日本語を対象としたLLMの大規模人手評価

井之上直也 [1,2]　安藤まや [1]　後藤美知子 [1]　関根聡 [1]　中山功太 [3]　宮尾祐介 [3]
[1] 株式会社いちから　[2]JAIST　[3]NII LLMC
naoya.inoue@ichikara.ai

概要

大規模言語モデル（LLM）の評価は，LLMによる自動評価が主流となっているが，自動評価には多くの課題が存在し，LLMの評価方法論そのものにおいて決定的な解は未だ得られていない．本研究では，今後のLLM評価に関する研究を支援することを目的として，484件の日本語プロンプトに対する10種類のLLMの応答を対象に，5つの評価項目に基づいた大規模な評価を実施し，その結果を公開する．本稿では，評価項目の設計方法と評価の実施手順を報告するとともに，構築した評価データに基づく予備的な分析として，評価項目間の関連性分析やLLMの性能比較についても触れる．

1　はじめに

近年，自然言語処理の分野では，大規模言語モデル（Large Language Models, 以下 LLM）の発展が著しい．LLMの品質評価は，LLM-as-a-Judgeをはじめとする LLMによる自動評価が主流となりつつあるが [1, 2]，自動評価には多くの課題があることが指摘されている [3, 4, 5]．

例えば，文献 [4] では，LLMに基づく自動相対評価において，応答の順序を入れ替えるだけで評価結果が大きく変化することが報告されている．また，文献 [3] では，要約タスクにおける自動評価が，使用する評価指標によってLLMの自動評価結果と人間の評価結果の相関が大きく異なることを指摘している．しかし，人手によるLLMの応答評価はコストが高く，自動評価を直ちに中止することは現実的ではない．したがって，LLMの評価の方法論そのものを抜本的に再検討することが喫緊の課題である．

そこで本研究では，LLMの応答を人手で評価した大規模なデータセットを構築する．具体的には，484件の日本語プロンプトに対する10種類のLLMの応答を，53名の作業者によって評価したデータセットを作成した．このデータセットは，図1に示すように，関連性 (Relevance)，流暢性 (Fluency)，正確性 (Factuality)，詳細性 (Detailedness)，総合評価 (Overall) という5つの観点から多様なLLMの応答を詳細に評価し，1応答あたり2人以上の評価者がスコアを付与したものとなっており，今後のLLM評価方法の研究に資することが期待される．本稿では，評価項目の設計方法 (§2.2)，評価の実施手順 (§2.3)，および最終的な成果物の概要 (§3) について報告する．さらに，評価結果の予備的な分析として，評価項目間の関連性の分析，LLMの性能比較についても述べる (§4)．構築したデータセットは，https://github.com/llm-jp/llm-human-eval-jp にて公開する．さらに，他グループによる関連研究として，安全性の観点からLLMの大規模評価を行う取り組みもあり [6]，こちらも併せて参照されたい．

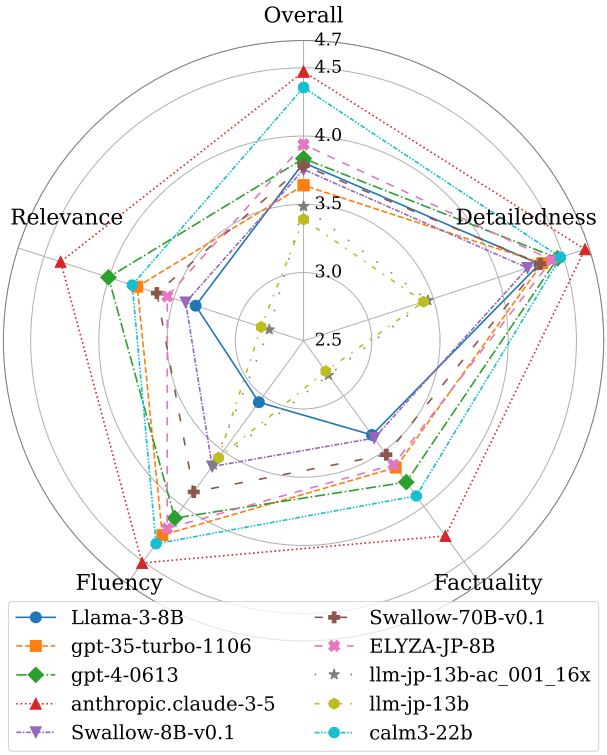

図1　484件の日本語プロンプトについて，10種類のLLMの応答を5つの評価項目で人手により評価した．

表 1　評価対象となるプロンプトの概要.

ソース	プロンプトのカテゴリ	件数	プロンプトの例
ichikara-instruction [7]	オープン QA，クローズド QA，ブレスト，創作，定義，要約，抽出，例示，分類，選択，穴埋め，書き換え，校正，翻訳，数学	104	近々結婚することになり、結婚式の計画を立て始めたところです。自分の結婚式のときには、プロのヘアメイクアップアーティストの方にお願いしたいと考えています。その場合、どのようにして自分に合う人を見つけてお願いをすればよいのでしょうか。また、個別に頼む際の金額の相場も知りたいです。
日本語 VicunaQA [8]	一般，知識，ロールプレイ，一般常識，フェルミ推定，反実仮想，コーディング，数学，創造的執筆	80	典型的な冬を考えた時、その間に降る雪の結晶は何個でしょうか？あなたの答えを説明してみてください。その際、あなたの推論過程を段階的に説明してください。
The Rakuda Benchmark [9]	歴史，社会，政治，地理	40	日本の「三位一体改革」について述べ、その経済に対する影響について解説してください。
ELYZA-tasks-100 [10]	要約修正，複雑な算数，ロールプレイ，情報抽象化，未知言語翻訳，対話生成，創造的生成	100	クマが海辺に行ってアザラシと友達になり、最終的には家に帰るというプロットの短編小説を書いてください。
Japanese MT-Bench [11]	人文，数学，推論，ロールプレイ，執筆，コーディング，抽出，STEM	80	量子物理学の中で、重ね合わせ状態とは何ですか？それはどのようにして量子もつれ現象と関連していますか？ / あなたの回答に含まれている前提は何ですか？それらは有効ですか？

表 2　評価対象とした LLM の一覧.

名前	サイズ
日本語非特化モデル:	
Llama 3	8B
GPT-3.5 (turbo-1106)	-
GPT-4 (0613)	-
Claude 3.5 Sonnet (20240620-v1:0)	-
日本語特化モデル:	
Swallow v0.1 (Llama 3 ベース)	8B
Swallow v0.1 (Llama 3 ベース)	70B
ELYZA LLM for JP (Llama 3 ベース)	8B
llm-jp v2.0 (16x)	13B
llm-jp v2.0	13B
CyberAgentLM3 (CALM3)	22B

2　評価方法

2.1　評価タスクとモデル

タスク　より現実的なシナリオにおいて LLM を対話システムとして評価するためには，ユーザが現実的に要求しうる多様なプロンプトを収集する必要がある．そこで，表 1 に示すように，既存の日本語 LLM 評価データセットから合計 404 件のプロンプトを収集した．なお，このうち Japanese MT-Bench の 80 件については，対話形式を想定した追加の 1 プロンプトがあるため，最終的にのべ 484 件のプロンプトが評価対象となる．これらを LLM に入力し応答を生成し，人手による評価を実施した．

LLM　多様なバリエーションの応答について評価を実施するため，モデルサイズ，オープンソースか否か，日本語特化型であるかといった観点に基づき，表 2 に示す 10 種類の多様な LLM を評価対象とした．詳細なモデル名は，付録 A を参照されたい．

2.2　評価指標

表 3 に示す 5 つの評価項目について，プロンプトに対する各システムの応答を 1〜5 の 5 段階評価で採点した．ここで，1 点は「その評価項目において，考慮の対象にならないレベルである」という最低点を示し，5 点は「その回答で十分であり，改良の必要性を感じないもの」という最高点を示している．

また，各評価項目は，他の評価項目の影響を受けることなく独立に評価するよう指示した．例えば，応答が事実として誤っていた場合，正確性のスコアは低くなるが，含まれる情報量が多ければ詳細性のスコアは高くなる，といった具合である．

流暢性についての補足　流暢性については，本研究が日本語 LLM を評価対象としているため，応答が日本語でない場合は基本的に 1 点とし，他の評価項目は「スコアなし」とすることとした．ただし，「特定の言語で回答せよ」といった指示があるプロンプトの場合は，その外国語での流暢性を評価する．また，外国語と日本語が混在するのが妥当と考えられるプロンプトについては，全体として流暢であるかどうかを評価した．

正確性についての補足　正確性については，応答に書かれた情報が事実であるかどうかを確認するため，検索エンジンや書籍などを用いて必ず事実確

表 3　評価項目の一覧.

項目	基準
関連性	プロンプトに対する応答として成立しているか．応答がプロンプトの要求（文字数制限，箇条書き等）を満たしているか．
流暢性	文章が日本語として正しいか（文法，ことばの使い方等），構成などが整っているか，自然で読みやすいか．
正確性	情報が事実として正しいか．
詳細性	情報量が多いか．
総合評価	上記 4 つの評価を踏まえた総合的な評価．

認を行うよう指示をした．また，前述のように，各評価項目は独立であるため，応答がプロンプトに対する回答として成立していなくても，応答に書かれた情報が正確であれば，高スコアを与えることとした．一方，プロンプトの内容によっては，応答の正確性を判断する必要がない場合もある．その場合は「スコアなし」とした．これには以下のような応答を期待するプロンプトが含まれる:

- 予想，想像上の事実，想像上の意見
- 反実仮想的な発言（「徳川家康が現在の日本の総理大臣になったらどのような政策をするでしょうか」等）
- 時間軸によって内容が変わるもので，質問や回答に時間軸の指定がないもの（「一番安い携帯電話会社はどこ？」等）

ただし，架空や想像上の人物（例えば，仮想的な人物設定やフィクションのアニメ・漫画・ドラマ・映画のキャラクタ）に関する質問の場合，その架空世界（例：アニメやドラマの設定）に照らして事実が正しいかどうかを確認し，評価することとした．

2.3　評価手順

自然言語処理のデータセット作成経験を持つプロの作業者を中心に，学部生や大学院生など 53 名の作業者を集めた．1 つのプロンプトに対して 10 種類の LLM の応答が存在するため（追加プロンプトがある 80 件については 20 個の応答），これらをまとめて 1 つの評価事例とし，各評価事例には 2 人または 3 人の評価者を割り当てて評価を実施した．

§2.2 で述べた評価基準に関するマニュアルを事前に配布し，各作業者は最初の 3 件の作業を終えた後，プロの作業者からフィードバックを受けることで，評価基準のキャリブレーションを行った．

また，評価作業中に疑問が発生した場合に対応できるよう，Slack 上に質問を随時投げられる環境を整備し，随時疑問を解消可能な体制を構築した．最終的に，Slack チャンネルには約 30 件の質問が寄せられ，第一著者がこれに回答した．

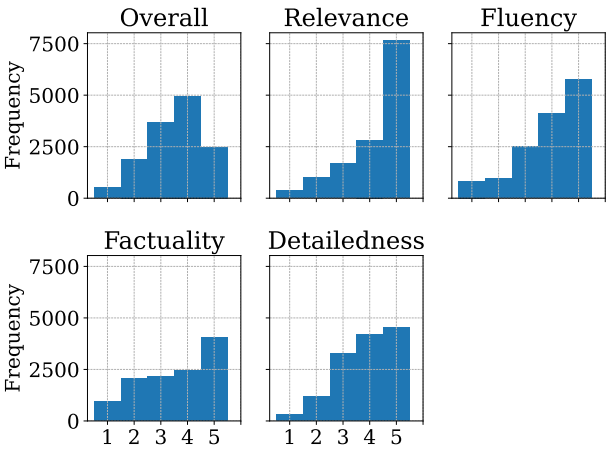

図 2　人手評価のスコアの分布.

3　評価データ

3.1　基本統計

最終的には，383 件の事例に 3 人の評価者，21 件の事例に 2 人の評価者が採点を行った．LLM に依らない全体の評価の傾向を把握するために，全 LLM の応答に付与された評価者のスコア分布を図 2 に示す．全体的な傾向として，関連性は非常に高く評価される傾向にあり，流暢性についても比較的高いスコアが付与されていることがわかる．一方で，総合評価については 4 点が大多数であり，LLM の応答には改善の余地があることが見て取れる．

3.2　作業者間の評価の一貫性

設計した評価項目の信頼性を見積もるため，各評価項目について評価者間のスコアのばらつきを調査した．ある評価項目 m において，プロンプト i に対する LLM j の応答に付与された評価者のスコアの集合を $S^m_{i,j}$，このスコアを範囲 [2,4] にクリップした値を $\hat{S}^m_{i,j} = \max(\min(S^m_{i,j}, 4), 2)$ とする．ここで，**評価の不一致度** d を以下のように定義する:

$$d^m(i,j) = \max(\hat{S}^m_{i,j}) - \min(\hat{S}^m_{i,j}) \in \{0,1,2\} \quad (1)$$

すなわち，すべての評価者のスコアが同一スタンスである場合（すべて 2 以下，すべて 4 以上，またはすべて 3）には不一致度は 0 となる．それ以外の場合，不一致度は評価者間のスコアのばらつきに応じた値を取る．なお，§2.2 で述べたように，流暢性が

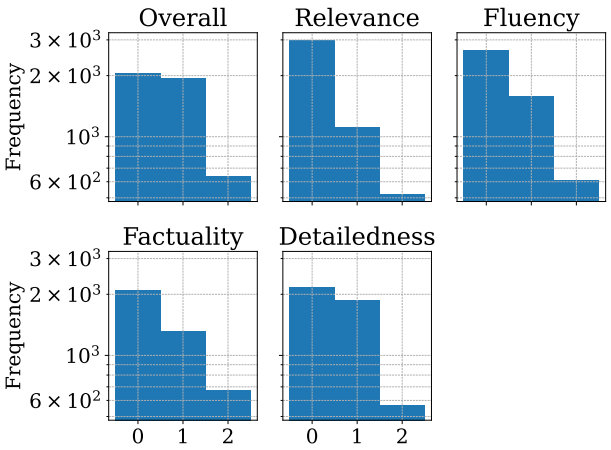

図 3　人手評価の評価者間の不一致度の分布.

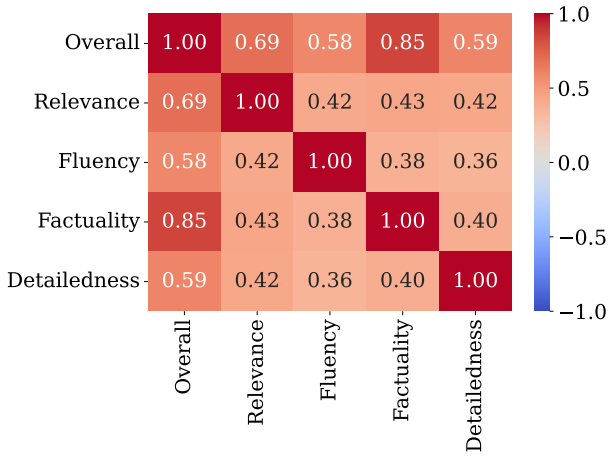

図 4　各評価項目のスコアの間の Pearson 相関係数.

1 の場合，その他の評価項目にスコアが付与されない場合があるため，常にすべての応答に対して複数のスコアが付与されているわけではない．不一致度の計算では，2 人以上の評価者がスコアを付与した場合のみを対象とした．

図 3 に，評価項目ごとの不一致度 $d^m(\cdot,\cdot)$ の分布を示す．すべての項目において，評価者のスコアが一致している事例 (不一致度 0) が大多数を占める一方で，総合評価，正確性，流暢性においては，評価者の間で逆のスタンスを取るようなケース (不一致度 2) が相対的に多く観察された．

そこで，特に不一致度の大きい正確性について，不一致度が高かったシステムの応答 637 件について，信頼性の高いプロの作業者による追加評価を実施した．この際，正確性の評価が総合評価に影響を与える可能性があるため，総合評価についても追加の再評価を行った．こうして得られた評価結果を，最終成果物として公開する．

4　予備的分析

4.1　評価項目間の関連性

各評価項目間の関係性を調べるため，各応答に付与された複数の作業者のスコアを平均し，それぞれの評価項目間の Pearson 相関係数を計算した．

結果を図 4 に示す．正確性と関連性の評価は特に総合評価と強い相関を持ち，LLM の応答の総合的な品質は，提示された情報の事実性やプロンプトとの整合性に大きく影響を受けていることが示された．一方で，個別の評価項目の間には強い相関は見られず，§2.2 で意図したとおり，評価項目間の独立性がある程度担保されていることが確認できた．

4.2　LLM の性能比較

各システムの評価結果の平均値を図 1 に示す．まず，日本語非特化モデルである Claude 3.5 Sonnet が，すべての評価項目で他の LLM を圧倒していることがわかった．次に，モデルサイズが比較的小さい日本語特化モデルの CALM3 が高い評価を得たものの，関連性や正確性の面では Claude 3.5 Sonnet や GPT-4 に差をつけられており，日本語特化 LLM にはプロンプトへの追従性や応答内容の事実性向上など，さらなる改善の余地があることが示された．

一方，小規模な日本語非特化型モデルである Llama-3-8B は，日本語の訓練データがほとんど含まれていないにもかかわらず，総合評価で一定のスコアを得ており，日本語のプロンプトをある程度理解し適切な応答を生成できていることがわかる．しかしながら，流暢性において特に低い評価を受けており，日本語データによる LLM の訓練が重要であることを示唆している．

5　おわりに

LLM の評価方法論に関する研究を促進することを目的に，LLM の応答を人手で大規模に評価したデータセットを構築した．本稿では，構築したデータの概要を報告するとともに，予備的な分析を通じてその活用例を示した．本データセットのさらなる活用例として，スコアの低い評価項目における応答の傾向分析，独自の LLM 自動評価モデルの訓練・評価データとしての利用などが考えられる．本データセットが，LLM の評価方法論のさらなる進展に貢献することを期待する．

謝辞

評価にご協力いただいた作業者 53 名の方々に深く感謝申し上げます.

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 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**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[2] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In **Proc. of EMNLP**, pp. 4334–4353, 2024.

[3] Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of EMNLP 2023**, 2023.

[4] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In **Proc. of ACL (Volume 1: Long Papers)**, pp. 9440–9450, 2024.

[5] 関根聡, 小島淳嗣, 貞光九月, 北岸郁雄. 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. 言語処理学会第 30 回年次大会論文集, 2024.

[6] 高橋哲朗, 鈴木久美, 関根聡. LLM の安全性における大規模人手評価. 言語処理学会第 31 回年次大会論文集, 2025.

[7] 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構築. 言語処理学会第 30 回年次大会論文集, 2024.

[8] 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 **Proc. of LREC-COLING 2024)**, 2024.

[9] YuzuAI. The Rakuda Benchmark. https://huggingface.co/datasets/yuzuai/rakuda-questions, 2023.

[10] 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. https://huggingface.co/elyza/ELYZA-tasks-100, 2023.

[11] Stability AI. Japanese MT-Bench. https://github.com/Stability-AI/FastChat/tree/jp-stable/fastchat/llm_judge, 2023.

A　評価対象の LLM のモデル名

- meta-llama/Meta-Llama-3-8B-Instruct
- openai/gpt-35-turbo-1106
- openai/gpt-4-0613
- anthropic/anthropic.claude-3-5-sonnet-20240620-v1:0
- tokyotech-llm/Llama-3-Swallow-8B-Instruct-v0.1
- tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1
- elyza/Llama-3-ELYZA-JP-8B
- llm-jp/llm-jp-13b-instruct-full-ac_001_16x-dolly-ichikara_004_001_single-oasst-oasst2-v2.0
- llm-jp/llm-jp-13b-instruct-full-dolly-ichikara_004_001_single-oasst-oasst2-v2.0
- cyberagent/calm3-22b-chat