YOMI-Bench: A Benchmark for Evaluating Kanji Reading and Phonological Understanding of LLMs for Japanese

Anonymous ACL submission

Abstract

We propose YOMI-Bench, a benchmark for evaluating kanji reading and phonological understanding of large language models (LLMs) for Japanese. In Japanese, a single kanji character often has multiple possible readings, making it difficult to infer the correct reading from surface-level text alone. Due to these linguistic characteristics, it is empirically known that LLMs exhibit low performance in kanji reading for Japanese. The proposed YOMI-Bench consists of four tasks specifically designed to evaluate kanji reading performance in Japanese. In our evaluation using YOMI-Bench, we assessed one multilingual open LLM, four Japanese-specific open LLMs, and five commercial LLMs. As a result, we found that even Japanese-specific models show low performance, and that commercial models also perform poorly on generation tasks that require consideration of kanji readings.

1 Introduction

The multilingual abilities of large language models (LLMs) have been analyzed from various perspectives (Zhu et al., 2024). Among these, we focus on the ability of LLMs to recognize and utilize the phonological readings of text, which is related to tasks such as grapheme-to-phoneme (G2P) estimation and phoneme-aware text generation. The relationship between written forms and their pronunciations varies substantially across languages, and even among languages that share the same writing system, notable differences can be observed. For example, although Chinese and Japanese share many kanji characters, nearly all kanji characters in Chinese correspond to a single pronunciation, except for approximately 10% of cases (Matsuo et al., 2010). In contrast, approximately 60% of Japanese kanji have multiple possible readings (see details in Sec. 3.1) , requiring more complex linguistic information for correct interpretation. For instance,

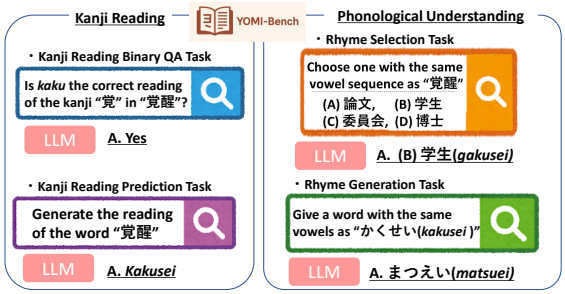

Figure 1: The overview of YOMI-Bench.

the kanji character "覚" has three possible readings: "*kaku*," "*obo*," and "*sa*." These readings vary depending on the word in which the character appears, such as "覚醒 (*kakusei*/Awakening)," "覚える (*oboeru*/Memorize)," and "覚める (*sameru*/Wake up)." Thus, in order to correctly predict the readings of kanji characters appearing within individual words, models should not only possess the knowledge of the multiple readings of each kanji character, but also predict the correct reading based on the word. For practical situations, the ability to correctly understand the readings of kanji is important for LLMs to solve classification and generation tasks that take rhyming into account, such as proofreading and generating lyrics (Potash et al., 2015; Nikolov et al., 2020), rap verses (Xue et al., 2021; Mibayashi et al., 2023), and advertising texts (Lei et al., 2022).

However, the reading ability of LLMs that considers such linguistic characteristics of Japanese remains largely unexplored. Therefore, as illustrated in Figure 1, we construct YOMI-Bench, a benchmark for evaluating Japanese reading performance of LLMs. YOMI-Bench is a multi-task evaluation set involving seven types of binary/multiple QA and text generation tasks. For example, the question "Generate the reading of the word 覚醒" asks LLMs to answer the correct kanji reading as a text generation task, where the gold answer is "*kakusei*." We also use YOMI-Bench to evaluate the reading

abilities of representative LLMs.

The contributions of this study include:

- We construct a challenging benchmark involving seven tasks that requires correct understanding of phonological readings in Japanese and release them on GitHub as publicly available linguistic resources.
- Using the benchmark, we evaluate one multilingual LLM, four Japanese-specific LLMs, and five commercial models, providing baselines for kanji reading performance.

2 Related Work

Grapheme-to-Phoneme (G2P) Benchmarks Grapheme-to-Phoneme (G2P) is a task that estimates a phoneme sequence from a given text. Its main applications include speech synthesis and speech recognition. Representative evaluation benchmarks for G2P include those based on the English-centered CMU Pronouncing Dictionary[1], as well as multilingual evaluation datasets released through shared tasks such as SIGMORPHON (Cotterell et al., 2018). These benchmarks have been widely used to evaluate G2P systems, particularly for languages with alphabetic writing systems.

However, existing benchmarks primarily target languages with alphabetic writing systems, such as English, and are often designed under the assumption that the correspondence between written forms and sounds is relatively regular. Therefore, it is not straightforward to directly apply these evaluation settings to languages such as Japanese and Chinese, where the correspondence between characters and pronunciations is often ambiguous and context-dependent. Furthermore, most existing G2P benchmarks are designed for speech processing models and do not aim to directly evaluate the reading ability of text-based models such as LLMs. To address this limitation, we construct a benchmark for evaluating Japanese reading ability in LLMs.

LLM Benchmarks in Japanese Several LLM evaluation benchmarks targeting Japanese have been released to date (Saito et al., 2025). These benchmarks mainly focus on question-answering (QA) tasks, evaluating responses to content ranging from Wikipedia to more specialized domains. A representative example is JMMLU (Yin et al., 2024), which is based on MMLU and consists of datasets that have been translated and adapted to reflect Japanese cultural contexts.

Task Name	Task Format	Size
Kanji Reading QA (Single)	Binary QA	240
Kanji Reading QA (Multiple)	Binary QA	450
Kanji Reading Prediction (Single)	Generation	120
Kanji Reading Prediction (Multiple)	Generation	120
Rhyme Selection (Kanji)	Multiple QA	120
Rhyme Selection (Hiragana)	Multiple QA	120
Rhyme Generation (Hiragana)	Generation	120

Table 1: YOMI-Bench dataset statistics.

In contrast to benchmarks that evaluate general Japanese language understanding, benchmarks related to reading have also been proposed. The kana-to-kanji conversion benchmark AJIMEE-Bench [2] has not been published as an academic paper, but it proposes tasks that convert kana characters into kanji based on their phonological readings, thereby involving reading-related processing. However, none of these datasets explicitly evaluate reading ability. However, in non-Roman-script languages, pronunciations often vary depending on context, making reading an important aspect of language understanding. Accordingly, we construct a dataset to evaluate Japanese reading ability.

3 YOMI-Bench

YOMI-Bench consists primarily of two types of tasks: tasks that evaluate kanji reading knowledge and tasks that require phonological understanding.

3.1 Kanji Data Collection

The kanji characters targeted in the benchmark are selected based on the Jōyō Kanji List, a set of 2,136 officially designated kanji characters published by the Agency for Cultural Affairs of Japan.[3] The Jōyō Kanji List contains a total of 2,136 kanji characters, of which 803 have a single reading and 1,333 have two or more possible readings. From this list, we randomly sample 100 kanji characters with a single reading and 100 kanji characters with multiple readings.

3.2 Multi-Prompt Evaluation

To mitigate biases arising from reliance on a single prompt (Wang et al., 2024; Zheng et al., 2024), we adopt a multi-prompt evaluation strategy. For each task, we design a base prompt and generate four paraphrased variants using ChatGPT while

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] https://github.com/azooKey/AJIMEE-Bench

[3] https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/joyokanjisakuin/index.html

preserving the original semantic intent. During evaluation, models are tested with five prompts, and the final score is computed as the average of their results.

3.3 YOMI-Bench Tasks

Kanji Reading Binary QA Task In this task, the model determines whether a given reading is a correct pronunciation of a kanji by answering Yes or No. For example, the model is asked a question such as: "Is *kaku* the correct reading of the kanji 覚 in the word 覚醒?" and must respond with either Yes or No. This task evaluates whether an LLM possesses knowledge of kanji readings.

In this task, Yes/No-style prompts are constructed based on the kanji dataset. Negative examples are generated by selecting readings from the set of readings in the kanji dataset that have the same number of readings as the target kanji and assigning them as incorrect readings. In addition, each prompt is evaluated using a 4-shot setting, where two positive examples and two negative examples are provided. The positive and negative examples are constructed by sampling four instances from the test dataset for each.

Kanji Reading Prediction Task The Kanji Reading Prediction Task requires models to output the correct reading for a given Japanese word composed of kanji characters. For example, given the input "Please provide the reading of the word 覚醒," the expected output is *kakusei*.

In this task, the predicted reading is extracted from the LLM output using regular expressions. A score of 1 is assigned if the extracted reading exactly matches the gold-standard reading, and 0 otherwise; accuracy is computed as the average score. For kanji with multiple readings, questions derived from the same kanji are treated as a single group, and the final score is computed by averaging the group-level accuracies.

Rhyme Selection Task The Rhyme Selection Task requires selecting a word that shares the same vowel sequence (i.e., forms a rhyme) with a given input word. For example, given the prompt: "Please select exactly one string from the following options that has the same vowel sequence as 覚醒 (*kakusei**). (A) 論文, (B) 学生, (C) 委員, (D) 博士," the correct answer is "(B) 学生 (*gakusei**)."[4]

[4]The readings of the options are: (A) 論文 (*ronbun** / Paper), (B) 学生 (*gakusei** / Student), (C) 委員会 (*iinkai** / Committee), and (D) 博士 (*hakase** / Doctor).

Since the Rhyme Selection Task is formulated as a multiple-choice problem, we extract one of the options from (A) to (D) from the LLM's output. If the extracted option matches the correct answer, a score of 1 is assigned; otherwise, a score of 0 is given. The final accuracy is calculated by dividing the total score by the number of evaluation instances. The target words are restricted to those containing kanji characters with a single possible reading, and we construct paired data consisting of one positive example and three negative examples.

Rhyme Generation Task The Rhyme Generation Task requires models to generate a hiragana word that shares the same vowel sequence with a given hiragana input word. For example, given the prompt "Please provide a word whose vowel sequence exactly matches that of かくせい (*kakusei*)," the expected output is はつめい (*hatsumei*).

In the Rhyme Generation Task, generated rhymes are evaluated by comparing vowel sequences. The rhyme is extracted from the model output using regular expressions and converted into a vowel sequence using the conversion table proposed by Mibayashi et al. (2025). A score of 1 is assigned if the resulting sequence exactly matches the gold-standard vowel sequence, and 0 otherwise; accuracy is computed as the average score.

4 Baseline Experiments

4.1 Overview

To demonstrate the difficulty and effectiveness of YOMI-Bench as a benchmark, we conduct a baseline evaluation. In this study, we conduct evaluations using a total of ten models: one multilingual open LLM that supports Japanese, four Japanese-specific open LLMs, and five commercial models. The specific models used in the evaluation are as follows: *Ministral-8B-Instruct-2410*, *Llama-3.1-Swallow-8B-Instruct-v0.5* [5], *Llama-3-ELYZA-JP-8B* [6], *llm-jp-3-7.2b-instruct3* [7], and *llm-jp-3-13b-instruct3* [8]. *Ministral-8B-Instruct-2410* is a multilingual model, while the remaining four models are specialized for Japanese. The following five models are used as commercial models in our evaluation: *claude-sonnet-4-5-20250929*, *mistral-medium-2508*, *gemini-2.5-flash*, *gpt-4o*, and *gpt-5*.

[5]https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

[6]https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B

[7]https://huggingface.co/llm-jp/llm-jp-3-7.2b-instruct3

[8]https://huggingface.co/llm-jp/llm-jp-3-13b-instruct3

Models	Kanji Reading QA		Kanji Reading Prediction		Rhyme Selection		Rhyme Generation
	Single	Multiple	Single	Multiple	Kanji	Hiragana	Hiragana
Ministral-8B	0.6640	0.6355	0.4679	0.3023	0.2960	0.3400	0.6519
Llama-3.1-Swallow-8B	0.7620	0.7782	0.8219	0.5116	0.2960	0.5020	0.8480
Llama-3-ELYZA-JP-8B	0.6980	0.6795	0.7900	0.4380	0.3040	0.5800	0.5439
llm-jp-3-7.2b	0.7350	0.6755	0.7999	0.5174	0.2279	0.2620	0.1660
llm-jp-3-13b	0.9620	0.8928	0.8960	0.5558	0.2780	0.3379	0.0980
claude	0.9140	0.7733	0.9960	0.9216	0.9559	0.9840	0.7280
mistral	0.9960	0.8737	0.8019	0.5410	0.3460	0.5740	0.4679
gemini	0.9980	0.9817	0.9820	0.8631	0.9380	0.9480	0.5920
gpt-4o	1.0000	0.9404	0.9620	0.6943	0.5820	0.2580	0.5980
gpt-5	0.9970	0.9848	0.9840	0.9678	1.0000	0.9980	0.7800

Table 2: The average quantitative evaluation results on YOMI-Bench.

The evaluation results for each task in our benchmark are shown in Table 2. Overall, across all tasks, commercial models exhibit higher reading performance than Japanese-specific open models, despite not being explicitly specialized for Japanese.

4.2 Discussion on Each Task

Kanji Reading Binary QA Task As shown in Table 2, all Japanese-specialized models except *llm-jp-3-13b* achieved performance of around 0.7. Most models performed worse under multiple conditions than under a single condition. This suggests that while models memorize single reading for kanji characters to some extent, they have difficulty predicting the correct reading according to the context. Furthermore, the multilingual model also did not achieve high performance. Since its performance is comparable to that of Japanese-specialized models, this result suggests that training with Japanese-specialized corpora alone may not necessarily lead to improved performance on kanji reading tasks.

Kanji Reading Prediction Task For kanji characters with a single reading, both Japanese-specialized LLMs and commercial LLMs showed high accuracy (more than 0.79). These results confirm that LLMs are generally able to generate correct readings for kanji characters with a single reading. In contrast, the multilingual LLM including Japanese showed a relatively low accuracy (0.4679). These results suggest that differences in the amount of Japanese training data may influence kanji reading performance. Furthermore, within the llm-jp series, which consists of models that share the same architecture but differ in parameter size, the 13B model outperformed the 7.2B model, suggesting that increasing the number of parameters may lead to improved model performance.

Rhyme Selection Task In this task, Japanese-specific models showed low performance overall. Notably, even models that achieve high performance on the Kanji Reading task, such as *llm-jp-13b* and *Llama-3.1-Swallow-8B*, exhibit low accuracy on rhyme selection. This pattern is also seen in commercial models such as *mistral* and *gpt-4o*. Although these models appear to know correct kanji readings, they do not perform well in both kanji and hiragana settings, suggesting that knowledge of readings is not necessarily linked to their use.

Rhyme Generation Task Finally, in the Rhyme Generation Task, overall performance is low, and in particular, Japanese-specific open models are found to be almost incapable of generating rhymes. Examples of outputs produced by Japanese-specific open models include generating "かがん (*kagan*)" for the input "かでん (*kaden*)", or "ももる (*momoru*)" for the input "たてこもる (*tatekomoru*)", indicating that these models fail to generate outputs with matching vowel sequences. Even commercial models sometimes generate unnatural words. For example, given the input "たてこもる (*tatekomoru*)", a commercial model generates "かけごろう (*kakegorou*)". Although the vowel sequences match in these cases, the generated outputs cannot be considered natural words.

5 Conclusion

We constructed YOMI-Bench, a benchmark specialized for evaluating kanji reading ability in Japanese LLMs. Baseline experiments showed that even Japanese-specific models show low performance, and that commercial models also perform poorly on generation tasks that require consideration of kanji readings. YOMI-Bench highlights the need for improving the reading-related abilities of LLMs.

Limitations

This study focuses on evaluating the kanji reading ability of large language models in Japanese, and the findings may not directly generalize to other languages. Our benchmark is constructed using the Jōyō Kanji set and common vocabulary, and therefore does not cover rare characters or domain-specific terms.

Ethical Considerations

This work does not involve human subjects, personal data, or sensitive content.

References

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. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27.

Zeyang Lei, Chao Zhang, Xinchao Xu, Wenquan Wu, Zheng yu Niu, Hua Wu, Haifeng Wang, Yi Yang, and Shuanglong Li. 2022. PLATO-Ad: A Unified Advertisement Text Generation Framework with Multi-Task Prompt Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 512–520.

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

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

Ryota Mibayashi, Takehiro Yamamoto, Kosetsu Tsukuda, Kento Watanabe, Tomoyasu Nakano, Masataka Goto, and Hiroaki Ohshima. 2023. Verse Generation by Reverse Generation Considering Rhyme and Answer in Japanese Rap Battles. In *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research*, pages 30–41.

Nikola I. Nikolov, Eetu Malmi, Curtis G. Northcutt, and Lorenzo Parisi. 2020. Rapformer: Conditional Rap Lyrics Generation with Denoising Autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 360–373.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. GhostWriter: Using an LSTM for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924.

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. 2025. Why we build local large language models: An observational analysis from 35 japanese and multilingual llms. *arXiv preprint arXiv:2412.14471*.

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

Liqiang Xue, Kaitao Song, Di Wu, Xu Tan, Ningyu Zhang, Tao Qin, Wentao Zhang, and Tie-Yan Liu. 2021. DeepRapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 69–81.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35.

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

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