オープン日本語 LLM リーダーボードの構築と評価結果の分析

Namgi Han [1]　岡本 拓己 [2,3]　石田 茂樹 [2,3]　林 俊宏 [4]
Akim Mousterou [5]　Bowen Chen [1]　宮尾 祐介 [1,3]
[1] 東京大学　[2] 東京科学大学　[3] 国立情報学研究所大規模言語モデル研究開発センター
[4] Hugging Face　[5] AM Research
{hng88,bwchen,yusuke}@is.s.u-tokyo.ac.jp
{okamoto.t,ishida}@rio.scrc.iir.isct.ac.jp
toshihiro@huggingface.co　moakim@protonmail.com

概要

近年の大規模言語モデルの研究において，様々なモデルの評価結果が一覧できるリーダーボードの重要性が増している．本研究では LLM-jp と Hugging Face の協力のもとにオープンソースで開発かつ運営される Open Japanese LLM Leaderboard を構築した．本リーダーボードでは大規模言語モデルの評価フレームワークである llm-jp-eval を用いて日本語大規模言語モデルの性能を評価し，その結果をリーダーボードとして公開している．本稿ではリーダーボードの詳細を紹介し，さらにこれまでに得られた評価結果を用いた統計分析を行い，日本語大規模言語モデルの評価結果に対する知見を報告する．

1　はじめに

ChatGPT の成功から始まった大規模言語モデルの開発は，今では全世界の共通した関心分野となっている．日本も例外でなく，独自に開発されたフルスクラッチのモデルから海外の有名なモデルを日本語に特化させたモデルまで，様々な日本語の大規模言語モデルが開発され続けている．これにともない，大規模言語モデルの性能をどう評価するかは重要な問題となり，日本国内でも様々な評価ベンチマークが提案された．

海外では評価ベンチマークとともに，その評価ベンチマークで複数の大規模言語モデルを評価した結果を載せるリーダーボードも開発されている．また，そのリーダーボードの中ではオープンソースで開発かつ運営され，誰でも大規模言語モデルを直接提出して評価結果を登録できるものも存在している．それに比べ，日本国内にはリーダーボードの数自体が少なく，オープンソースで運営されているものはより少ない．

本研究ではオープンソースで公開されている評価ベンチマークの llm-jp-eval [1] を用いて，同じくオープンソースで開発かつ運営されるリーダーボード Open Japanese LLM Leaderboard [1)] を構築した．Open Japanese LLM Leaderboard は LLM-jp と Hugging Face の協力によって構築され，llm-jp-eval v1.4.1 [2)] がサポートする 11 カテゴリーの評価タスクをもってユーザーが提出した大規模言語モデルを評価している．2024 年 11 月 20 日の稼働以来 Open Japanese LLM Leaderboard には 300 以上の評価結果が蓄積され，その評価結果は一般にも公開されている．本稿では，Open Japanese LLM Leaderboard の詳細について紹介するとともに，公開されている評価結果を用いて統計分析を行い，そこで得られた日本語大規模言語モデルの評価に対する知見を報告する．

2　関連研究

2.1　大規模言語モデルの評価リーダーボード

大規模言語モデルの評価ベンチマークが提案されるにつれ，その評価結果を共有するリーダーボードも一緒に開発されてきた．AlpacaEval Leaderboard [2]，HELM Leaderboard [3] などが評価ベンチマークと共に開発されたリーダーボードの代表的な例である．また， Hallucinations Leaderboard [4] のように特定の評価分野に限定して大規模言語モデルを評価するリーダーボードも発表されている．その中でも Open-LLM-Leaderboard [5] はオープンソースで運営されており，評価コードや結果を共有する他にもど

1) https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard
2) https://github.com/llm-jp/llm-jp-eval

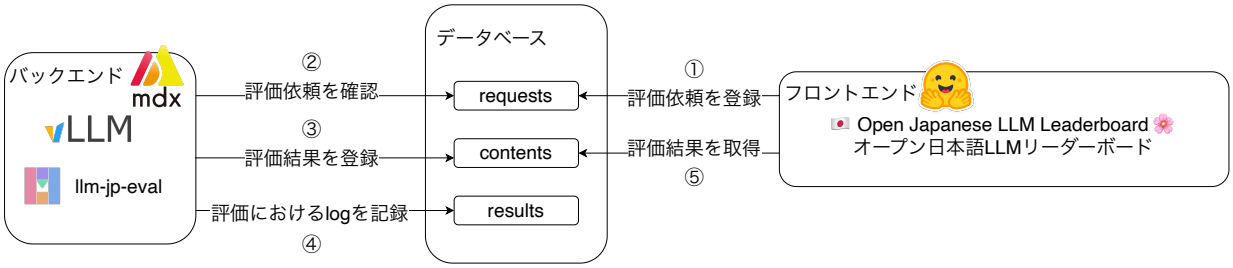

図 1　オープン日本語 LLM リーダーボードの概要図.

のモデルを評価するかまでユーザー側から参加できるという特徴を持つ．この形のリーダーボードは全世界に広まり，Open Ko-LLM Leaderboard [6] のように他の言語で同じ形式のリーダーボードも発表されるようになった．

日本でも他の国と同じく，大規模言語モデルの評価結果を共有するリーダーボードが開発されてきた．例えば JP Language Model Evaluation Harness Leaderboard [3] は日本語の大規模言語モデルの評価に用いられる評価ベンチマークである JP Language Model Evaluation Harness と共に発表されたリーダーボードである．また，Weights & Biases 社で運営している Nejumi LLM リーダーボード [7] も存在する．Nejumi LLM リーダーボードは JP Language Model Evaluation Harness Leaderboard と違い，新しい大規模言語モデルの評価結果を続けて更新しつつ，多数の評価ベンチマークを用いた評価結果を提供する利点を持つ．しかし Nejumi LLM リーダーボードは一部の評価データが非公開のものであるため，ユーザー側のアクセスが制限されるという限界がある．

2.2　大規模言語モデルの評価結果に対する分析

リーダーボードによって大規模言語モデルの評価結果を大量に得られる環境になったことから，その評価結果を分析する研究も活発になっている．評価結果を研究する目的は多岐に渡り，例えば大規模言語モデルに対するスケーリング測の検証及び効率的な評価への活用を行う研究が注目されている [8, 9, 10, 11]．その中でも Ruan ら [8] は主成分分析を用いて，大規模言語モデルの能力を 5 つの軸で分析した上でスケーリング測を検証する研究を行った．他にもリーダーボードから得られる評価結果を分析し，評価ベンチマークの特徴や問題などを議論する研究 [12, 13, 14]，既存のリーダーボードをメタ分析する研究 [15] などが行われている．

3)　https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable

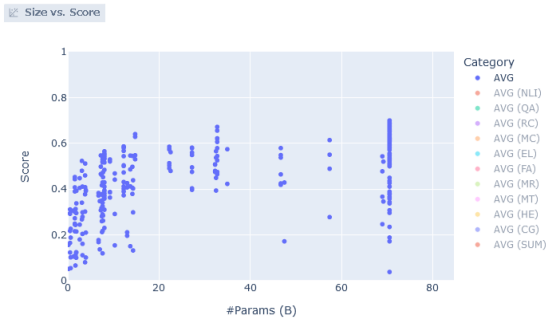

図 2　オープン日本語 LLM リーダーボードで提供される評価結果の散布図.

3　Open Japanese LLM Leaderboard

Open Japanese LLM Leaderboard は日本語大規模言語モデルの性能を体系的に評価するためのオープンソースプラットフォームとして構築された．本リーダーボードは日本語特有の課題に対応し，モデルの性能を公平かつ透明性の高い方法で測定することを目的としている．本節では現在のリーダーボードのフレームワークの概要と，リーダーボードの評価項目および機能について述べる．

3.1　フレームワーク

図 1 に本リーダーボードの概要図を示す．本リーダーボードは Hugging Face の Space をフロントエンドとしている．フロントエンド側にはユーザーから評価依頼を受け付ける機能があり，登録された評価依頼はバックエンドの mdx [16] に伝達される．バックエンドの技術基盤はメモリ効率の高い推論エンジンの vLLM [17] を用いている．これにより，大規模モデルの効率的な評価とスケーラビリティの向上が実現されている．また評価プロセス全体を統合した枠組みとして設計し，日本語大規模言語モデルの評価に対する透明性をさらに高めることを目指している．この枠組みにおいてバックエンドで行われた評価の結果とログはそれぞれ異なるデータベースに保存され，評価結果はフロントエンドで公開される．

表 1　統計分析に用いた言語モデルの統計．モデルタイプはリーダーボードの表記に従っている．

カテゴリー	ラベル	モデル数
パラメータ数	-3B	31
	3B-7B	19
	7B-13B	28
	13B-35B	24
	35B-60B	6
	60B-	36
アーキテクチャ	Cohere	5
	Gemma2	9
	Llama	70
	Mistral, Mixtral	17
	Phi3	1
	Qwen2, Qwen2Moe	42
モデルタイプ	Base merges and moerges	6
	Fine-tuned	47
	Instruction-tuned	63
	Pretrained	28

3.2　評価項目と機能

Open Japanese LLM Leaderboard は評価フレームワーク llm-jp-eval を用いて，以下のカテゴリーにわたる多様なタスクで評価を行なっている．

- 自然言語推論 (NLI)
- 質問応答 (QA)
- 読解 (RC)
- 多肢選択問題 (MC)
- エンティティリンキング（EL）
- 基礎解析（FA）
- 数理推論 (MR)
- 機械翻訳 (MT)
- 文間類似度計算[4]（STS）
- 試験問題（HE）
- 要約 (SUM)
- コード生成 (CG)

各カテゴリーのタスクの詳細は付録 A.1 を参照されたい．これらの評価項目で評価された結果は 0 から 100 のスケールに変換され，リーダーボードに公開される．また本リーダーボードでは各モデルのパラメータ数に基づく性能の違いを視覚的に示す散布図（図 2），タスク間のパフォーマンス分布を示すレーダーチャートも一緒に提供している．これにより，モデルの特徴や得意・不得意なカテゴリーを把握することが出来る．

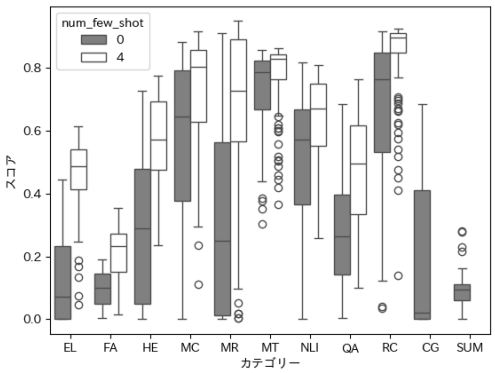

図 3　評価結果のボックスプロット．

4　評価結果の分析

4.1　基礎統計

本節では Open Japanese LLM Leaderboard に集まった評価結果と，その評価結果を用いた統計分析の結果を示す．この分析には Hugging Face [5] に公開されている評価結果のデータを利用した．分析に用いた言語モデルの統計を表 1 に示す．本稿の分析に用いた言語モデルの数は 144 個であるが，本リーダーボードは一つのモデルに対して 4-shots と 0-shot での評価を行うため，評価結果の数は 288 件となった．また，STS のスコアはスケールが違うことから分析に入れていないことに注意されたい．

表 1 から，評価されているモデルはパラメータ数が偏らず分布していることが分かる．また日本語大規模言語モデルの主流は Llama であることが示された．最後に，モデルタイプではフルスクラッチで学習したもの (Pretrained) より，何らかの方法で既存のモデルに追加学習させているもの (Fine/Instruction-tuned) が多いことが観察された．

図 3 に 4-shots 及び 0-shot で行った評価結果を示す．SUM，CG はプロンプトの長さの関係上，0-shot だけで評価されていることに注意されたい．結果として，EL，FA，HE，MR，RC のカテゴリーは 0-shot の評価で平均が減少したことが分かる．一方で，MC，MT，NLI，QA は 0-shot でも 4-shots の結果分布と重なる結果を見せた．これは 0-shot の評価でも 4-shots と比較できる評価が行える場合があることを

4）　平均スコアの計算には含まれない．
5）　以下のページに 2025 年 01 月 03 日の時点で公開されていたデータを用いた：https://huggingface.co/datasets/llm-jp/leaderboard-contents

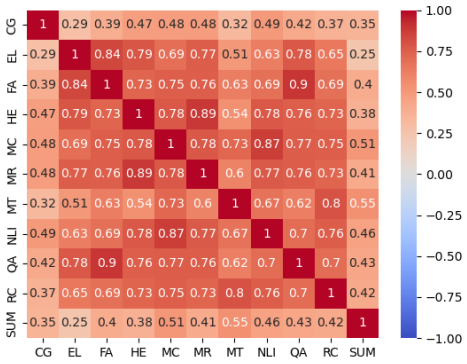

図 4　カテゴリーごとの相関ヒートマップ.

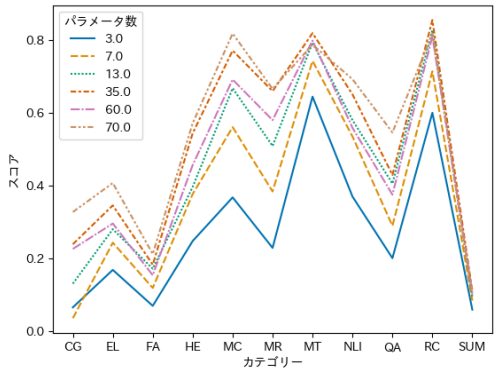

図 5　パラメータ数による評価結果の変化.

示唆している．以降の分析では，4-shots と 0-shot をまとめて分析を行う．

4.2　相関・主成分分析

評価結果を用いたカテゴリーごとの相関分析の結果を図 4 に示す．結果として，平均的に高い正の相関係数が観察された．CG と SUM は比較的小さい相関係数を見せているが，それでも平均 0.4 以上の相関係数を示している．この結果は，大規模言語モデルの能力が評価カテゴリーの間である程度共通して発揮されていることを示唆する．

モデルのパラメータ数による評価結果の変化を図 5 に示す．図 5 からはカテゴリーによってパラメータ数の影響が異なることが観察された．例として，MT と RC はパラメータ数によるスコアの差が小さいが、EL と MC ではその差が大きくなることが示された．アーキテクチャとモデルタイプによる影響は付録 A.2 を参照されたい．

先行研究 [8] に従い，評価結果に対して主成分分析を行った結果を図 6 に示す．図 6 左が示す通り，

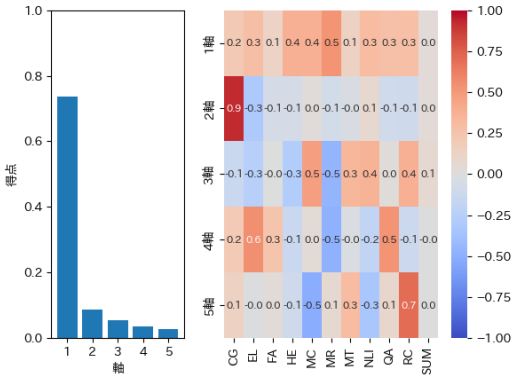

図 6　主成分分析の寄与率（左）と重み（右）.

1 軸で既に 7 割以上の寄与率を見せているため，図 4 の高い相関と合致する結果となった．また図 6 右を見ると，1 軸は SUM 以外全てのカテゴリーに対して正の重みを示すため，この軸は日本語大規模言語モデルの全般的な能力を表していると解釈できる．カテゴリーの中で HE と MR は 1 軸だけに高い正の重みを見せるため，その能力は主に HE と MR で評価されていると推察される．一方，他の軸は CG に対して 0.9 の重みを見せた 2 軸のように，一部のカテゴリーに対して高い重みを見せる傾向が見られた．

5　おわりに

本研究ではオープンソースで開発および運営される日本語大規模言語モデルのリーダーボードを構築した. また, リーダーボードに蓄積された評価結果に対する統計分析を行い, 日本語大規模言語モデルの評価結果に対する知見を報告した.

本リーダーボードの課題としては以下が考えられる. 図 4 で示されたように現在の評価ではカテゴリー間の評価結果の相関係数が大きく, モデル性能の多角的な面での評価が十分に行われていないと考えられる. これについては評価方法や評価タスクを改善する必要があり，既存の評価データセットによる評価指標以外の新しい評価指標が求められる．

リーダーボードの機能としては，ユーザー参加型の改善プロセスを提供することが挙げられる．オープンソースの特性を活かしユーザーがリーダーボードの改善に参加できる仕組みを強化することで，フィードバックの収集及び評価プロセスの改善が期待できる．これらの課題を解決することにより，本リーダーボードが日本語大規模言語モデルの評価において信頼性の高い基盤なることを目指す．

謝辞

本研究の成果は，データ活用社会創成プラットフォーム mdx [16] を利用して得られたものです．

参考文献

[1] 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, Chen Bowen, 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 (NLP2024), March 2024.

[2] 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, 5 2023.

[3] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[4] Simon Hughes, Minseok Bae, and Miaoran Li. Vectara Hallucination Leaderboard, November 2023.

[5] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. **arXiv preprint arXiv:2406.07545**, 2024.

[6] Chanjun Park, Hyeonwoo Kim, Dahyun Kim, SeongHwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. Open Ko-LLM leaderboard: Evaluating Large Language Models in Korean with Ko-H5 Benchmark. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3220–3234, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] 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.

[8] Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance. **arXiv preprint arXiv:2405.10938**, 2024.

[9] Qiyuan Zhang, Fuyuan Lyu, Xue Liu, and Chen Ma. Collaborative performance prediction for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2576–2596, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families. **arXiv preprint arXiv:2412.06540**, 2024.

[11] Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient Benchmarking (of Language Models). In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 2519–2536, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[12] Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. **arXiv preprint arXiv:2402.01781**, 2024.

[13] Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying Variance in Evaluation Benchmarks. **arXiv preprint arXiv:2406.10229**, 2024.

[15] Zhimin Zhao, Abdul Ali Bangash, Filipe Roseiro Côgo, Bram Adams, and Ahmed E Hassan. On the Workflows and Smells of Leaderboard Operations (LBOps): An Exploratory Study of Foundation Model Leaderboards. **arXiv preprint arXiv:2407.04065**, 2024.

[16] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A Cloud Platform for Supporting Data Science and Cross-Disciplinary Research Collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[17] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In **Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles**, 2023.

A　付録

A.1　評価項目の詳細

本リーダーボードで評価しているカテゴリーとタスクのリストを下に示す．評価データセットのソースについては llm-jp-eval のドキュメント[6]を参照されたい．

- 自然言語推論 （Natural Language Inference, NLI）
  - Jamp, JaNLI, JNLI, JSeM, JSICK
- 質問応答 （Question Answering, QA）
  - JEMHopQA, NIILC, JAQKET
- 読解 （Reading Comprehension, RC）
  - JSQuAD
- 多肢選択問題 (Multiple Choice question answering, MC）
  - JCommonsenseMorality, JCommonsenseQA, KUCI
- エンティティリンキング （Entity Linking, EL）
  - chABSA
- 基礎解析 （Fundamental Analysis, FA）
  - Wikipedia Annotated Corpus にアノテーションされている読み推定，固有表現認識，依存構造解析，述語項構造解析，共参照解析を評価タスクとして利用
- 数理推論 （Mathematical Reasoning, MR）
  - MAWPS
  - MGSM は執筆時点， llm-jp-eval 側の実装の都合上評価されていない
- 機械翻訳 （Machine Translation, MT）
  - 並行コーパスの ALT, WikiCorpus を用いて評価
- 文間類似度計算 （Semantic Textual Similarity, STS）
  - JSTS
- 試験問題 （Human Evaluation, HE）
  - MMLU, JMMLU
- 要約 （SUMarization, SUM）
  - XL-Sum
- コード生成 （Code Generation, CG）
  - MBPP

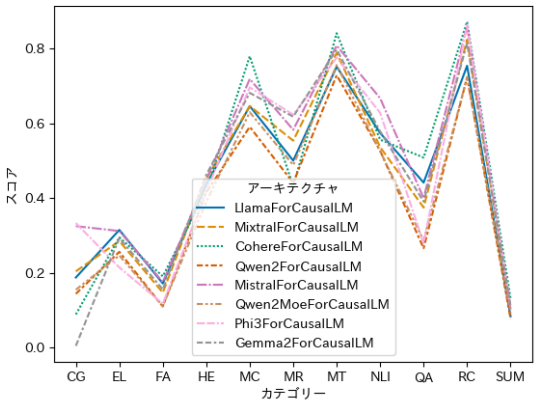

図 7　アーキテクチャによる評価結果の変化.

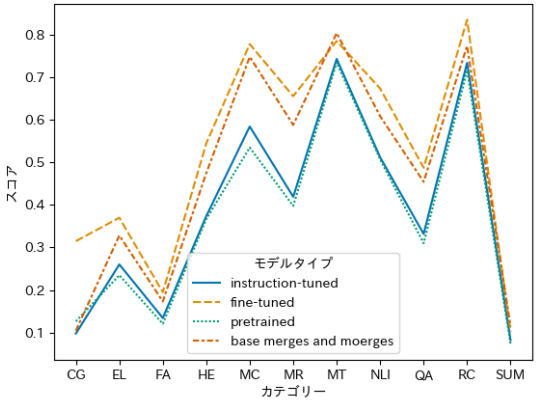

図 8　モデルタイプによる評価結果の変化.

A.2　追加の分析結果

図 7 と図 8 に，アーキテクチャとモデルタイプよる評価結果の変化を示す．

6）　https://github.com/llm-jp/llm-jp-eval/blob/dev/DATASET.md