# 日本語大規模言語モデルの事前訓練過程における下流タスク性能の網羅的な分析

西田 悠人 [1,3]　小田 悠介 [1,3]　Namgi Han[2]　高木 優 [3]　宮尾 祐介 [2,3]
[1] 奈良先端科学技術大学院大学　[2] 東京大学
[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
nishida.yuto.nu8@naist.ac.jp　{odashi,yu-takagi}@nii.ac.jp
{hng88,yusuke}@is.s.u-tokyo.ac.jp

## 概要

大規模言語モデル (LLM) は目覚ましい発展を遂げているが，その高い性能がどのように獲得されていくのかについての理解は不十分である．本稿では，日本語を多く含む大規模テキストを用いて学習された LLM-jp-3 モデルを対象に，様々なモデルサイズ・下流タスクにおけるモデルの学習過程を分析し，日本語 LLM の内部メカニズムの洞察を深める．分析の結果，LLM の学習過程の下流タスクのスコアの軌跡は，タスクの種類によっていくつかの典型的なパターンに分類できることが示唆された．

## 1　はじめに

大規模言語モデル (LLM) は，自然言語生成タスクにおいて目覚ましい成果を挙げている．しかし，LLM の高い性能が事前訓練の過程でいつ達成されるのか，また，モデルサイズや学習データ，タスクの種類による挙動の違いについての理解は，依然として不十分である．これまで，LLM の学習過程についての分析は，主に英語または中国語中心の LLM を対象として取り組まれてきた [1, 2, 3, 4]．Pythia プロジェクト [1] は，14M から 12B までの 10 個のモデルサイズそれぞれについて 154 個のチェックポイントを保存し，それらを通じて LLM の学習過程を詳細に分析する枠組みを提供した．

本稿では，日本語を多く含む大規模コーパスを用いて学習された LLM-jp-3 モデル[1)]を対象に，様々なモデルサイズについて学習過程を分析する．また，従来の取り組みを日本語 LLM で行うだけでなく，LLM の性能評価ツールである llm-jp-eval [5] を活用して広範な下流タスクを分析の対象とし，タスクの種類によって学習過程にどのような特徴が現れるかを観察する．さらに，様々なモデルサイズ・下流タスクにおける学習過程の軌跡の類型化を試み，推論能力の獲得についての洞察を深める．

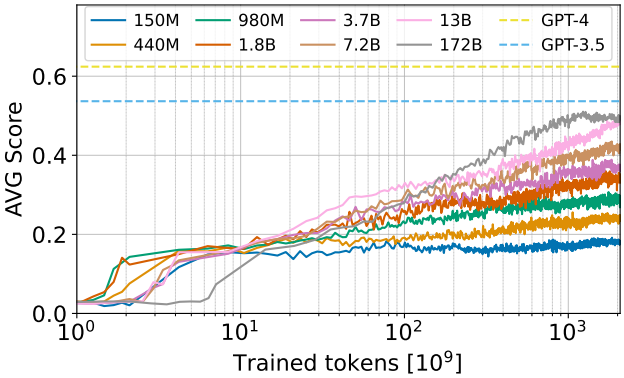


図 1: LLM-jp-3 の学習過程における llm-jp-eval v1.4.1 の平均スコアの推移

分析の結果，LLM の下流タスクにおける学習過程の性能変化は，学習初期から後期にかけて徐々にスコアが向上するもの，学習中期にスコアの向上が停滞するもの，学習後期からスコアが向上し始めるもの，の 3 つの典型的なパターンに分類できることが示唆された．また，LLM の指示追従能力に関する考察から，タスク固有の推論能力においては後者 2 つの類型を同一視できることを示す．

## 2　実験設定

### 2.1　LLM-jp-3 モデルの概要

LLM-jp-3 モデルの事前学習には，日本語を多く含む 1.7T トークンの大規模コーパスである LLM-jp Corpus v3[2)]が用いられている．LLM-jp-3 モデルは

1) https://llmc.nii.ac.jp/topics/post-707/, https://llmc.nii.ac.jp/topics/llm-jp-172b/

2) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

その一部をアップサンプリングした 2.1T トークンにより学習されており，その過程のチェックポイントが数百から一千個程度保存されている．本稿執筆時点では事前学習済みモデルとして 1.8B, 3.7B, 13B, 172B の最終チェックポイントがリリースされており[3]，本稿では LLM-jp モデル学習ワーキンググループの追加実験で得られた 150M, 440M, 980M, 7.2B の 4 つを追加した計 8 つのモデルを対象として分析を行う．なお，指示学習済みモデルについてもリリースされているが，本稿では学習過程に着目し，事前学習済みモデルのみを対象とする．LLM-jp Corpus v3 には llm-jp-eval のスコアを向上させるようなデータ[4]を意図的に含めることはしておらず，本稿で評価する事前学習済みモデルは純粋な言語モデルと見做せる．

付録の表 2 に各モデルにおいて保存されているチェックポイントと学習設定について記載した．

### 2.2 下流タスクの評価

下流タスクの評価には llm-jp-eval v1.4.1[5]を用いた．SUM, CG, STS を除く 9 つのカテゴリについて，step 1 以降の各チェックポイントを対象に 4-shot 推論による評価を行った[6]．各カテゴリに含まれる評価データセットの一覧を付録の表 1 に示した．

## 3 下流タスクにおける性能

下流タスクのスコアをタスクカテゴリ，モデルサイズ，学習過程の観点から分析する．なお，比較対象として，GPT-3.5 (`gpt-35-turbo-16k-0613`) と GPT-4 (`gpt-4-0613`) の評価結果も付記する．

### 3.1 全体の平均スコア

図 1 にスコアの推移を示す．ほとんどのモデルサイズにおいて，学習が進むにつれてスコアは概ね片対数スケールで線形的に上昇する特徴が観測されたが，172B モデルでは学習がおよそ 1T トークンまで進んだあたりでスコアが頭打ちとなった．それ以外のモデルについては，本実験の範囲内ではモデルサイズを大きくするほど性能が向上する傾向がみられた．また，小さいモデルほど，学習トークン数の基準ではスコアの立ち上がりが早い傾向がみられる．

### 3.2 タスクカテゴリごとのスコア

図 2 に各タスクカテゴリにおけるスコアの推移を示す．なお，多くのタスクでスコアの大幅な振動がみられたため，大局的な推移を観察するために HE, MC, NLI を除いて窓幅 20 の移動平均線を示した[7]．

**EL (Entity Linking)** 与えられた文章から全ての固有表現を抽出し，それぞれに極性を付与するタスク．Set F1 により評価する．150M モデルではほとんど解くことができず，150M と 440M の間には顕著な性能差が見られる．1.8B 以下ではモデルサイズが大きいほど性能が向上するが，それよりも大きいモデルではその関係は破れ，たとえば 13B モデルが 172B モデルよりも優れた性能を示している．

**FA (Fundamental Analysis)** 与えられた文章について共参照解析や依存構造解析などを行うタスク．Set F1 または char F1 により評価する．7.2B モデルまでは，モデルサイズが大きいほど性能が向上する傾向が見られるが，それ以降は停滞する．また，440M と 980M の間でスコアが大きく向上する．

**HE (Human Examination)** 与えられた質問と 4 つの選択肢から適切な回答を選択するタスク．完全一致により評価する．学習初期にスコアが一定値に達した後，いずれのモデルでもスコアの上昇が停滞する．13B 以上のモデルでは学習後期にスコアが再び上昇する．

**MC (Multiple Choice QA)** 与えられた質問と 2~5 個の選択肢から適切な回答を選択するタスク．完全一致により評価する．HE タスクと同様に，学習初期でスコアが向上した後に停滞し，7.2B 以上のモデルでは学習後期に再びスコアが向上する．7.2B 以上では，モデルサイズが大きいほど最終的なスコアが高くなるが，172B モデルでは学習の最終段階でスコアが下がり始める傾向がみられる．

**MR (Mathematical Reasoning)** 与えられた計算問題に対する数値回答を生成するタスク．完全一致により評価する．980M モデルまではタスクをほとんど解けない．1.8B 以上のモデルでは，100G トークンを学習し終わった付近からスコアが大きく向上する．172B モデルでは，学習最終段階でスコアが頭打ちとなるが，モデルサイズが大きいほど最終的なスコアは高い．

3） https://huggingface.co/collections/llm-jp/llm-jp-3-pre-trained-models-672c6096472b65839d76a1fa
4） たとえば llm-jp-eval の評価形式に基づく指示学習データ．
5） https://github.com/llm-jp/llm-jp-eval
6） なお，全体の平均スコアには SUM の評価も含まれているが，全てのモデルでスコアがほとんど 0 であるため大局に影響しない．
7） HE, MC, NLI については学習初期に特徴的なスコア変動がみられたため，移動平均を用いずに表示した．

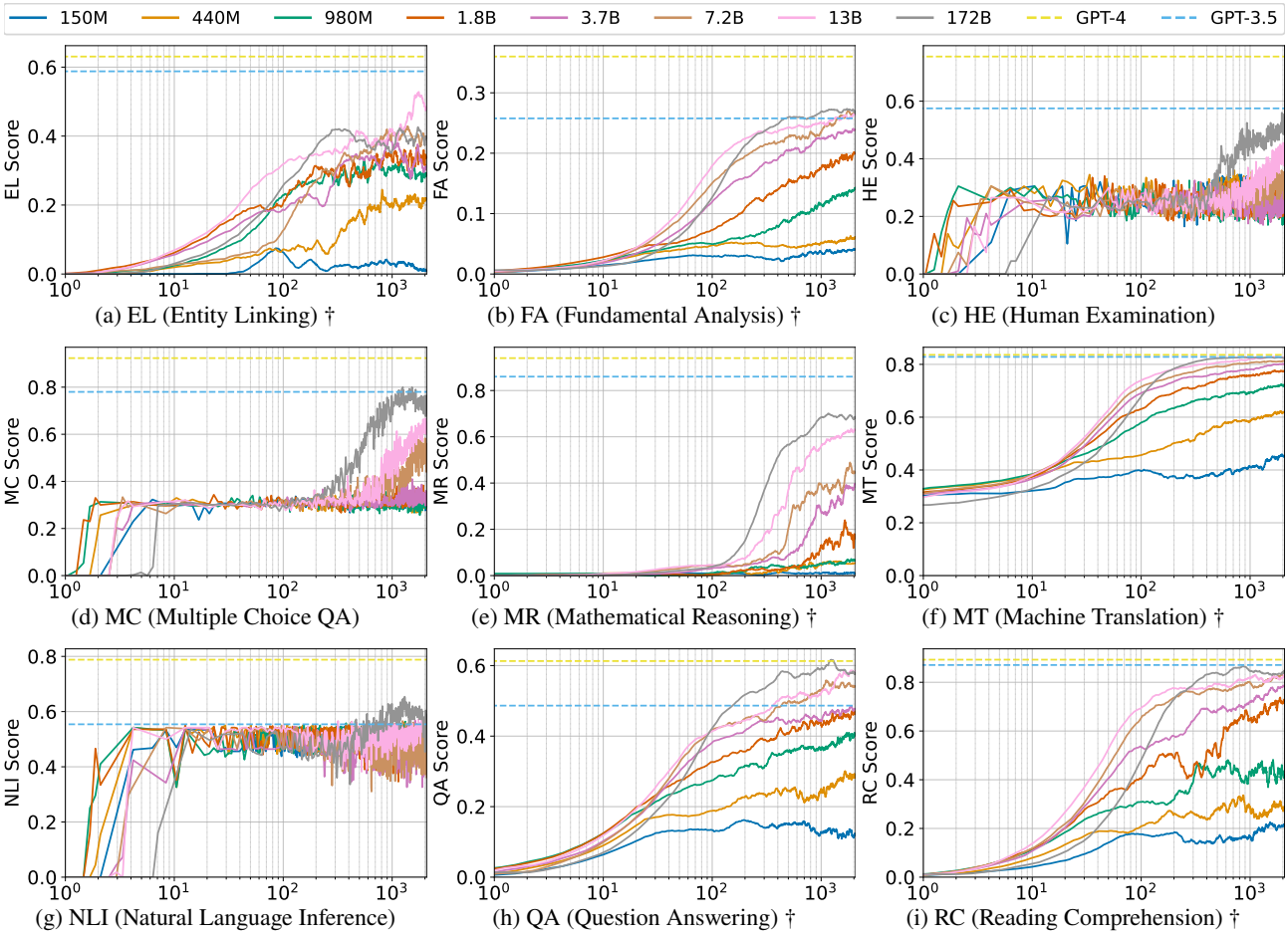


図 2: 各タスクカテゴリにおける下流タスクスコアの推移（† の付くカテゴリでは窓幅 20 の移動平均を表示）

**MT (Machine Translation)** 日英/英日翻訳を行うタスク．COMET [6] により評価する．どのモデルも初期段階から 0.3 程度のスコアを示し，学習全体を通じて少しずつスコアが向上する．1.8B モデルまではモデルサイズの増加による顕著な性能向上がみられるが，それ以上ではスコアが頭打ちとなる．13B モデルや 172B モデルでは，本尺度の上では GPT-3.5 や GPT-4 と同等の性能が達成される．

**NLI (Natural Language Inference)** 与えられた前提と仮説の関係（含意か矛盾かなど）を選択肢の中から回答するタスク．完全一致により評価する．HE や MC タスクと同様に，学習初期でスコアが上昇した後に停滞する．172B モデルでは学習後期に再びスコアが向上するものの，最終段階でスコアが低下する挙動が観測される．

**QA (Question Answering)** 与えられた質問に対して回答を生成する自由記述型の質問応答タスク．char F1 により評価する．150M モデルを除いて学習全体を通じてスコアが緩やかに向上する．13B モデルまではモデルサイズが大きいほど最終的なスコアが高くなる．最高スコアは 13B モデルよりも 172B モデルのほうが高いが，172B モデルは学習の最終段階でスコアが低下する．

**RC (Reading Comprehension)** 質問に対する回答（名詞）を文章から抽出するタスク．char F1 により評価する．7.2B モデルまではモデルサイズが大きいほどスコアが向上し，980M モデルと 1.8B モデルの間に顕著な性能差がみられる．学習中期までスコアが緩やかに向上し，特に大きいモデルでは学習後期にはスコアが停滞する傾向が確認された．

## 3.3 タスクカテゴリの類型

本節では，タスクカテゴリのスコアの推移についての類型化を試みる．図 2 に示したスコア推移の形状から，（学習中期から学習後期を通じてスコアが停滞しないような十分大きいモデルサイズ[8]において）次の 3 つに分類できると考える．1 つ目は，学

8） 本稿の設定では，EL, QA, RC では 440M 以上，FA では 980M 以上，HE, MC では 7.2B 以上，MR では 1.8B 以上，MT では全モデルサイズ，NLI では 172B のモデルが該当する．

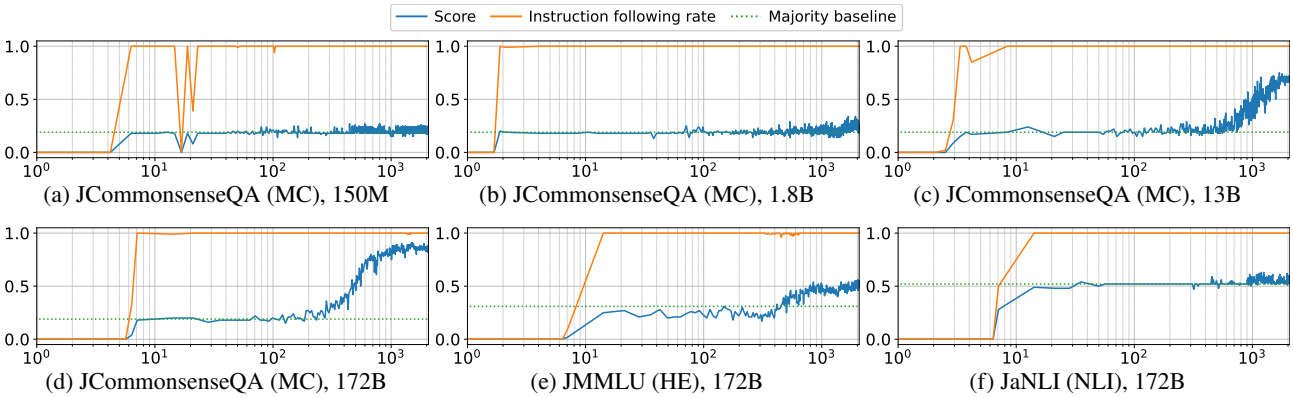


図 3: 指示追従率 (instruction following rate) とスコアの関係

習初期から後期にかけて徐々にスコアが向上するもの（漸進型）であり，EL, FA, MT, QA, RC が該当する．2 つ目は，学習初期に急激にスコアが上昇し，停滞を経たのちに学習後期に再びスコアが徐々に向上し始めるもの（二段階型）であり，HE, MC, NLI が該当する．3 つ目は，学習初期は全くタスクが解けず，学習後期から徐々にスコアが向上し始めるもの（後発型）であり，MR が該当する．

二段階型と後発型は漸進型と異なり，完全一致を評価尺度としている．このため 1 つの事例に対するスコアの変化が急激であり，漸進型のような学習過程全体を通じた緩やかなスコアの上昇が発生しづらいと考えられる．また，二段階型と後発型はタスクが多肢選択式かどうかという相違がある．このように，これらの類型は，タスクや評価尺度の性質で特徴づけられると考えられ，本稿で扱ったタスク以外においても同じ傾向がみられることが予想できる．

### 3.4 指示追従率の分析

二段階型に共通するのは，評価尺度が完全一致の多肢選択式タスクであることであり，タスクへの回答を生成するのに十分な知識がなかったとしても，選択肢のなかから適当に選ぶというナイーブな戦略によって一定程度のスコアに到達可能である．そのため，二段階型の挙動は，学習初期に選択肢に沿って生成できるようになることで一定のスコアに到達し，学習後期に適切な選択肢を選べるような知識を獲得し始めてさらにスコアが向上する，ということによって説明できると考えられる．本仮説を検証するために，モデルの生成した出力の指示追従率の推移を観察する．また，比較対象として，few shot として与えられた事例のなかで最頻の解答を常に出力する majority baseline を導入する．

図 3 に結果を示す．MC カテゴリのタスクのひとつである JCommonsenseQA [7]（4 択の常識推論質問応答）の結果（図 3a~3d）に着目すると，学習初期に指示追従率が 100%まで急上昇すると同時にスコアがベースライン付近まで向上し，停滞を経たのちに大きいモデルでは再びスコアが徐々に向上するという推移を示すことがわかる．172B モデルの結果（図 3d~3f）に着目すると，どのカテゴリにおいても同様の推移が観察される．指示追従率が 100%まで急上昇するタイミングは，モデルサイズやタスクカテゴリによって多少の違いはあるものの，どの場合でも共通して 1B トークンから 10B トークンを学習する間であることが観察された．

したがって，二段階型における LLM の推論結果の学習初期から中期にかけての挙動は，LLM の推論能力の獲得を反映しているわけではなく，指示への追従能力の獲得によってナイーブなベースラインと同程度のスコアを示すようになるというシナリオで説明できる．二段階型のタスクに対する LLM の本来の推論能力は学習後期になって初めて獲得され始めるため，推論能力の推移の観点からは，二段階型の挙動と後発型の挙動は同一視できるといえる．

## 4 おわりに

本稿では，LLM-jp-3 モデルの分析を下流タスク，モデルサイズ，学習過程の観点から網羅的に行った．分析により，下流タスクの種類や評価尺度，モデルサイズによってスコア推移の軌跡が異なり，いくつかの典型的なパターンを示すことが観察された．この知見は LLM は目的とするタスクによって適した訓練計画は異なりうることを示唆するものであり，訓練の条件と下流タスクにおける性能の関係性についての更なる分析は今後の課題である．

# 参考文献


[1] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: a suite for analyzing large language models across training and scaling. In **Proceedings of the 40th International Conference on Machine Learning**, ICML'23. JMLR.org, 2023.

[2] Dirk Groeneveld, Iz Beltagy, Pete Walsh, et al. OLMo: Accelerating the Science of Language Models. **arXiv preprint**, 2024.

[3] DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. **arXiv preprint arXiv:2401.02954**, 2024.

[4] Chen Yang, Junzhuo Li, Xinyao Niu, et al. The fine line: Navigating large language model pretraining with down-streaming capability analysis. **arXiv preprint arXiv:2404.01204**, 2024.

[5] Namgi Han, 植田暢大, 大嶽匡俊ほか. 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, pp. 2085–2089, 2024.

[6] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[7] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[8] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会, pp. 237–240, 2020.

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training Verifiers to Solve Math Word Problems. **arXiv preprint arXiv:2110.14168**, 2021.

[10] Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. Jamp: Controlled Japanese Temporal Inference Dataset for Evaluating Generalization Capacity of Language Models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**, pp. 57–68, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] Hitomi Yanaka and Koji Mineshima. Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference. In **Proceedings of the 2021 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP2021)**, 2021.

[12] 川添愛, 田中リベカ, 峯島宏次, 戸次大介. 形式意味論に基づく含意関係テストセット構築の方法論. 人工知能学会全国大会論文集 第 29 回 (2015), pp. 1K31–1K31. 一般社団法人 人工知能学会, 2015.

[13] Hitomi Yanaka and Koji Mineshima. Compositional Evaluation on Japanese Textual Entailment and Similarity. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1266–1284, 2022.

[14] 石井愛, 井之上直也, 関根聡. 根拠を説明可能な質問応答システムのための日本語マルチホップ QA データセット構築. 言語処理学会第 29 回年次大会論文集, 2023.

[15] 関根聡. 百科事典を対象とした質問応答システムの開発. 言語処理学会第 9 回年次大会, 2003.

[16] 竹下昌志, ジェプカラファウ, 荒木健治. JCommonsenseMorality: 常識道徳の理解度評価用日本語データセット. 言語処理学会第 29 回年次大会, pp. 357–362, 2023. in Japanese.

[17] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. A Method for Building a Commonsense Inference Dataset based on Basic Events. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2450–2460, Online, November 2020. Association for Computational Linguistics.

[18] Takahiro Kubo and Hiroki Nakayama. chABSA: Aspect Based Sentiment Analysis dataset in Japanese, 2018. https://github.com/chakki-works/chABSA-dataset/blob/master/doc/chabsa-aspect-based.pdf.

[19] 萩行正嗣, 河原大輔, 黒橋禎夫. 多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析. 自然言語処理, Vol. 21, No. 2, pp. 213–247, 2014.

[20] 堀尾海斗, 村田栄樹, 王昊ほか. 日本語における Chain-of-Thought プロンプトの検証. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 3T1GS602–3T1GS602, 2023.

[21] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Introducing the Asian Language Treebank (ALT). In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, et al., editors, **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 1574–1578, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[22] 国立研究開発法人情報通信研究機構. Wikipedia 日英京都関連文書対訳コーパス, 2010. https://alaginrc.nict.go.jp/WikiCorpus/.

[23] Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring Massive Multitask Language Understanding. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[24] 尹子旗, 王昊, 堀尾海斗, 河原大輔, 関根聡. プロンプトの丁寧さと大規模言語モデルの性能の関係検証. 言語処理学会第 30 回年次大会, pp. 1803–1808, 2024.

[25] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, et al. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics.

表 1: 本稿で対象とする評価データセット

| タスクカテゴリ | データセット | サブタスク | 評価指標 |
|---|---|---|---|
| Natural Language Inference (NLI) | Jamp [10] | - | 完全一致 |
| | JaNLI [11] | - | 完全一致 |
| | JNLI [7] | - | 完全一致 |
| | JSeM [12] | - | 完全一致 |
| | JSICK [13] | - | 完全一致 |
| Question Answering (QA) | JEMHopQA [14] | - | Char. F1 |
| | NIILC [15] | - | Char. F1 |
| Reading Comprehension (RC) | JSQuAD [7] | - | Char. F1 |
| Multiple Choice question answering (MC) | JCommonsenseMorality [16] | - | 完全一致 |
| | JCommonsenseQA [7] | - | 完全一致 |
| | KUCI [17] | - | 完全一致 |
| Entity Linking (EL) | chABSA [18] | - | Set F1 |
| Fundamental Analysis (FA) | Wikipedia Annotated Corpus [19] | NER | Set F1 |
| | | PAS | Set F1 |
| | | Coreference | Set F1 |
| | | Dependency | Set F1 |
| | | Reading | Char. F1 |
| Mathematical Reasoning (MR) | MAWPS [20] | - | 完全一致 |
| Machine Translation (MT) | ALT [21] | Ja→En | COMET [6] |
| | | En→Ja | COMET |
| | WikiCorpus [22] | Ja→En | COMET |
| | | En→Ja | COMET |
| Human Evaluation (HE) | MMLU [23] | - | 完全一致 |
| | JMMLU [24] | - | 完全一致 |
| Summarization (SUM) | XL-Sum [25] | - | ROUGE |

表 2: LLM-jp-3 モデルの学習設定および保存したチェックポイント

| サイズ | 保存したチェックポイントのステップ数（総数） | バッチサイズ | 最大トークン数 |
|---|---|---|---|
| 150M | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 988000, 988240 (1,017 個) | 512 | 4,096 |
| 440M | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 988000, 988240 (1,017 個) | 512 | 4,096 |
| 980M | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 988000, 988240 (1,017 個) | 512 | 4,096 |
| 1.8B | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 988000, 988240 (1,017 個) | 512 | 4,096 |
| 3.7B | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 494000, 494120 (523 個) | 1,024 | 4,096 |
| 7.3B | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 494000, 494120 (523 個) | 1,024 | 4,096 |
| 13B | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 494000, 494120 (523 個) | 1,024 | 4,096 |
| 172B | 0, 1, ⋯, 9, 10, ⋯, 90, 100, ⋯, 900, 1000, ⋯, 275000, 275500, 275750, ⋯, 292000, 292812 (373 個) | 1,728 | 4,096 |

# A 実験設定の詳細

表 1 に，本稿でモデルの評価に用いたデータセットとその評価指標の一覧を示す．llm-jp-eval の公式ドキュメント[9)]では QA カテゴリの JAQKET [8] と MR カテゴリの MGSM [9] にも対応していると表記されているが，執筆時点での実装の都合上，本稿の評価には含まれていない．

表 2 に，各モデルにおいて保存されているチェックポイントと学習設定について示す．

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