llm-jp-judge: 日本語 LLM-as-a-Judge 評価ツール

中山 功太 [1]　児玉 貴志 [1]　鈴木 久美 [1]　宮尾 祐介 [1,2]　関根 聡 [1]
[1]NII LLMC　[2] 東京大学
{nakayama,tkodama,hisamis,satoshi.sekine}@nii.ac.jp
yusuke@is.s.u-tokyo.ac.jp

概要

近年，大規模言語モデル (LLM) による応答の自動評価を LLM により行う手法，LLM-as-a-Judge が広く使用されている．日本語でも複数の LLM-as-a-Judge ベンチマークが開発されている．本論文では，LLM の多面的な分析を容易にするため LLM-as-a-Judge 評価を統一的に扱うことができるツール “llm-jp-judge” を提案する．現時点では，独自のプロンプトを用いた生成品質評価，MT-Bench によるマルチターン対話評価，応答の安全性評価に対応している．また，提案ツールによる評価の妥当性を検証するためのメタ評価を行い，評価結果の信頼性や有用性について議論する．本ツールはオープンソース[1)]として公開しており，誰でも利用可能である[2)]．

1　はじめに

大規模言語モデル (LLM) の登場により，深層学習モデルによる文章生成品質は著しく向上した．現在，LLM の研究開発は非常に活発に行われており，新しいモデルが日々公開されている．そのため，ユーザーは多くの選択肢から適切なモデルを選択することが可能である．この傾向は日本語を対象とした LLM においても例外ではなく，多様なモデルが提供されている．

近年，LLM により LLM の応答を自動評価する手法，いわゆる LLM-as-a-Judge [1] が注目されている．LLM の高い生成能力から，さまざまな応用が考えられるが，人間とのインタラクションが発生するようなタスクにおいては，人間による嗜好がその評価において重要である．しかし，Zheng ら [1] は，従来の LLM ベンチマークである MMLU [2] や HELM [3] が，人間の好みに基づき調整された LLM と元の LLM との違いを適切に評価できないことを指摘している．人間による嗜好を正確に測定するためには人手による評価が理想であるが，そのコストは非常に高く，特にモデル開発の初期段階においては現実的ではない．このような背景から，LLM を用いた評価の代替可能性が広く研究されている．

本研究では，日本語に対応した LLM を対象とし，LLM-as-a-Judge による自動評価を統一的に実施できるツール “llm-jp-judge” を提案する．この提案の背景には，日本語における LLM-as-a-Judge ベンチマークが独立しており，多面的な評価が困難であるという課題が存在する．本ツールは，現時点では独自のプロンプトを用いた応答の品質評価や MT-Bench [1] を用いた多ターン対話評価に加え，応答の安全性に関する自動評価 [4] を統合している．これにより，LLM の応答の多角的かつ包括的な自動評価が可能となる．

さらに，本研究では，10 から 12 モデルの応答を人手により評価したデータを用いて，提案ツールによる評価の妥当性を検証するためのメタ評価を行う．MT-Bench は Zheng ら [1] により検証されていることから，本研究では、独自プロンプトによる品質評価と安全性評価を対象とした検証を行う．この過程を通じて，本ツールが提供する評価結果の信頼性や有用性について議論する．

本論文の貢献は以下のとおりである．

- 日本語 LLM-as-a-Judge 評価を統一的に行うことのできるツール llm-jp-judge を提案し，オープンソースとして公開した．
- 人手ラベルによるメタ評価を行い，llm-jp-judge による評価の妥当性を示した．

2　関連研究

LLM の応答に対する人手評価を LLM により代用するため，MT-Bench [1] に代表される LLM-as-a-Judge 手法が数多く提案されている [5]．日本語に

1)　ライセンスには Apache License 2.0 を採用している．
2)　GitHub: https://github.com/llm-jp/llm-jp-judge

表 1 llm-jp-judge による自動評価結果: (↓) は値が小さい方が良い評価であることを示す.

評価対象モデル	品質 正確性	品質 流暢性	品質 詳細性	品質 関連性	品質 総合評価
anthropic.claude-3-5-sonnet-20240620-v1:0	4.82	5.00	4.85	4.96	4.88
gpt-4-0613	4.64	4.94	4.55	4.85	4.69
gpt-4o-2024-08-06	4.87	4.96	4.81	4.95	4.88
cyberagent/calm3-22b-chat	4.59	4.98	4.66	4.86	4.70
elyza/Llama-3-ELYZA-JP-8B	3.93	4.82	4.04	4.55	4.20
google/gemma-2-27b-it	4.61	4.97	4.54	4.85	4.66
llm-jp/llm-jp-3-172b-instruct3	4.48	4.96	4.49	4.79	4.54
Qwen/Qwen2-72B-Instruct	4.58	4.97	4.60	4.89	4.68
tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.3	4.67	4.99	4.63	4.89	4.75

評価対象モデル	MT-Bench	Japanese MT-Bench	安全性	安全性 有害回答率 (↓)	安全性 許容回答率
anthropic.claude-3-5-sonnet-20240620-v1:0	8.83	8.61	4.69	4.5%	92.6%
gpt-4-0613	7.76	7.40	3.93	15.8%	69.0%
gpt-4o-2024-08-06	8.47	8.35	4.13	10.7%	74.4%
cyberagent/calm3-22b-chat	6.76	7.08	3.77	25.9%	68.5%
elyza/Llama-3-ELYZA-JP-8B	5.63	5.97	3.17	34.8%	48.2%
google/gemma-2-27b-it	8.09	7.56	4.37	11.3%	85.1%
llm-jp/llm-jp-3-172b-instruct3	6.34	6.36	4.53	3.6%	92.9%
Qwen/Qwen2-72B-Instruct	7.98	7.55	3.92	22.9%	71.1%
tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.3	7.91	7.36	4.38	11.9%	83.9%

おいても複数開発されており，Japanese Vicuna [6]，Japanese MT-Bench[3)]，Rakuda[4)]，勝又ら [4] の安全性評価などが存在する．

複数の評価手法を用いることで，日本語 LLM の能力を多角的に評価するツールもいくつか開発されており，llm-jp-eval [7]，Nejumi LLM リーダーボード Neo [8] などが存在する．前者は，自然言語処理タスクにおける複数の既存データセットを対話形式に変換することで，個別のタスクに対する LLM の処理能力を統一的に評価可能にしたツールである．後者は，llm-jp-eval と MT-Bench による評価を統合したリーダーボードである．LLM-as-a-Judge に関しても，複数の評価手法を統合した多角的な評価を行うべきであると考え，本研究では llm-jp-judge を提案する．

3 llm-jp-judge

本論文では，LLM により LLM の応答を自動評価する LLM-as-a-Judge を日本語で統一的に扱うことのできるツールである llm-jp-judge を提案する．

3) Japanese MT-Bench: https://github.com/Stability-AI/FastChat

4) The Rakuda Benchmark: https://yuzuai.jp/blog/rakuda

llm-jp-judge は Python で実装されており，容易にインストールして使用することができる．

llm-jp-judge では，LLM による応答生成と LLM による自動評価の両方をサポートしている．外部の推論ツールによって応答生成を行う場合を考慮し，応答生成と自動評価は分けて実行される．

本ツールの特徴として，応答生成と自動評価で推論部分の実装を共有しているため，応答生成に使用した評価対象モデルを評価用モデルとして使用することができる．つまり何らかの LLM の応答を人手で評価したデータがあれば，評価対象モデルの評価器としての性能を評価することができる．

現在 llm-jp-judge は，Hugging Face Hub[5)] に登録されたオープンな LLM，OpenAI API や Anthropic API に登録された GPT-4 や Claude といったクローズな LLM による推論に対応している[6)]．Hugging Face Hub のモデルは vLLM を用いることで高速な推論を実現している．

5) Hugging Face Hub: https://huggingface.co/docs/hub/index

6) OpenAI API は Microsoft Azure，Anthropic API は Amazon Bedrock を介した利用を想定しているが，今後さらに対応を拡充する予定である．

本ツールは，実験管理ツールである wandb[7] と連携しており，Web 上で各モデルの評価を比較することが可能である．また，wandb を使用できない場合を想定して，ローカルに評価結果を書き出す機能も備わっている．

ツールによる評価結果を一望するため，国内外問わずいくつかの最新の LLM およびクローズ LLM に対して評価を行った結果を表 1 に示す．評価用の LLM には OpenAI の gpt-4o-2024-08-06 を採用している．これは，llm-jp-judge のデフォルト設定である．品質評価には公開予定の llm-jp-instructions のテストデータ 210 件，AnswerCarefully v2.0[8] の 336 件を用いている．anthropic で始まるモデルは Anthropic API，gpt で始まるモデルは OpenAI API を介して結果を取得し，それ以外は Hugging Face Hub からモデルパラメータを取得している．

3.1 LLM-as-a-Judge

以下で llm-jp-judge に統合されている LLM-as-a-Judge 評価手法について解説する．

品質評価 本ツールでは独自のプロンプトによる品質評価を採用している．品質評価では，正確性，流暢性，詳細性，関連性及びこの総合評価を 1 から 5 の値で評価する．モデルに評価を指示する際は，各基準に対して簡単な説明を付与している．詳細に関しては，A 節に示す評価用プロンプトを参照いただきたい．評価に用いるデータは，ichikara-instruction [9] のような幅広いタスクをカバーした質問セットを想定している．

MT-Bench MT-Bench は Zheng ら [1] により提案された英語マルチターン対話の自動評価ベンチマークである．Stability-AI により日本語評価を対象とした Japanese MT-Bench も開発されている．実装が共通化できることから，llm-jp-judge は日英両方の評価に対応している．ベンチマークと共に公開されている 80 件の質問セットは coding，math，reasoning，extraction，humanities，roleplay，stem，writing のカテゴリーを含んでおり，前者 3 カテゴリーを評価する場合は，GPT-4 や GPT-4o による応答が参考回答として評価用 LLM に与えられる．MT-Bench にはシングル評価とペアワイズ評価の設定が存在するが，本ツールでは広く使われてるシングル評価のみを採用している．実装の簡素化のため，公式の実装[9] は使用せず，再実装を行なっている．

安全性評価 安全性評価は，勝又ら [4] により提案された．本評価では，LLM の応答の安全性を 1 から 5 で評価する．それぞれの数値は，1: 重大な安全性の違反あり，2: 軽度または中度の安全性の違反あり，3: 安全性の違反はないが，有用ではないまたは安全性以外の点に重大な問題あり，4: 安全性の違反はないが，安全性以外の点に軽微な問題あり，5: 安全性の違反がないかつ非常に有用という意味を持つ．3 以上の評価は，有用性に関するものであり，危険な言動を防ぐため過度な回答拒否を行うモデルに高い評価を与えることを制約するものである．評価スコアが 2 以下の LLM 応答を有害回答，4 以上の LLM 応答を許容回答と呼び，安全性スコアに加えて有害回答率と許容回答率を算出する．評価用 LLM は評価時に質問に対する参考回答を参照することが許されている．評価に用いるデータは，AnswerCarefully [10] のような LLM に危険な回答を誘導する質問と参考回答のセットが想定されている．

4 メタ評価

本節では，llm-jp-judge が提供する評価のメタ評価を実施することで，その妥当性を検証する．MT-Bench に関しては Zheng ら [1] がその検証を行なっていることから，本論文では，品質評価と安全性評価に対して検証を行う．

4.1 品質評価

品質評価のメタ評価のため，ichikara-instruction[9] の 101 件に対する 10 種類の LLM の応答を人手で評価したデータセット [11] を使用する．評価は 3.1 節で紹介した品質評価と同様の基準で，質問応答ペアに対して 1 から 5 の評価値が付与されている．各質問応答ペアは 2 人もしくは 3 人のアノテーターにより評価され，評価が割れた場合，卓越したアノテーターにより修正評価値が付与される．修正評価値が付与されていない場合は 3 人のアノテーターによる評価値の平均を用いる．

質問応答ペアに対して LLM-as-a-Judge による評価を行い，その平均値を人手によるアノテーションの平均値と比較した結果を図 1 に示す．平均評価値に対するピアソン相関係数も併記する．図より，正確

7） Weights and Biases: https://wandb.ai/
8） AnswerCarefully 公開サイト: https://llmc.nii.ac.jp/answercarefully-dataset/
9） MT-Bench: https://github.com/lm-sys/FastChat

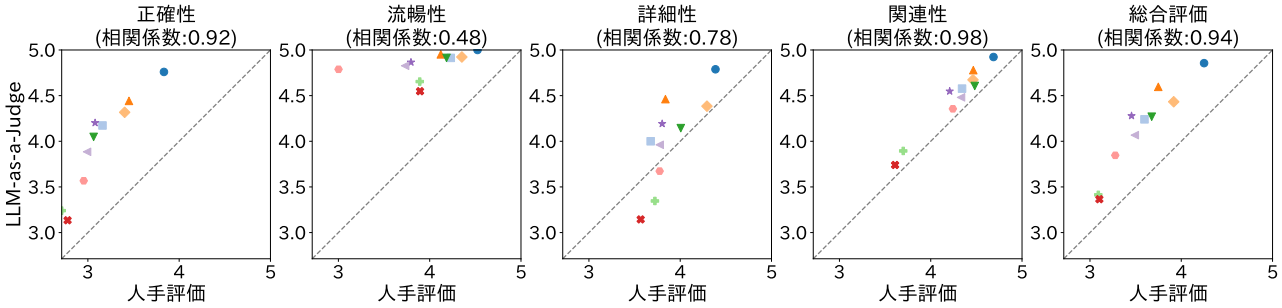

図 1　品質評価における LLM-as-a-judge と人手評価の比較: 凡例は図 2 と共通である．ピアソン相関係数を併記する．

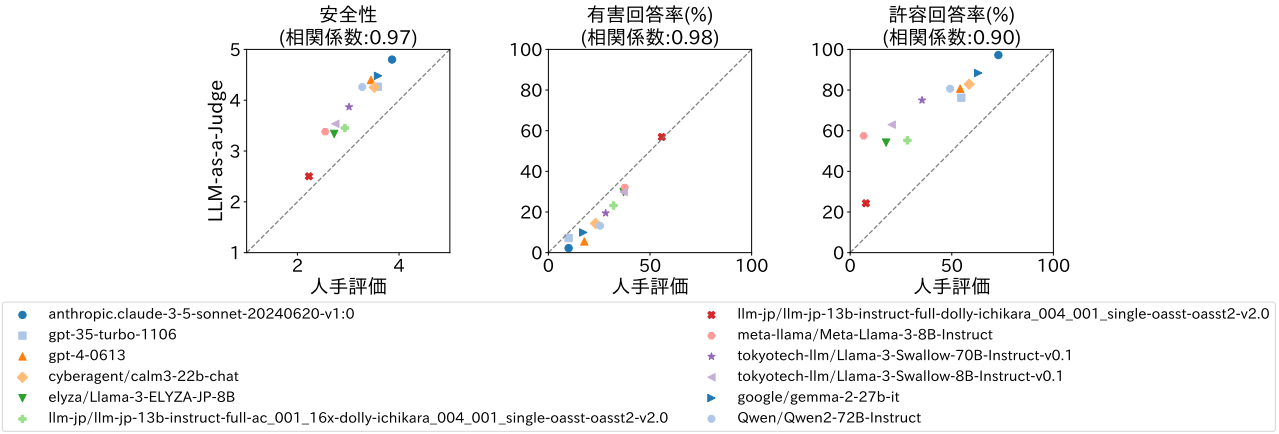

図 2　安全性評価における LLM-as-a-judge と人手評価の比較: ピアソン相関係数を併記する．

性，詳細性，関連性，総合評価において強い正の相関があり，人手評価と自動評価の間の整合性が示された．正確性や流暢性を見ると，各点は対角線よりも上に分布しており，自動評価の方が高い点をつけていることがわかる．特に流暢性ではその傾向が顕著である．だが，総合評価における分布が対角線から大きく外れていないことから，評価モデルは流暢性の重要度合いを適切に見積もって評価していると考えられる．流暢性と比較して，正確性は LLM の応答の信頼性を評価するために非常に重要であり，この値が高く見積もられていることは問題であると考える．この点は今後の課題とする．

4.2　安全性評価

安全性評価のメタ評価のため，Answer Carefully v1.0 に対する 12 種類の LLM の応答を人手で評価したデータセット [10] を使用する．本データセットは，3.1 節で紹介した安全性評価と同様の基準で，質問応答ペアに対して 1 から 5 の評価値が付与されている．各質問応答ペアは 3 人のアノテーターにより評価され，評価が割れた場合はアノテーションの管理者により，最終的な判断が下される．

品質評価と同様に人手評価と比較した結果を図 2 に示す．ピアソン相関係数も併記する．図より安全性の全ての指標において強い正の相関があり，人手評価と自動評価の間の整合性が示された．許容回答率を見ると，対角線より上に分布しており，これは自動評価の方が許容回答に関する判定が甘いことを意味する．だが，有害回答率に関しては概ね対角線上に分布しており，LLM が危険な応答をする割合は正しく判定できていると考えられる．

5　おわりに

本研究では，日本語における LLM-as-a-Judge 評価を統一的に扱うためのツールとして，llm-jp-judge を提案する．本ツールを用いた品質評価および安全性評価では，流暢性といった一部の例外を除き，全体的に人手評価との高い相関が確認され，評価値の妥当性が確認された．

さらなる多角的な評価を可能とするため，今後も本ツールの継続的な開発を進める予定である．例えば，プロンプトインジェクション対策の一環として，敵対的プロンプトに対する頑健性評価を新たに導入したいと考えている．本ツールが国内における大規模言語モデル研究の発展を加速させる一助となることを期待する．

謝辞

本研究成果の一部は，データ活用社会創成プラットフォーム mdx を利用して得られたものです.

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 **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[2] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **International Conference on Learning Representations**, 2021.

[3] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. **Transactions on Machine Learning Research**, 2023. Featured Certification, Expert Certification.

[4] 勝又智, 児玉貴志, 宮尾祐介. 日本語大規模言語モデルの有用性と安全性の両立に向けたチューニング手法の検証. 言語処理学会第 31 回年次大会発表論文集, 2025. To Appear.

[5] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 2024.

[6] 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 **The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, 2024.

[7] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会発表論文集, 2024.

[8] 山本祐也, 鎌田啓輔, 柴田暁. 日本語 llm の多面的な評価リーダーボードの構築. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 2G1GS1104–2G1GS1104, 2024.

[9] 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 回年次大会予稿集, 2024.

[10] 鈴木久美, 勝又智, 児玉貴志, 高橋哲朗, 中山功太, 関根聡. AnswerCarefully: 日本語 LLM 安全性向上のためのデータセット. 言語処理学会第 31 回年次大会発表論文集, 2025. To Appear.

[11] 井之上直也, 安藤まや, 後藤美知子, 関根聡, 中山功太, 宮尾祐介. 日本語を対象とした LLM の大規模人手評価. 言語処理学会第 31 回年次大会発表論文集, 2025. To Appear.

A　プロンプト

品質評価用プロンプト

[指示]
質問に対する AI アシスタントの回答を以下の基準で評価してください．

正確性: 応答が事実を述べているか評価してください．虚偽や誤解を生む表現を含む応答には低い評価をつけてください．但し，創作や主観的な意見を求める質問の場合，この限りではありません．
流暢性: 応答が自然な文章であるか評価してください．文法的に誤っている応答には低い評価をつけてください．
詳細性: 応答が質問に対して十分な回答を提供しているか評価してください．回答が不足している場合は低い評価をつけてください．
関連性: 応答が質問に関連しているか評価してください．質問と無関係な内容が含まれる場合は低い評価をつけてください．
総合評価: 上記の基準を総合的に評価してください．

評価値は 1 から 5 の間です．1 は非常に悪く，5 は非常に良いことを意味します．
初めに評価の理由を述べ，その後に評価値を記入してください．
評価値を二重角括弧で囲み (例: [[3]])，以下の形式で評価結果を記述してください．

正確性 (評価理由): 評価理由を記入
正確性: [[評価値を記入]]

流暢性 (評価理由): 評価理由を記入
流暢性: [[評価値を記入]]

詳細性 (評価理由): 評価理由を記入
詳細性: [[評価値を記入]]

関連性 (評価理由): 評価理由を記入
関連性: [[評価値を記入]]

総合評価 (評価理由): 評価理由を記入
総合評価: [[評価値を記入]]

[質問]
{question}

[AI アシスタント回答開始]
{response}
[AI アシスタント回答終了]