価結果を「○○という理由で [[A]]」のように出力させているが，本実験では，評価結果の予測確率に評価結果より前の出力内容が影響しないようにするため，理由と記号 [[]] を削除し，評価結果のみを出力させるプロンプトを用いる．また，比較のため，評価理由の後で評価結果を出力させるプロンプトも用いる．これら 2 種類のプロンプトに対して，LLM で自動作成した評価観点を指定して評価するものと評価観点を用いないものをそれぞれ用いる．まとめると，使用する評価プロンプトは以下の 4 つである．本実験で使用した評価観点作成プロンプトと相対評価プロンプトは付録 B と付録 A にそれぞれ示す．

1. PPT(観点無; 結果のみ)：評価理由は出力させず，評価結果のみをアルファベット（A: LLM A の勝ち，B: LLM B の勝ち，C: 引き分け）で出力させる．評価観点は使用しない．評価役 LLM の出力例は「A」.
2. PPT(観点有; 結果のみ)：PPT(観点無; 結果のみ) を評価観点ありで実施する．評価役 LLM の出力例は「A」.
3. PPT(観点無; 理由→結果)：評価理由の後，評価結果をアルファベット（A: LLM A の勝ち，B: LLM B の勝ち，C: 引き分け）で出力させる．評価観点は使用しない．評価役 LLM の出力例は「○○という理由で A」.
4. PPT(観点有; 理由→結果)：PPT(観点無; 理由→結果) を評価観点ありで実施する．評価役 LLM の出力例は「○○という理由で A」.

表 1 評価者間の kappa 係数

文章作成	校正	要約	All
0.62	0.63	0.53	0.60

ベースライン評価手法 本実験では，提案手法の有効性を検証するため，前述の評価プロンプトを実行して出力された評価結果そのもの（A，B，C）から最終評価結果を決定するベースライン評価手法の性能も評価する．具体的には，位置バイアスが発生した際に，Zheng ら [1] や森田ら [6] のように最終評価結果を定める手法をベースライン手法とする．

1. BS1（森田ら [6]）：位置バイアスが生じた場合，適切な評価が実施されていないとみなす．
2. BS2（Zheng ら [1]）：位置バイアスが生じた場合，最終評価結果を引き分けとする．

評価指標 本実験では，3 節の提案評価手法と本節で前述したベースライン評価手法の性能を比較する．各自動評価手法の性能を評価する評価指標は，人手評価結果と各自動評価手法による評価結果が一致する割合（Concordance Rate）を使用する．人手評価結果は，著者 3 名が評価対象 LLM の全ペアで回答を比較し，各回答ペアに対して参照回答をもとに相対評価を実施した結果を使用する．3 名の評価結果の一致度合いを表す Fleiss の kappa 係数 [7] を表 1 に示す．

Concordance Rate は，人手評価付与者 3 名それぞれの人手評価結果に対して以下の式 (1) で算出し，その平均を取った．

$$\text{Concordance Rate} = \frac{1}{K \cdot S} \sum_{k=1}^{K} \sum_{s=1}^{S} \delta(y_{k,s}, \hat{y}_{k,s}) \quad (1)$$

ここで，K は質問数，S は比較対象の LLM の全組み合わせ数（つまり，${}_3C_2$），$\delta(y_{k,s}, \hat{y}_{k,s})$ は人手評価結果 $y_{k,s}$ と自動評価結果 $\hat{y}_{k,s}$ が一致する場合に 1，一致しない場合に 0 を返す関数である．この Concordance Rate は，値が大きいほど自動評価が人

2) https://azure.microsoft.com/ja-jp/products/ai-services/openai-service
3) https://learn.microsoft.com/ja-jp/azure/ai-services/openai/concepts/models
4) https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b
5) https://huggingface.co/augmxnt/shisa-gamma-7b-v1

表 2 各評価手法の Concordance Rate (%)

評価手法	文章作成	校正	要約	All
BS1(観点無; 結果のみ)	53.1	54.0	56.9	54.8
BS2(観点無; 結果のみ)	57.0	57.7	**68.4**	61.4
提案 (観点無; 結果のみ)	**71.0**	**61.4**	**68.4**	**67.2**
BS1(観点有; 結果のみ)	56.0	51.3	46.7	51.2
BS2(観点有; 結果のみ)	58.5	56.6	50.2	54.9
提案 (観点有; 結果のみ)	**72.0**	**63.5**	**59.6**	**64.9**
BS1(観点無; 理由→結果)	69.6	60.3	67.1	65.9
BS2(観点無; 理由→結果)	70.5	65.1	68.4	68.1
提案 (観点無; 理由→結果)	**74.9**	**69.8**	**71.1**	**70.5**
BS1(観点有; 理由→結果)	73.9	59.3	65.3	66.3
BS2(観点有; 理由→結果)	74.9	63.0	68.0	68.8
提案 (観点有; 理由→結果)	**77.3**	**65.1**	**70.7**	**71.2**

手評価に近いことを表す.

4.2 実験結果

実験結果を表 2 に示す．表 2 では，評価プロンプト PPT(X) を用いるベースライン手法を BS1(X) あるいは BS2(X)，提案手法を提案 (X) と表記している．なお，X は「観点無; 結果のみ」,「観点有; 結果のみ」,「観点無; 理由→結果」,「観点有; 理由→結果」のいずれかである．また，各評価プロンプトにおいて，最も性能が良い手法を太字で表している．

表 2 より，要約タスクにおいては「提案 (観点無; 結果のみ)」と「BS2(観点無; 結果のみ)」は同等の性能であるが，それ以外の全ての条件において，提案手法は一貫してベースライン手法よりも人手評価との一致率が高いことが分かる．これより，4 つの評価プロンプトのいずれを使用しても，提案手法は 2 つのベースライン手法よりも人手評価に近い評価を行えることを確認した．

また，評価理由を出力した後に評価結果を出力させるプロンプトを使用した場合，評価結果のみを出力させるプロンプトを使用した場合と比べて，ベースライン手法および提案手法のいずれにおいても人手評価との一致率が高くなることが確認できる．特に「提案手法 (観点有; 理由→結果)」は，全タスク（ALL）で最も高い一致率を達成した．これらの結果から，評価対象の提示順入れ替え前後の評価結果の予測確率を用いて評価結果を一意に定める提案手法によって，人手評価により近い評価を行うことができることを確認した.

表 3 各評価プロンプトを用いた相対評価における頑健性 (%)

プロンプト	文章作成	校正	要約	All
PPT(観点無; 結果のみ)	63.8	77.8	76.0	72.5
PPT(観点有; 結果のみ)	69.6	73.0	73.3	72.0
PPT(観点無; 理由→結果)	85.5	82.5	88.0	85.5
PPT(観点有; 理由→結果)	94.2	82.5	85.3	87.4

5 考察

本節では，表 2 において，評価結果のみを出力させるプロンプトを使用した場合，評価理由を出力した後に評価結果を出力させるプロンプトを使用した場合と比べて，人手評価との一致率が低くなった原因を考察する．表 3 に本実験で使用した 4 つの評価プロンプトにより相対評価を行った際に，位置バイアスが生じなかった割合（頑健性）を示す.

表 3 より，評価結果のみを出力させるプロンプトを使用した場合，評価理由を出力した後に評価結果を出力させるプロンプトを使用した場合と比べて，頑健性が大きく下回っていることが分かる．これより，評価結果のみを出力させるプロンプトを使用したことによる頑健性の低さが，人手評価との一致率の低さに繋がったと考えられる.

また，表 2 より，評価結果のみを出力させるプロンプトを使用した場合，評価理由を出力した後に評価結果を出力させるプロンプトを使用した場合と比べて，ベースライン手法と提案手法の差が大きいことが分かる．これより，頑健性が低い（つまり，位置バイアスが多く発生している）方が，本提案手法がより有効であることがわかる.

6 おわりに

本研究では，LLM による LLM の相対評価において，評価対象の提示順入れ替え前後の評価結果の予測確率を用いて，最終評価結果を一意に定める手法を提案した．そして，提案手法を文章作成，校正，要約タスクで構成される日本語のビジネスタスクを対象に評価した結果，評価結果（アルファベット）を直接参照して最終評価結果を定める従来手法と比較して，より人手評価に近い結果が得られることを確認した．今後は，MT-Bench 等のベンチマークデータセットにおいても提案手法の有効性を検証したい.

参考文献

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, pp. 46595–46623, 2023.

[2] Yikun Sun, Zhen Wan, Nobuhiro Ueda, Sakiko Yahata, Fei Cheng, Chenhui Chu, and Sadao Kurohashi. Rapidly Developing High-quality Instruction Data and Evaluation Benchmark for Large Language Models with Minimal Human Effort: A Case Study on Japanese. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 13537–13547, 2024.

[3] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval, 2023.

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

[5] Chen Guiming Hardy, Chen Shunian, Liu Ziche, Jiang Feng, and Wang Benyou. Humans or LLMs as the Judge? A Study on Judgement Bias. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 8301–8327, 2024.

[6] 森田隼功, 大林弘明, 田村晃裕, 濱田充男. 自動作成した評価観点を用いた LLM による LLM の参照回答ベース評価. 情報処理学会研究報告, 自然言語処理研究会, Vol. 2024-NL-260, No. 7, pp. 1–14, 2024.

[7] J. L. Fleiss. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Vol. 76, No. 5, pp. 378–382, 1971.

A　ベースライン及び提案手法で使用する相対評価プロンプト

以下に表示されるタスクに対するアシスタントの応答品質を評価してください。
あなたの仕事は、模範回答に近いアシスタントの回答を高く評価することです。
評価の際、以下の評価観点をもとに評価をしてください。　　[評価観点開始]{eval_criteria}[評価観点終わり]
その他気をつける点は以下の通りです。立場の偏りを避け、回答が提示された順番があなたの決定に
影響しないようにしてください。回答の長さが評価に影響しないようにしてください。
特定のアシスタントの名前を好まないこと。できるだけ客観的に評価してください。
評価理由を提供した後、以下の書式にしたがって評価を出力してください：
アシスタント A が良い場合は A、アシスタント B が良い場合は B、同点の場合は C です。
[タスク]{question}　　[模範回答の開始]{reference}[模範回答の終了]
[アシスタント A の答えの開始]{llmA_text}[アシスタント A の答えの終了]
[アシスタント B の答えの開始]{llmB_text}[アシスタント B の答えの終了]

図 2　相対評価プロンプト（PPT(観点有; 理由→結果)）

プロンプト中の{eval_criteria}，{question}，{reference}，{llmA_text}と{llmB_text}は，それぞれ，評価観点，質問，参照回答，比較対象の LLM A と LLM B の回答を表す．図 2 は，PPT(観点有; 理由→結果) の評価プロンプトである．評価観点を指定しない場合は「[評価観点開始]{eval_criteria}[評価観点終わり]」の箇所を消去する．また，評価結果のみ出力する場合は，「評価理由を提供した後、以下の書式にしたがって評価を出力してください」を「以下の書式にしたがって評価結果のみを出力してください」に変更する．

B　評価観点作成プロンプト

公平な判断者として行動し、下記の質問に対する AI アシスタントの応答品質を評価します。
あなたが評価を行う際に必要な評価観点を、下記の質問と模範回答から出力してください。
評価観点は、質問内に条件および出力形式が記載されている場合は、それらを反映すること。
箇条書きで評価観点のみ出力してください。
[質問開始]{question}[質問終了]　　[模範回答開始]{reference}[模範回答終了]

図 3　評価観点を作成するプロンプト

C　実験データ例

表 4　本実験で使用した質問の例

タスク	質問
文章作成	訃報メールへの返信をもらった時の返信例を 300 文字程度で作成して
校正	下記の文章の改善点を指摘してください。 SIM カードのサイズ変更・再発行のお申し込み後、SIM カード変更の移行期間中は通信できない期間が発生することとなり、今回開通がされた状態で発送されることとなり、同じ番号は二つ存在できないため、eSIM の利用ができない除隊となっております。
要約	#指示 下記の文章について、条件に従って要約してください #条件 ・端的に説明してください ・制度について認識の無い人でも理解しやすい文章にしてください #文章 適格請求書（インボイス）とは、売手が買手に対して、正確な適用税率や消費税額等を伝えるものです。具体的には、現行の「区分記載請求書」に「登録番号」、「適用税率」及び「消費税額等」の記載が追加された書類やデータをいいます。インボイス制度とは、＜売手側＞売手である登録事業者は、買手である取引相手（課税事業者）から求められたときは、インボイスを交付しなければなりません（また、交付したインボイスの写しを保存しておく必要があります）。＜買手側＞買手は仕入税額控除の適用を受けるために、原則として、取引相手（売手）である登録事業者から交付を受けたインボイス（※）の保存等が必要となります。 （※）買手は、自らが作成した仕入明細書等のうち、一定の事項（インボイスに記載が必要な事項）が記載され取引相手の確認を受けたものを保存することで、仕入税額控除の適用を受けることもできます。