YOMI-Bench：LLM の日本語読み理解に向けた評価用ベンチマーク

三林亮太 [1,2]　高村大也 [2]　谷中瞳 [3,4,5]
[1] 神戸大学　[2] 産業技術総合研究所　[3] 東京大学　[4] 理化学研究所　[5] 東北大学
mibayashi@people.kobe-u.ac.jp　takamura.hiroya@aist.go.jp
hyanaka@is.s.u-tokyo.ac.jp

概要

本研究では，日本語の読み評価に特化した LLM ベンチマークである YOMI-Bench を提案する．日本語における漢字は同じ 1 文字でも読みの候補が複数あり，表層上のテキストからは読みを想起するのが難しい．このような言語的側面から，LLM においても日本語の漢字の読みに関する性能が低いことが経験的に知られている．提案する YOMI-Bench は日本語の読み性能評価に特化した 3 種類のタスクで構成されたベンチマークである．本ベンチマークを用いた評価実験の結果，日本語に特化したローカル LLM であっても読み性能は限定的であることが明らかとなった．さらに，読みを明示的に考慮する生成タスクにおいては，商用モデルにおいても十分な性能が得られていないことが確認された．

1　はじめに

近年，大規模言語モデルの多言語に対する処理能力の分析が多角的に進められている [1]．その中でも，本研究では言語処理能力として「読み能力」に着目する．読み能力は言語ごとに性質が大きく異なり，同一の文字体系を共有する言語間であっても，読みと音の対応関係には顕著な差異が存在する．たとえば，中国語と日本語は多くの漢字を共有しているが，中国語における漢字 1 文字は，約 10%の文字を除いて，ほとんどが 1 つの音にのみ対応している [2]．それに対して，日本語は約 60%の漢字に対して読みが複数存在し，より複雑な情報を必要としている．このように，同じ漢字でも日本語の漢字は読みを推定することが難しく，LLM においても読みに関する性能が低いことが経験的に知られている．同様の理由で，読みを考慮したようなテキスト生成も難しく，例えば押韻のような，同じ音を持つ単語の生成は未だ課題である．

そこで，本研究では日本語の読み理解に着目し，LLM の日本語読み性能を評価できるベンチマークである YOMI-Bench を作成する．本ベンチマークは再現性の確保および研究利用を目的として公開予定である．YOMI-Bench では，例えば「覚醒という単語の読みはなにか？」という問いに対して，「読みはかくせい」のように回答するタスクを含むマルチタスクな評価セットである．また本ベンチマークを用いて，代表的な LLM によるタスクの評価結果を示すことで，日本語における読みが正しく LLM によって理解できているかを定量的に評価する．

2　関連研究

2.1　LLM 評価ベンチマーク

LLM 評価における代表的なベンチマークとして，幅広いトピックを網羅した MMLU [3] があり，これを日本語に適応した JMMLU [4] も提案されている．これらの大規模ベンチマークは，LLM の全体的な性能を評価する上で非常に有用である一方で，言語特有の現象に対する細粒度な評価を行うことは難しい．そのため，本研究では日本語の言語特性に特化した評価用ベンチマークの構築に取り組む．

2.2　日本語の LLM 評価ベンチマーク

日本語を対象とした LLM 評価用ベンチマークはいくつか提案されている [5]．これらの多くは QA タスクを中心としており，Wikipedia や専門的知識に基づく回答能力を評価するものである．代表的な例としては，MMLU を基に，日本語の文化的側面に配慮した翻訳および調整が施された JMMLU [4] が挙げられる．

一方で，一般的な日本語理解を評価するベンチ

表 1　ベンチマークにおける各タスクの例

タスク	プロンプト例	正解
読み推定	覚醒という単語の読みを答えてください	かくせい
押韻選択	覚醒と同じ母音列を持つ文字列を以下から 1 つ選択してください (A) 論文，(B) 学生，(C) 委員会，(D) 博士	(B) 学生
押韻生成	かくせいという単語と母音が完全に一致する単語を答えてください	はつめい

マークとは異なり，読みに関連する能力に着目したベンチマークも一部提案されている．Mizumoto らは，日本語のなぞなぞに特化した NazonazoBench [6] を提案しており，漢字の構造理解を必要とするタスクが含まれている．また，研究論文としては発表されていないものの，かな文字から漢字への変換を扱う AJIMEE-Bench [1] も公開されている．

しかし，これらのデータセットはいずれも，漢字等の読みを明示的に評価する目的としたものではない．非ローマ字言語においては，文脈に応じた読みの変化が理解において重要であるため，読み能力は評価すべき重要な側面である．本研究ではこの点に着目し，日本語の読みに特化した評価ベンチマークの構築に取り組む．

2.3　G2P 評価ベンチマーク

G2P（Grapheme-to-Phoneme）は，与えられたテキストから対応する音韻列を推定するタスクであり，音声合成や音声認識などを主な応用先とする．G2P における代表的な評価ベンチマークとしては，英語を中心とした CMUDictionary [2] を用いたものや，SIGMORPHON [7] を代表とするシェアードタスクにおいて公開されている多言語対応の評価データが挙げられる．これらの研究ではアルファベット表記を前提とした言語にて高い有効性が示されている．

しかし，既存の G2P ベンチマークは，英語のように 1 文字が 1 つの音に対応する言語構造を前提としている場合が多く，日本語や中国語のように 1 文字に対して複数の読みを持つ言語にそのまま適用することは難しい．また，これらのベンチマークは主に音声処理モデルを対象としており，LLM のようなテキスト処理モデルにおける読み能力を直接評価することを目的としていない．そこで本研究では，LLM における日本語の読み評価に特化したベンチマークの構築を行う．

1）https://github.com/azooKey/AJIMEE-Bench
2）http://www.speech.cs.cmu.edu/cgi-bin/cmudict?stress=-s&in=CITE

2.4　LLM 評価ベンチマークのタスク設計

LLM 評価用ベンチマークにおけるタスク設計では，多肢選択形式のタスクが広く用いられている [8]．一方で，多肢選択タスクでは，プロンプト設計に起因する回答バイアスが存在することが指摘されている [9, 10]．そのため，本研究ではプロンプトを複数用意することで，このようなバイアスの影響を受けにくい評価を行う．

3　読み評価ベンチマークの作成

読み評価ベンチマークは，日本語を対象とした読みに関する「読み推定タスク」「押韻選択タスク」「押韻生成タスク」の 3 つのタスクで構成される．それぞれのタスクの詳細を各節で示す．

3.1　読み推定タスク

読み推定タスクは，入力された単語に対する正しい読みを出力するタスクである．例えば，表 1 に示すように，「覚醒」という単語に対して「かくせい」という正しい読みを出力することが求められる．

本タスクでは，漢字 1 文字に対する読みの候補が 1 つのみである漢字で構成された単語（Single）と，読みの候補が 2 つ以上存在する漢字で構成された単語（Multiple）の両方を対象とする．これにより，読みが一意に定まる場合に正しい読みを推定できるかに加えて，複数の読み候補を持つ漢字に対して適切な読みを推定できるかを評価する．

3.2　押韻選択タスク

押韻選択タスクは，入力された単語と同じ母音列を持つ単語を，4 つの選択肢の中から 1 つ選択するタスクである．例えば，表 1 に示すように，「覚醒（kakusei）」という単語に対しては「学生（gakusei）」を選択することが求められる．

本タスクでは，漢字表記の単語とひらがな表記の単語の 2 種類を対象とした．漢字表記の場合には，漢字からひらがなへの読み変換が必要となるのに対

し，ひらがな表記の場合にはその変換を必要としないため，漢字表記より容易なタスクとなっている．

3.3 押韻生成タスク

押韻生成タスクは，入力されたひらがな表記の単語に対して，同一の母音列を持つひらがなの単語を生成するタスクである．例えば，表 1 に示すように，「かくせい（kakusei）」という入力に対しては「はつめい（hatsumei）」のような母音列（auei）が同一である単語を出力することが求められる．

同一の母音列を持つ単語を生成するためには，入力となるひらがなの背後にある読みを正しく推定し，その読みを生成過程において一貫した制約として保持する必要がある．したがって，本タスクは LLM が読み知識を有しているかどうかに加えて，その読み情報を生成時に適切に利用できるかを評価することを目的としている．

4 実験

4.1 漢字データの収集

ベンチマークで対象とする漢字は，日本の文化庁によって公開されている常用漢字表を参考に選択した．常用漢字表には全部で 2,136 文字の漢字が収録されており，読みが 1 つの漢字が 803 文字，読みが 2 つ以上の漢字が 1,333 文字収録されている．このうち，「愛（ai）」という漢字に対して「愛媛（ehime）」のような特殊な読みが網羅されていない漢字が一部存在したため，それら 504 件については除外した．この中から，読みが 1 つの漢字 100 文字と，読みが 2 つ以上の漢字 100 文字をランダムに選択し，合計 200 文字の漢字を対象とした．

4.2 プロンプトの作成

本研究では，単一のプロンプトに依存することによるバイアスを避けるため，各タスクに対して複数種類のプロンプトを作成した．まず，各タスクのベースとなるプロンプトを作成し，その文意を維持したまま ChatGPT を用いて言い換えを行った．この言い換えによって作成されたプロンプト 4 件とベースとなるプロンプト 1 件の合計 5 件のプロンプトを各タスクごとに用意した．実験ではこれら 5 件のプロンプトを用いて評価を行い，得られたスコアの平均を最終的な評価結果として用いた．具体的なプロンプトは付録 A にて示す．

4.3 ベンチマークのベースライン評価

YOMI-Bench のベンチマークとしての難易度および有効性を示すために，ベースライン評価を行った．多言語に対応した一般的なモデルでは，日本語の読み能力を十分に獲得していない可能性があるため，日本語に特化したモデルと，多言語対応モデルの双方を評価対象とした．本研究では，日本語を含む多言語対応のローカル LLM1 件，日本語に特化したローカル LLM4 件，および商用モデル 5 件の合計 10 件のモデルを用いて評価を行った．

読み推定タスクでは，BIG-Bench Extra Hard の実装を参考に，LLM が生成したテキストから，読みに該当する部分のみを正規表現を用いて抽出した．抽出した読みが正解の読みと完全に一致する場合にのみスコアを 1 点とし，データ数で割ることで正解率を算出した．読みが複数存在する漢字については，各漢字を 1 つのグループとして扱い，グループ内の問題に対する正解率の平均を算出し，そのグループスコアを用いて全体の平均正解率を求めた．

押韻選択タスクは 4 択形式の多肢選択問題であるため，LLM の出力結果から (A) から (D) までのいずれかの選択肢を抽出し，正解と一致した場合にスコアを 1 点とした．最終的な正解率は，全データに対する平均として算出した．対象とする単語は，読みが 1 つのみの漢字を含む単語に限定し，正例 1 件と負例 3 件からなるデータを作成した．

押韻生成タスクでは，生成された押韻表現を母音列に変換し，その一致によって評価を行った．まず，生成テキストから押韻に該当する部分を正規表現により抽出し，秋永の変換表 [11] を参考に作成された Mibayashi ら [12] の変換表を用いて母音列へ変換した．生成された母音列が正解の母音列と完全に一致する場合にのみスコアを 1 点とし，事例数で割ることで正解率を算出した．

5 結果と考察

本ベンチマークにおける各タスクの評価結果を表 2 に示し，各タスクにおける詳細な評価結果について，以下の各節で述べる．

5.1 読み推定タスク

読み推定タスクにおいて，読みが単一の漢字を対象とした場合には，日本語特化の LLM および商用 LLM のいずれにおいても，正解率 0.79 以上と高い

表 2 ベンチマークを用いた各モデルの正解率

	読み推定タスク		押韻選択タスク		押韻生成タスク
Models	Single	Multiple	漢字	ひらがな	ひらがな
Ministral-8B-Instruct-2410	0.4679	0.3023	0.2960	0.3400	0.6519
Llama-3.1-Swallow-8B-Instruct-v0.5	0.8219	0.5116	0.2960	0.5020	0.8480
Llama-3-ELYZA-JP-8B	0.7900	0.4380	0.3040	0.5800	0.5439
llm-jp-3-7.2b-instruct	0.7999	0.5174	0.2279	0.2620	0.1660
llm-jp-3-13b-instruct	0.8960	0.5558	0.2780	0.3379	0.0980
claude-sonnet-4-5-20250929	0.9960	0.9216	0.9559	0.9840	0.7280
mistral-medium-2508	0.8019	0.5410	0.3460	0.5740	0.4679
gemini-2.5-flash	0.9820	0.8631	0.9380	0.9480	0.5920
gpt-4o	0.9620	0.6943	0.5820	0.2580	0.5980
gpt-5	0.9840	0.9678	1.0000	0.9980	0.7800

性能が確認された．この結果から，LLM は読みが一意に定まる漢字に対しては，表層的な文字情報や語彙知識に基づいて，概ね正しい読みを生成できていることがわかる．一方で，日本語を含む多言語対応のローカル LLM では，正解率 0.4679 と相対的に低い結果となっており，日本語に特化した学習データ量の違いが，漢字の読み性能に影響している可能性が示唆される．また，同一アーキテクチャでパラメータ数のみが異なる llm-jp シリーズにおいては，13B モデルが 7.2B モデルを上回る性能を示しており，モデル規模の増加が読み推定性能の向上につながる可能性が示された．一方，読みが複数存在する漢字を対象とした場合には，すべてのモデルにおいて正解率が大きく低下する傾向が確認された．この結果は，LLM が単一の正解に対応する読みについては適切に生成できる一方で，複数の読み候補の中から文脈に即した読みを選択する能力は十分に獲得できていないことを示している．すなわち，LLM は漢字の「読みの存在」を知識として保持している可能性があるものの，その曖昧性を解消するための言語的・語彙的手がかりを十分に活用できていないことが示唆される．

5.2 押韻選択タスク

押韻選択タスクにおいては，*mistral-medium-2508* と *gpt-4o* を除く商用モデルは高い正解率を示した一方で，日本語特化のローカルモデルはいずれも性能が低い結果となった．また，漢字とひらがなの表記の差に着目すると，漢字設定よりもひらがな設定の方が高い性能を示す傾向にあった．通常，漢字表記で母音列を推定するためには，まず漢字の読みを推定し，その後に音韻情報を導出する必要があるのに対し，ひらがな表記では読みが明示されているため，押韻判断に必要な音韻情報を取得しやすかったためと考えられる．この点から，ひらがな設定の方がタスクとしての難易度は低く，性能が高くなる結果は妥当であると考えられる．

5.3 押韻生成タスク

押韻生成タスクは，性能がモデルによって大きく分かれる結果となった．具体的な例としては，「やく」という入力に対して「かし」，「せっけん」に対して「けん」といった出力が見られ，母音列がすべて一致するような生成が行われていない．また，母音列が一致する生成についても，商用モデルですら，「せつな」に対して「めぐや」，「きょうくん」に対して「ようりゅん」といった生成が見られ，これらは母音列は一致しているものの，自然な単語とは言えない結果となった．これらの結果から，LLM は読み知識を有している可能性がある一方で，それを生成過程において適切に活用する能力は依然として限定的であることが示唆される．

6 おわりに

本研究では，日本語 LLM の読みに関する能力を評価するためのベンチマーク YOMI-Bench を構築した．本ベンチマークを用いて，代表的な LLM を対象に評価を行った結果，いずれのタスクにおいても，日本語に特化したローカルモデルは読み性能が十分でないことが明らかとなった．また，押韻生成タスクにおいては，商用モデルにおいても性能のばらつきが大きく，読み情報を生成過程で適切に活用することが依然として難しいことが示された．

今後の課題として，タスクとデータ数の拡充や，中国語やアラビア語などの非ローマ字言語を対象に，読み評価の範囲を拡大する予定である．

謝辞

本研究は JSPS 科研費 JP25K03229, JP24K03228, JP25K03228，JST CREST JPMJCR2565，および，国立研究開発法人産業技術総合研究所事業の令和 7 年度覚醒プロジェクトの助成 (FY2025) の助成を受けたものです.

参考文献

[1] Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. Multilingual large language models: A systematic survey. **arXiv preprint arXiv:2411.11072**, 2024.

[2] Kayako Matsuo, Shen-Hsing Annabel Chen, Chih-Wei Hue, Chiao-Yi Wu, Epifanio Bagarinao, Wen-Yih Isaac Tseng, and Toshiharu Nakai. Neural substrates of phonological selection for Japanese character kanji based on fMRI investigations. **NeuroImage**, Vol. 50, pp. 1280–1291, 2010.

[3] 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**, ICLR 2021, pp. 1–27, 2021.

[4] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In James Hale, Kushal Chawla, and Muskan Garg, editors, **Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)**, pp. 9–35, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Koshiro Saito, Sakae Mizuki, Masanari Ohi, Taishi Nakamura, Taihei Shiotani, Koki Maeda, Youmi Ma, Kakeru Hattori, Kazuki Fujii, Takumi Okamoto, Shigeki Ishida, Hiroya Takamura, Rio Yokota, and Naoaki Okazaki. Why we build local large language models: An observational analysis from 35 Japanese and multilingual llms, 2025.

[6] Masaharu Mizumoto, Dat Nguyen, Zhiheng Han, Jiyuan Fang, Heyuan Guan, Xingfu Li, Naoya Shiraishi, Xuyang Tian, Yo Nakawake, and Le Minh Nguyen. The nazonazo benchmark: A cost-effective and extensible test of insight-based reasoning in LLMs. **arXiv preprint arXiv:2509.14704**, 2025.

[7] Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In Mans Hulden and Ryan Cotterell, editors, **Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection**, pp. 1–27, Brussels, October 2018. Association for Computational Linguistics.

[8] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In **Proceedings of the First Conference on Language Modeling**, COLM 2024, pp. 1–31, 2024.

[9] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. LLMs may perform MCQA by selecting the least incorrect option. **arXiv preprint arXiv:2402.01349**, 2024.

[10] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In **International Conference on Representation Learning**, ICLR 2024, pp. 19426–19454, 2024.

[11] 秋永一枝. 日本語の音節 (拍) は幾つか. 講座日本語教育第 5 分冊 早稲田大学語学教育研究所. pp, 11–21, 1969.

[12] Ryota Mibayashi, Takehiro Yamamoto, and Hiroaki Ohshima. Japanese rhyme generation based on mora similarity and generation probability. In **Proceedings of the 27th International Conference on Information Integration and Web Intelligence**, iiWAS 2025, pp. 95–111, 2025.

A　各タスクにおけるプロンプト例

表 3 に，各タスク使用したプロンプトのテンプレート例を示す.

表 3　各タスクにおけるプロンプトテンプレート例

タスク	プロンプトテンプレート例
読み推定	「{word}」という単語の「{char}」という漢字の読みをひらがなで答えて下さい。 「{word}」に含まれる漢字「{char}」の読みを、ひらがなで答えてください。 単語「{word}」の中で使われている「{char}」の読みを、ひらがなで示してください。 「{word}」という語における漢字「{char}」の読みを、ひらがなで記してください。 単語「{word}」中の漢字「{char}」は、ひらがなでどのように読まれますか。
押韻選択	「{word}」という単語と同じ母音で構成される単語を以下から 1 つだけ選択してください。 以下の単語候補のうち、「{word}」と母音構成が一致するものを 1 つだけ選択してください。 以下の選択肢の中から、「{word}」と同一の母音で構成される単語を 1 つ選んでください。 「{word}」と母音の並びが同じ単語を、次の候補から 1 つだけ選択してください。 次に示す候補の中から、「{word}」と同じ母音列を持つ単語を 1 つ選んでください。
押韻生成	「{word}」という単語と同じ母音で構成される単語（韻）を生成してください。 「{word}」と同一の母音列をもつ単語を、ひらがなで 1 語生成してください。 「{word}」と母音の並びが一致する韻語を、ひらがなで 1 つ出力してください。 「{word}」と同じ母音構成になる単語を、ひらがな表記で 1 つ作ってください。 「{word}」と母音列が同一となる単語を 1 つ、ひらがなで生成してください。