

LLM の出力結果に対する人間による評価分析と GPT-4 による自動評価との比較分析

関根聡¹ 小島淳嗣² 貞光九月² 北岸郁雄²

¹理化学研究所 AIP ²マネーフォワード

satoshi.sekine@riken.jp

kojima.atsushi@moneyforward.co.jp

概要

本論文では、2つの LLM の出力結果に対して「複数の人間による評価の分析」および「人間の評価と GPT-4 による自動評価結果の比較分析」を報告する。人間による評価は5人で行い、回答が質問に答えているかという関連性、回答内容の正確性、回答の流暢性、情報量を5スケールで評価した上で、出力の優劣を判定した。このデータを概観し、優劣の判定がどのようになされたかについての仮説を立てた。また、この人間による評価結果と GPT-4 による評価結果は大きく食い違っていたが、その原因を突き止めた。一方の LLM は情報量が多いが正確性に欠けるものであり、もう一方が一般的なことしか書かれていないため正確性には問題が少なかったシステムであった。最初のシステムに対し、人間の評価では情報が不正確であると判断されて低い評価を受けたが、GPT-4 は正確性が判断できず「具体性があり情報量が多い」というコメントと共に高い評価を下した。GPT-4 は正確性の判断に弱いということは、評価コメント中に認められたハルシネーションからも観察され、これが判定の食い違いの大きな理由であったことが分かった。これらの分析を基に今後の LLM 開発における課題を考察した。

1. 背景

GPT を始めとする生成 AI (LLM) が、数多くのタスクにおいて非常に高い精度で対話的な言語生成を行い、実応用化が広がるなどの影響を与えている。その開発に取り組む際には、出力の評価が重要であるが、これには困難が伴う。単語表現などの抽出タスクや factoid などの質問応答、選択肢から正解を選ばせる分類問題は自動的な評価が容易であるが、対話型生成 AI の本来の魅力のごく一部しか評価できない。

この問題に対して、生成 AI を利用して評価を行うという考え方があり[1]、そのサービスも展開されている[2]。つまり、質問文と2つの生成 AI の出力を入力とし、どちらの出力の方がより良いものかを生成 AI に判定させて評価を行うというものである。しかしながら、たとえば[1]で報告されている「人間同士の評価一致率と、人間と生成 AI との評価の一致率がほぼ等しい」という報告をそのまま鵜呑みにし、どのような面で評価が一致しているのか、一致していないのか、および、出力のどういった面に注目して人間や生成 AI は評価を行なっているのかといった分析は著者の知る限り、存在しない。本論文では、日本語の生成 AI の出力を GPT-4[3]で評価する Rakuda ベンチマーク[2]の評価タスクを題材に取り、「複数の人間による評価の分析結果」および「人間の評価と GPT-4 の評価結果の比較分析」を行った。

2. 実験セッティング

2.1 2つの生成 AI のシステム

本実験では、2つの生成 AI システムを利用して評価実験を行なった。一つは OpenAI が提供している GPT3.5-turbo-1106[4]である。もう一つは、マネーフォワードが構築した houou-instruction-7b-v2 (以下、houou-v2) である[5][6]。このシステムは、Llama-2[7]をベースに日本語英語の追加事前学習を行なった rinna/yourri-7b[8]を元にし、理研 AIP を中心に構築された”ichikara-instruction (4802)”[9][10]で SFT の学習を行なったシステムである。”ichikara-instruction”の有用性は、以前のバージョン(houou-instruction-7b-v1)でも検証されており、このインストラクションの影響が大きいことは分かっている。

2.2 Rakuda ベンチマーク

Rakuda ベンチマークは、日本語生成 AI の出力を評価するために作られたベンチマークで、日本の地理、歴史、政治、経済に関する各 10 問、全体で 40 問の質問からなる。それぞれの分野の知識を問いながら説明も求める形式の問題が多く、分野は偏るものの単純なクイズや質問応答と比較すると、生成 AI の特徴を捉えた評価セットだと言える。図 1 に質問の例を示す。また、Rakuda ベンチマークは GPT-4 により評価を行なっている。この際に利用されるプロンプトを付録 A.1 に載せる。

日本の三大都市圏（首都圏、近畿圏、中部圏）の各中心都市と、それぞれの都市圏を構成する主要な都道府県を列挙してください。
日本の都道府県の中で人口が最も少ないのはどこですか？その理由となる地理的な特性について説明してください。
日本の主要な海運路を一つ挙げてください。また、その海運路が日本の物流や交通にどのように寄与しているかについて説明してください。
熊野古道は、日本のどの地方に位置し、主に何のために使われていましたか
日本の三権分立について説明し、それぞれの権力がどのように機能しているか述べてください。
戦後の日本政治において最も影響力のあった政治家を一人挙げ、その貢献について詳しく述べてください。

図 1. Rakuda ベンチマークの質問例

2.3 評価アノテーター

評価実験には 5 名のアノテーターが参加した。大文学部生 4 名と元国語講師の主婦 1 名である。これまで、「ichikara-instruction」の構築に参加しており、しっかりとした言語知識を持ち、信頼できるデータを提供してくれている方々である。

3. 人間による評価結果の分析

本節では、5 名のアノテーターによる評価結果と、アノテーターの判定の一致度、人間の優劣判定導出に関する仮説など、人間の評価の分析について報告する。

3.1 評価の方法

アノテーターには、表 1 に示した 5 つの項目に関する判定と自由記述のコメントを各データ（質問と回答のペア）に対して記載するよう依頼した。回答はどのシステムが出力したものか分からないよう、ランダムに「A」、「B」と記号化されている。

評価項目	内容	出力
関連性	質問に対する回答か	1~5 (5が最善)
正確性	情報の正確性	
流暢性	言語の正しさ、読みやすさ	
情報量	詳細で有用な情報か	
優劣判定	どちらの出力がより良いか	A/B

表 1. 評価項目

アノテーターによる評価結果の例を付録 A.2 に載せる。そこにあるように、アノテーターは正確性などについても細かく調べ、1 つの質問に対して平均 15 分で評価を行なった。

3.2 評価結果

5 人による 5 つのスコアの平均値および判定数の平均値は表 2 の通りであった。「H」は houou、「G」は GPT-3.5、「同」は同等と判定された数。セル内の上側は平均スコア、下側はそのシステムが勝った数。この結果から、「関連性」は過半数が同等であるが、それ以外はやや GPT の方が優れている。実際、両方共、質問に対して妥当な回答をしている場合が多い。「流暢性」はほぼ同程度である。「正確性」は GPT-3.5 が非常に優れていて、「情報量」は houou の方が非常に優れている。全体的な「優劣」では、GPT が優れているという結果であった。

	関連性	正確性	流暢性	情報量	優劣
H	3.9 4.0	2.9 4.2	4.1 13.0	3.6 19.2	12.8
G	4.2 11.6	3.7 23.2	4.1 11.4	3.30 7.0	22.4
同	24.4	12.6	15.6	13.8	4.8

表 2. 人間による評価結果の平均値

3.3 統合的な優劣判定

表 3 で示された平均値ではなく、GPT-4 による自動判定結果と比較するためにも、個々の質問に対して、5 名のアノテーターの結果をまとめた統合的な優劣判定を作成した。まず、いずれかのシステムを「優」と判定した人が過半数いて、「劣」と判定した人が 1 名以下の場合、「優」と判定されたシステムを勝者とする。2 つのシステムで票が割れた場合

(つまり「優」と「劣」が3対2、もしくは2対3となった場合)には第一著者がデータおよびアノテーターのコメントを見て判定したものを統合結果とした。そのような場合は6件あり、そのうち3件は多数決そのまま、残りの3件は多数側のアノテーターの判定に疑問があり「同等」となった。3人以上が同じシステムを「優」と判定しなかった場合には、2つのシステムは同等とした。この結果を表3に示す。平均値の結果(表2)とほぼ同様である。

システム	勝率 (数)
houou-7b-instruction-v2	27.5 (11)
GPT-3.5-turbo-1106	55.0 (22)
引き分け	17.5 (7)

表3. 人間による統合的な評価結果

3.4 アノテーター同士の一致度

アノテーター同士の評価の一致度は[1]の論文同様、異なるシステムを2人の評価者が選んだ場合には0、同じものを選んだ場合には1、同等が含まれる場合には0.5として、平均値を計算すると66.75%となった。[1]の論文の概要では一致度が80%以上と述べられているが、この論文のAppendix D.3を読むと、この数字は「同等を除外した場合」の値であり、本論文と同じ評価では63%と報告されており、ほぼ同じ結果である。統合判定とそれぞれのアノテーターの判定との一致度は79.25%であり、これも[1]の論文の81%と同程度の数字となっている。

3.5 優劣判定導出に関する仮説

各アノテーターがシステムの優劣をどのように導出したかを考察する。各アノテーターは4つの評価項目とコメントを記載し、優劣判定を行っている。この評価項目が優劣の判定に関係していると考えられるが、コメントも参考にしながら、後述のようなアルゴリズムを考案した。このアルゴリズムに基づく、4例を除いて全ての優劣判定が説明できることが分かった(4例は全て「同等」が絡んでいる)。このアルゴリズムは直感的でもあり、自然言語で表現すると下記のように表現できる。

<アルゴリズム>

1. 関連性の平均の差が0.5以上だった場合には、それが良いほうが「良い回答」
2. 残りの内、正確性の平均の差が0.8以上だっ

た場合には、それが良いほうが「良い回答」

3. 残りの内、情報量の平均の差が1以上だった場合には、それが良いほうが「良い回答」
4. 残りの内、流暢性の平均の差が1以上だった場合には、それが良いほうが「良い回答」
5. そうでない場合には同等

<自然言語による記述>

まず、質問に対する回答でなければダメで、次に、そこに書かれている情報が正しいこと。それらが満たされていれば、より詳しく書いてあり、流暢である方がより良い回答である。

4. GPT-4の自動評価との比較分析

Rakuda ベンチマークでの(GPT-4による)評価結果を表4に示す。houou-v2システムがGPT-3.5に優っているという結果が得られた。この結果は、表3に示した人間による統合的な評価結果とは全く逆である。詳細な分析を行った。まず、表5に2種類の評価のconfusion matrixを示す。

システム	勝率 (数)
houou-7b-instruction-v2	67.5 (27)
GPT-3.5-turbo-1106	30.0 (12)
引き分け	2.5 (1)

表4. Rakuda ベンチマークによる評価結果

	houou	GPT-3.5	同等	合計
houou	9	13	5	27
GPT-3.5	2	9	1	12
同等	0	0	1	1
合計	11	22	7	

表5. 人間評価結果(横軸)とGPT-4による評価結果(縦軸)のconfusion matrix

表5における人間とGPT-4の評価結果の大きな相違点としては、GPT-4が「hououが良い」と判定した27件の内13件について人間は「GPT-3.5が良い」と判定し、同じく5件について人間は「同等である」と判定していることが分かる。付録A.2はこの13件の内の一例であり、同じ質問に対するGPT-4の出力を付録A.3に載せる。最初の13件について、出力、評価結果、コメントを全て読み、詳細な調査したところ、以下のことが分かった。

- 正確性の値は、GPT-3.5の方が1.39の差をつけて優れていると判定されている。この値はそれ以外の平均の0.60と比較して非常に大きい。つまり、人間はこれらのデータに対し、正確性の点でhououの出力は大きく劣ると判定している。
- 多くの場合、hououは関連した数値や具体的情報を記載しているが、誤っている場合がある^{※i}
- 多くの場合、GPT-3.5の出力は一般的な説明のみであり、正確性に影響するような数値や具体的な情報を使った説明が少ない。そのため正確性が比較的にならなると考えられる^{※i}
- GPT-4による判定説明文の中に、「hououの説明の方が具体的で詳細が記されていて優れている」という意味の記載が半数以上にあり、情報量について優れていると判定している。逆に、情報が間違っている場合でも「この記述は正確であり」という言及が見られ、正確性の判断ができていないことが伺える。

付録A.2,A.3に記載した例や対象の13例のみではなく、全体的に見ても、GPT-3.5の出力には具体的な数字や情報の記載が少なく、一般的なことのみが書かれている場合が多い。他方、hououでは、数値や具体的な情報による詳細な記述をしている場合が多い。これは、ichikara-instructionのアノテーターリーダーに問い合わせたところ「インストラクションの回答にはその説明の確証になるようなデータや事実を入れるように指導している」と述べており、その結果がhououの出力の特徴に表れていると考えられる。これらをまとめると、下記のことが導ける。

- hououは具体的な数値や具体例を含んだ情報量の多い出力をしているが、そこには誤りが含まれている。その誤りを人間は見抜いたが、GPT4は見抜けず、優劣判定の差ができた。
- 正確性についても、インストラクションが大きな影響を与えていた。つまり、情報量と正確性は、完全なシステムではない限り、トレードオフの関係にあり、hououは情報量は豊かであるが、そのために正確性を損ねた。
- 知識を問う質問が多い「地理」の10問中5問で、審判であるGPT4のコメントにハルシネー

ションが含まれていた。上記の分析も含め、やはり、GPT4は正確性に弱いと考えられる。

- よく言われる「GPTは長い回答に良いスコアを与える」というのは正しく因果関係を捉えておらず、「GPTは正確性の判断は苦手で、それ以外の要素で判断しており、長く詳細な回答は結果的に良いスコアを得ている」と考えられる。

これらのことを鑑みると、GPTはハルシネーションを避けるために、具体的な数値や細かい説明は避け、一般的な説明に終始するインストラクションチューニングをしているのではないかと推測できる。

5. 考察

本論文の分析を今後のLLMの開発に繋げるための考察を行う。まずは、人間の判定アルゴリズムが推測できたことにより、新たなLLM評価方法の模索が望まれる。ただし、現状の生成AIが直接的には正確性を判定することができないことが課題になるかもしれない。しばらくは人間による評価も大きな価値があると考えられる。また、インストラクションの重要性が再認識されたことにより、注意深くその設計方針を再検討する必要がある。ichikara-instructionでは、各質問回答ペアに様々な種類のタグを付与する計画であるが、少なくともそこには「詳細度」のようなタグの種類を考案し、現在のGPTの出力タイプ（一般的な説明）とhououの出力タイプ（具体的な説明も提示）の間を調整可能な仕組みの研究が一つのテーマになるであろう。

また、現状では深く分析ができておらず、結論を導く段階にはないが、興味深い現象が観測されている。例えば、Rakudaベンチマークにある4つのカテゴリ毎に人間の評価、GPT-4の評価ともに大きな違いがある。また、アノテーターによる判定の違いはLLMのパーソナライゼーション研究の取っ掛かりとなると思われる。GPT-4の出力はシステムの順番を変えると評価結果が異なる現象を観察している。「関連性」の評価結果とインストラクションのカバレッジにも関係性がありそうである。日本語のLLMをしっかりと構築していくためには、着実な歩みを進めていくことが非常に重要だと考える。

※i 付録A.2の例で説明する。hououの出力(B)には、憲法43条（正しくは14条）、小委員会の名前と設置日（このような名前の小委員会存在しない）など付加的で不正確な情報が記載されている。一方、GPT-3.5の出力

にはこのような「具体的な情報」はほぼない。付録A.3に示した通り、GPT-4はこの誤りが含まれるhououの出力に対し、「正確さと相関性も（中略）ユーザーの質問に対して十分でした。」と書いている。

謝辞

本論文の評価分析に関し、LLM 勉強会での議論が参考になった。本研究に参加した5人のアノテーターの地道な努力に感謝する。また、ichikara-instruction 作成チームのアノテーターをはじめとするメンバーの努力や、本論文に対する有益なコメントにも感謝する。

参考文献

1. Zheng et.al. “Judging LLM-as-aJudge with MT-Bench and Chatbot Arena”. 37th Conf. of NeurIPS. 2023.
2. Rakudaベンチマーク. <https://yuzuai.jp/benchmark>
3. GPT-4. <https://openai.com/gpt-4>
4. GPT-3.5-turbo-1103. <https://platform.openai.com/docs/models/gpt-3-5>
5. 小島、北岸. “大規模言語モデルhouou (鳳凰): 理研ichikara-instruction データセットを用いた学習と評価”. 言語処理学会第30回年次大会, 2024.
6. houou-7b-instruction-v2. <https://huggingface.co/moneyforward/houou-instruction-7b-v2>
7. Llama-2. <https://ai.meta.com/llama/>
8. rinna/youri-7b. <https://rinna.co.jp/news/2023/10/20231031.html>
9. ichikara-instruction. LLMのための日本語インストラクション作成プロジェクトホームページ: <https://liat-aip.sakura.ne.jp/wp/llmのための日本語インストラクションデータ作成>.
10. 関根聡, 安藤まや, 後藤美知子, 鈴木久美, 河原大輔, 井之上直也, 乾健太郎. “ichikara-instruction : LLM のための日本語インストラクションデータの作成”. 言語処理学会第 30 回年次大会. 2024.

付録 A.1 Rakuda ベンチマークで利用している prompt

prompt = ""あなたは、回答の質をチェックするための審判員です。

[質問]

{question}

[アシスタント 1 の回答の開始]

{as1}

[アシスタント 1 の回答の終了]

[アシスタント 2 の回答の開始]

{as2}

[アシスタント 2 の回答の終了]

上に表示されたユーザーの質問に対する 2 つの AI アシスタントのパフォーマンスについて、あなたのフィードバックをお願いします。回答の有用性、関連性、正確性、詳細度、日本語能力を評価してください。まず、アシスタントの有用性、関連性、正確性、詳細度、日本語能力の評価を提供してください。評価の包括的な説明も提供してください。ユーザーは日本語しか話さないで日本語で書かれていない回答には低評価をつけてください。 偏見を避け、回答の提示された順序があなたの判断に影響を及ぼさないことに気をつけてください。両方の解答を慎重に評価した後、評価が高い方のアシスタントの解答を選び、アシスタント 1 の回答であれば 1 を、アシスタント 2 の回答であれば 2 を、そしてアシスタント 1 とアシスタント 2 の間から選ばない場合は 3 を最後の行に出力してください""

付録 A.2 人間による評価結果の例

prompt	A	Aコメント	B	Bコメント	流暢性		正確性		情報量		関連性		優
					A	B	A	B	A	B	A	B	
一票の格差問題とは何か説明し、これが日本の政治にどのような影響を及ぼしているかを論じてください。	一票の格差とは、選挙において1人の有権者の投票価値が他の有権者よりも大きいことを指します。例えば、選挙区ごとの有権者数が大きく異なる場合、人口の多い選挙区の有権者の一票の価値が低くなってしまふことがあります。これは選挙民の平等な権利を侵害することになります。 日本の場合、選挙区ごとの人口密度の差が大きく、それによって一票の格差が生じています。例えば、人口の少ない地方の選挙区では、一票の価値が人口の多い都市部の選挙区に比べて大きくなってしまふことがあります。 この一票の格差が日本の政治に与える影響としては、人口の多い都市部の選挙区の影響が相対的に弱まることや、地方の利益が過大に反映されることが挙げられます。また、一票の格差が選挙制度の公平性を問題視されることで、政治の信頼性や選挙制度への不信感が生じる可能性もあります。 このような影響を考慮し、一票の格差を解消するための選挙制度改革や選挙区の再編成などが議論されています。	問題点や改善案も書かれている点が良い	一票の格差問題とは、選挙区ごとの人口や面積等に格差があるため、有権者ひとりあたりの投票価値が違ってしまう問題のことです。 この問題は、日本の国会である参議院において、2015年（平成27年）以降、抜本的な改革が行われていないために、常態化しています。 一票の格差が最大となった2022年（令和4年）7月の参議院議員通常選挙については、最高裁判所大法廷において、投票価値の平等を定めた憲法第43条に違反する無効な選挙とする判決が出されています。これを受けて、参議院は、2022年（令和4年）9月27日に「参議院議員の選挙区選出議員の選挙に関する規定に係る憲法の規定に基づく参議院選挙制度の見直しに関する小委員会」（小委員会）を設置し、「一票の格差」の解消に向けた議論を行っています。 日本の政治において、一票の格差問題は、選挙の度に論点となり、抜本的な改革が求められています。	詳細な内容に踏み込んでいるが、その内容が正確でない	5	5	5	2	5	5	5	5	A

付録 A.3 上記の同じ問題に対する GPT-4 の評価結果の出力

“アシスタント 1 の説明は具体的で詳細であり、一票の格差問題の定義を明らかにしています。また最高裁判所の判決、参議院の対応、その影響、そして日本の政治における重要性について具体的な事例を用いて説明しています。一方、アシスタント 2 も一票の格差問題とその影響について明確に説明していますが、具体的な事例や裁判の判決などの具体的な情報は提供していませんでした。その結果、こちらの説明は抽象的で一般的なものとなってしまっています。¥n¥n 両者とも日本語の表現は適切で、正確さと相関性も一票の格差問題が日本の政治にどのような影響を与えているかというユーザーの質問に対して十分でした。しかしながら、アシスタント 1 の方が具体的な情報を提供しているため、より詳細性と有用性が高いと言えます。¥n¥n したがって、アシスタント 1 がユーザーの質問に対してより有用で詳細な情報を提供していると考えます。ですので、この質問に対する優れた回答はアシスタント 1 のものであると評価します。”

(アシスタント 1 が A.2 における B のシステム)