JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models

Junfeng Jiang[†] **Jiahao Huang**[†] **Akiko Aizawa**[§†]
[†]The University of Tokyo [§]National Institute of Informatics
jiangjf@is.s.u-tokyo.ac.jp
jiahao-huang@g.ecc.u-tokyo.ac.jp
aizawa@nii.ac.jp

Abstract

Recent developments in Japanese large language models (LLMs) primarily focus on general domains, with fewer advancements in Japanese biomedical LLMs. One obstacle is the absence of a comprehensive, large-scale benchmark for comparison. Furthermore, the resources for evaluating Japanese biomedical LLMs are insufficient. To advance this field, we propose a new benchmark including eight LLMs across four categories and 20 Japanese biomedical datasets across five tasks. Experimental results indicate that: (1) LLMs with a better understanding of Japanese and richer biomedical knowledge achieve better performance in Japanese biomedical tasks, (2) LLMs that are not mainly designed for Japanese biomedical domains can still perform unexpectedly well, and (3) there is still much room for improving the existing LLMs in certain Japanese biomedical tasks. Moreover, we offer insights that could further enhance development in this field. Our evaluation tools tailored to our benchmark as well as the datasets are publicly available to facilitate future research.[1][2]

1 Introduction

Large language models (LLMs) show excellent performances in various tasks in general domains including Question Answering (QA) (Brown, 2020; Taori et al., 2023), Machine Translation (MT) (He et al., 2024), Summarization (Ravaut et al., 2024), Machine Reading Comprehension (MRC) (Zhou et al., 2023), Sentiment Analysis (Zhang et al., 2024), and so on. Some researchers design proper prompts for solving biomedical tasks (Singhal et al., 2023; Liévin et al., 2024; Nori et al., 2023). However, most of the existing LLMs have been pretrained with texts in general domains, lacking domain-specific knowledge. To overcome this challenge, biomedical LLMs are proposed through pretraining on biomedical corpora (Chen et al., 2023; Wu et al., 2024), fine-tuning with instruction data (Han et al., 2023), or reinforcement learning with human feedback (Yang et al., 2024b).

With the chain-of-thought prompting technique, Liévin et al. (2024) have achieved 60.2% accuracy on USMLE-QA (Jin et al., 2021), passing the medical licensing examination in the United States. In the most recent work, with the help of multiple agents, Nori et al. (2023) have achieved 93.06% accuracy on the USMLE-QA dataset, similar to the performance of a human expert. With this series of techniques, biomedical LLMs are greatly promoted in English biomedical tasks. However, biomedical LLMs in other languages still have much room for improvement (e.g., Japanese, Chinese, French, etc.). Besides the relative unpopularity of existing Japanese LLMs, another important obstacle is the lack of a comprehensive benchmark for evaluation and comparison. Therefore, in this paper, we focus on constructing a benchmark for evaluating Japanese biomedical LLMs.

We selected five tasks that are widely used for evaluating LLMs and real-world applications, including multi-choice question-answering (MCQA), named entity recognition (NER), machine translation (MT), document classification (DC), and semantic text similarity (STS). Since there are only a few Japanese biomedical datasets exist and they are generally small (e.g., IgakuQA (Kasai et al., 2023) only has 1,600 samples for testing), to reduce the fluctuation of evaluation results, we translate large-scale and high-quality datasets from other languages (e.g., English) to Japanese, augmenting the scale of our benchmark. Furthermore, in the field of Japanese biomedical LLM, a solid leaderboard is missing. Therefore, we select eight representative models to conduct extensive experiments, providing a standard for comparison. We hope our work

[1]https://huggingface.co/datasets/Coldog2333/JMedBench
[2]https://github.com/nii-nlp/med-eval

Proceedings of the 31st International Conference on Computational Linguistics, pages 5918–5935
January 19–24, 2025. ©2025 Association for Computational Linguistics

can make future comparisons more convenient and fair, promoting the development in this field.

In summary, our contributions are in three folds.

- We construct a large-scale benchmark including 20 Japanese biomedical datasets across five tasks for a comprehensive evaluation.
- We evaluate eight representative models across four categories in our benchmark to provide a standard for future comparison.
- We conduct extensive analysis from aspects of the dataset, model, and prompt template, providing valuable insights for future researchers.

2 Related Works

Benchmarking is essential for the development of a specific field. ImageNet Challenge (Deng et al., 2009) is a famous benchmark in Computer Vision. Many remarkable works on image recognition have been proposed (Krizhevsky et al., 2012; He et al., 2016; Tan, 2019) throughout history and the development has increased rapidly. One reason for this success is the convenience of comparison and evaluation in this field. The GLUE (Wang, 2018) is another famous benchmark for evaluating and analyzing natural language understanding (NLU) systems to promote research in developing general and robust NLU systems. However, these works mainly focus on English tasks, limiting the scope of evaluating other languages like Japanese. Kurihara et al. (2022) constructed the JGLUE from scratch without using any translation, including six datasets, which facilitates the research in Japanese natural language processing (NLP) (Yano et al., 2024; Enomoto et al., 2024; Aizawa et al., 2024).

Considering the wide applications of language models (LMs), researchers are trying to explore LMs' power in biomedical tasks. Gu et al. (2021) collected 13 biomedical NLP datasets in six tasks from different isolated work to form a benchmark called BLURB for evaluating biomedical models. MMLU (Chang et al., 2024) is a benchmark consisting of multiple topics. Specially, it contains some biomedical questions like medical questions at the college level. DrBenchmark (Labrak et al., 2024) is an NLU benchmark for evaluating French biomedical models. However, they are not applicable in Japanese. JMMLU[3] is a translated version of the MMLU. The researchers recruited human translators to check and remove those that were difficult to translate, irrelevant, or inconsistent with the Japanese culture. Recently, Qiu et al. (2024) have proposed a multilingual benchmark with six languages for evaluating medical LMs. These benchmarks reflect some shortages of existing LLMs and provide insights into improving the Japanese biomedical LLMs, but they only focus on the MCQA tasks, which hinders the completeness of the evaluation. Considering these shortages, in this paper, we are dedicated to constructing a large-scale benchmark with diverse tasks for evaluating Japanese biomedical large language models. Table 1 shows a comparison of these benchmarks.

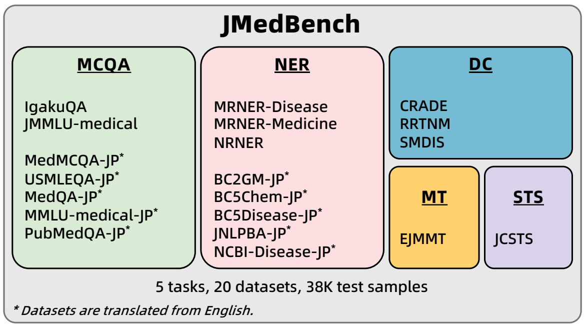

Figure 1: Overview of JMedBench

3 JMedBench

Our benchmark construction consists of two parts. The first part is the dataset collection, while another part is the protocol for evaluation. Firstly, we introduce the rationality of dataset selection and how we augment our benchmark with datasets from other languages. Then, we propose a protocol to obtain robust evaluation results and discuss its necessity for evaluating Japanese biomedical LLMs. Figure 1 is an overview of our benchmark.

3.1 Datasets

In the JMedBench, we include 20 datasets across five tasks containing 38K testing samples. We collect some human-manufactured Japanese datasets, like IgakuQA (Kasai et al., 2023). We also translate some high-quality large-scale English datasets into Japanese to enhance the robustness of JMedBench. Considering the convenience and performance of using OpenAI's API, we use ChatGPT[4] and GPT-4 (Achiam et al., 2023) to create our evaluation datasets when translation is needed. To ensure the quality of the translated testing sets, we use the most powerful model from OpenAI,

[3] https://github.com/nlp-waseda/JMMLU

[4] https://openai.com/index/chatgpt/

Benchmark	Language	Domain	Task MCQA	Task Others	#Dataset	#Sample	Creator
BLURB (Gu et al., 2021)	English	Biomedical	✓	✓	13	65,146	Human
MMLU (Chang et al., 2024)	English	Mixed	✓	✗	1	14,042	Human
JMMLU	Japanese	Mixed	✓	✗	1	7,097	Translation
DrBenchmark (Labrak et al., 2024)	French	Biomedical	✓	✓	20	10,519	Human
MMedBench (Qiu et al., 2024)	Multilingual	Biomedical	✓	✗	6	8,518	Human
JMedBench	Japanese	Biomedical	✓	✓	20	38,130	Mixture

Table 1: Comparison of existing benchmarks.

the GPT-4[5], to perform machine translation. In-context learning is a common practice for adapting an LLM to an unseen task. Therefore, we also translate the training or validation sets to support few-shot evaluation. Due to the limitation of our budgets, we use the cheapest API[6] from OpenAI to translate these samples. Though the translation may not be perfect, producing unfaithful content sometimes, it is good enough to provide information like some domain-specific knowledge and task format during the few-shot evaluation. Previous works (Hendy et al., 2023; Sanz-Valdivieso and López-Arroyo, 2023; AlAfnan, 2024) also have similar findings that ChatGPT has already had a comparable MT performance with specialized Neural Machine Translation systems. Here listed are the involved biomedical tasks and corresponding datasets. Detailed statistics can be found in Table 5 in the Appendix.

- **MCQA** is one of the most commonly used tasks for evaluating LLMs since other tasks can be easily reformulated into the MCQA task. We included IgakuQA (Kasai et al., 2023), JMMLU-medical[7], and translated MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), USMLE-QA, PubMedQA (Jin et al., 2019), and MMLU-medical (Hendrycks et al., 2021b,a).

- **MT** is an important natural language generation (NLG) task. In the biomedical domain, researchers usually need to refer to some English terminologies or communicate with other researchers. Therefore, we expect LLMs can handle cross-lingual tasks besides monolingual tasks. We included the EJMMT (Hayakawa and Arase, 2020) dataset to evaluate the cross-lingual ability of LLMs.

- **NER** is an NLU task aiming to extract named entities like biomedical terminologies, medicines, etc. We included three Japanese medical NER datasets from JMED-LLM[8]: MRNER-disease, MRNER-medicine, and NRNER. To improve the diversity of the dataset, we also follow the BLURB benchmark and include translated BC2GM (Smith et al., 2008), BC5Chem, BC5-Disease (Li et al., 2016), JNLPBA (Collier et al., 2004), and NCBI Disease (Doğan et al., 2014).

- **DC** aims to classify documents into specific categories. We include three datasets from JMED-LLM: CRADE, RRTNM, and SMDIS.

- **STS** is a regression task to compute the semantic similarity between two biomedical sentences. We reformulate it as a classification task to output the discrete level of similarity. We include the JCSTS (Mutinda et al., 2021).

3.2 Evaluation Dataset Augmentation

To enlarge the size of JMedBench for obtaining robust evaluation results, we select several biomedical datasets in English, because of its popularity.

3.2.1 Multi-choice Question-Answering

Different from previous works that usually conduct machine translation at the sentence level, we perform translation at the instance level. Specifically, we translate questions and options meanwhile, so that LLM can understand the scenario better to provide more correct translations. Detailed prompt template can be found in Table 6 in the Appendix.

3.2.2 Named Entity Recognition

We also translate the NER datasets from the BLURB benchmark to improve the amount and diversity of JMedBench. There are three fields in the NER samples: entity type, text, and entities.

[5]We used `gpt-4-0613` checkpoint.
[6]We used `gpt-3.5-turbo-1106` checkpoint.
[7]https://github.com/nlp-waseda/JMMLU
[8]https://github.com/sociocom/JMED-LLM

To ensure the consistency of the translated entity types, we manually translate them into Japanese based on a dictionary (e.g., gene →遺伝子). As for the text and entities, we also perform translation at the instance level, as described in Section 3.2.1. The prompt template for translating the biomedical NER datasets is also shown in Table 7 in the Appendix.

One of the challenges is that the translated entities may not appear in the translated text. To solve this issue, we conduct the translation in two phases: machine translation and manual modification. We first use ChatGPT and GPT-4 to translate the training and testing sets, respectively. We then collect all the invalid samples, mainly due to JSON format error and failure to include the translated entities, and re-translate them using GPT-4. We increase the temperature to 0.5 and call the GPT-4 API again at most 5 times to seek a valid sample. After the machine translation phase, 223 translated entries (0.34%) still remain invalid and then we manually modify these entries to make them valid.

During machine translation, we find that translating entities first instead of text first can reduce about 10% of invalid samples. We speculate that with the entity-first prompt, LLM can refer to the already translated entities when translating the text, thus, the translated entities are usually contained in the following translated sentence. However, since this is not the main focus of this paper, we did not conduct further analysis to verify this hypothesis. We hope this finding can inspire future researchers when performing instance-level machine translation. Though there is a risk of the translation quality from neural translation system (Naraki et al., 2024) and we met a small number of failure cases during the machine translation phase (some bad cases can be found in the Appendix A.3), we realized that the translation quality is still high when we conduct the manual modification, which also reflects the reliability of our data augmentation method.

3.3 Evaluation Protocols

LLMs are usually sensitive to the prompt templates, especially in zero-shot evaluation (Gan and Mori, 2023). To obtain a robust and fair result, we suggest reporting the maximal score of multiple runs using diverse prompt templates for benchmarking. We have also considered computing an average score using different templates, whereas this reported performance may be easily implicated by inappropriate prompts (e.g., using an English-centric prompt for a Japanese-only LLM). In the following evaluation, we use four types of prompt templates:

- **Minimal**: We include information as little as possible in the prompt. For example, for completing the MCQA task, we only input the question, and compute the likelihood of each possible option, namely, `{question}\n`.

- **Standard**: We use commonly used prompt templates in each task. For example, we follow (Robinson and Wingate, 2023) for evaluating MCQA tasks.

- **English-centric**: Some of the existing Japanese LLMs were continually pre-trained from English-centric LLMs. Therefore, we intend to explore whether an English-centric prompt template is beneficial.

- **Instructed**: Besides the standard input, we include a brief task instruction, evaluating the instruction-following ability of LLMs.

As for the MCQA and DC tasks, it is difficult to constrain the auto-regressive LLMs to generate one of the given options or classes. Therefore, we follow Gao et al. (2024) to compute the likelihood perplexity of each possible answer and select the one that has the highest generation possibility as the final answer. We report accuracy on these two tasks. As for the STS task, we also calculate the likelihood perplexity of generating 0-5 and select the one that has the highest generation possibility as the final output. We use the Pearson Correlation as the evaluation metric. As for the MT and NER tasks, we generate the output and compute the BLEU (Papineni et al., 2002) score and entity-level F1 score, respectively.

4 Experiments

4.1 Comparison Methods

In our experiments, we included four categories of popular and excellent LLMs to construct our benchmark, including **general LLMs in other languages**: Llama2 (Touvron et al., 2023), Llama3 (Dubey et al., 2024), Qwen-2 (Yang et al., 2024a), Mistral (Jiang et al., 2023); **biomedical LLM in other languages**: Meditron (Chen et al., 2023); **Japanese general LLMs**: llm-jp (Aizawa et al., 2024), SwallowLM (Fujii et al., 2024); and **Japanese biomedical LLM**: MMed-Llama3 (Qiu et al., 2024). The specific checkpoints are listed

Accuracy (%)	IGA	JMM	MedM	USM	MedQ	MML	Pub	Aver (Micro)
Zero-shot Evaluation								
Llama2-7B	22.69	26.20	30.31	27.81	23.17	29.77	63.50	30.91
Llama3-8B	26.19	35.09	31.94	32.21	26.00	36.77	62.30	34.51
Qwen2-7B	**41.25**	**44.06**	**38.03**	**38.49**	**31.03**	**49.01**	68.90	**42.58**
Mistral-7B	25.19	30.68	30.60	28.44	23.57	32.82	68.80	32.74
Meditron-7B	21.94	25.65	28.31	26.39	21.92	25.65	56.50	28.56
llm-jp-13B	31.00	36.51	30.46	31.66	25.29	35.54	**73.60**	35.17
SwallowLM-7B	27.88	29.50	29.26	27.73	22.39	29.88	<u>70.70</u>	31.86
MMed-Llama3-8B	<u>35.56</u>	<u>37.45</u>	<u>35.43</u>	<u>36.92</u>	<u>29.54</u>	<u>38.86</u>	70.00	<u>38.64</u>
Few-shot Evaluation								
Llama2-7B	23.56	29.35	29.95	29.07	24.43	32.71	55.80	31.28
Llama3-8B	36.31	37.77	36.77	35.04	29.30	43.77	<u>72.50</u>	39.97
Qwen2-7B	**51.75**	**51.61**	**42.74**	**42.42**	**35.51**	**61.04**	<u>72.50</u>	**49.03**
Mistral-7B	30.31	33.60	31.80	29.62	23.96	37.20	72.40	35.07
Meditron-7B	22.31	28.25	28.57	27.73	24.19	28.92	55.80	29.80
llm-jp-13B	36.06	37.37	32.54	33.62	26.32	39.44	**75.90**	37.54
SwallowLM-7B	29.00	33.67	32.23	30.32	23.41	37.89	71.40	35.16
MMed-Llama3-8B	<u>45.37</u>	<u>46.42</u>	<u>38.54</u>	<u>41.95</u>	<u>34.88</u>	<u>50.29</u>	<u>72.50</u>	<u>44.64</u>

Table 2: Benchmark results on Japanese biomedical MCQA tasks, including IgakuQA (**IGA**) and JMMLU-medical (**JMM**), as well as the translated versions of MedMCQA (**MedM**), USMLE-QA (**USM**), MedQA (**MedQ**), MMLU-medical (**MML**), and PubMedQA (**Pub**). We report the highest accuracy among four prompt templates as discussed in Section 3.3. The best and second-best performances are highlighted in bold and underlined, respectively.

in Table 9 in the Appendix. Due to the computation resources, we only evaluate LLMs with around 7 $\sim$ 8B parameters. Llm-jp is a representative LLM that was pre-trained from scratch with Japanese and English texts. Although it does not have the 7B version of the model, we still include the llm-jp with 13B parameters in our benchmark.

4.2 Experimental Results

4.2.1 Multi-choice Question-Answering

Table 2 shows the benchmark results on Japanese biomedical MCQA tasks. Surprisingly, Qwen2 outperforms all models in MCQA, followed by MMed-Llama3. Note that Qwen2 was primarily pre-trained with Chinese and English texts. We hypothesize that one reason for its success is the considerable overlap in tokens between Chinese and Japanese. MMed-Llama3 was continually pre-trained on biomedical texts in multiple languages including Japanese, explaining its superior performance over Llama3. These observations highlight the importance of understanding the Japanese language and injecting domain knowledge. With few-shot demonstrations, all models have improved. We attribute this to the task format (Min et al., 2022) and some domain-specific knowledge provided by the demonstrations. Comparing Llama2 and Llama3, we find that the performance gap under the zero-shot setting is larger than that under the few-shot setting. The additional improvement should be attributed to the improved in-context learning (ICL) ability of Llama3, highlighting the need to enhance the ICL ability of LLMs. Moreover, we can also observe a large improvement from the zero-shot setting to the few-shot setting for Qwen2, showing its superior ICL ability.

Although there is a human-translated version of MMLU-medical, namely, the JMMLU-medical dataset, we still translate the original MMLU-medical dataset using GPT-4 to enrich our benchmark. According to the performances of these two datasets (i.e., JMM & MML in Table 2), the differences between performances on these two datasets do not exceed 5% of accuracy. Furthermore, the ranking of the performances on the translated MMLU-medical dataset also reflects the ranking on the human-translated JMMLU-medical dataset. These observations confirm the quality and the applicability of our translated datasets.

Meditron was continually pre-trained with large-scale English biomedical texts from the Llama2 checkpoint. Chen et al. (2023) showed that Meditron has been successfully shifted to the biomedical domain, outperforming the vanilla Llama2 in various biomedical MCQA tasks. However, we realize that Meditron performs worse than Llama2 in the JMedBench. Such multilingual ability degradation is probably due to the catastrophic forgetting is-

F1-entity (%)	MRD	MRM	NRN	B2G	B5C	B5D	JNL	NCB	Aver (Micro)
Zero-shot Evaluation									
Llama2-7B	0.74	18.99	10.12	32.37	58.74	38.33	7.76	36.21	34.74
Llama3-8B	3.57	18.43	<u>14.97</u>	36.17	58.67	<u>40.91</u>	**24.69**	**52.70**	**40.69**
Qwen2-7B	3.06	15.02	9.54	**39.88**	52.26	38.40	8.51	40.13	35.43
Mistral-7B	**16.75**	**30.21**	11.33	35.61	52.37	38.92	7.12	46.65	34.65
Meditron-7B	1.94	4.78	5.17	15.31	31.12	17.71	12.89	18.29	19.14
llm-jp-13B	<u>8.80</u>	11.99	14.58	29.31	<u>59.15</u>	37.62	22.52	43.55	37.41
SwallowLM-7B	2.20	23.74	11.79	31.18	58.22	**41.76**	13.22	34.85	36.26
MMed-Llama3-8B	3.77	<u>26.85</u>	**17.25**	<u>39.70</u>	**61.85**	39.21	<u>16.48</u>	<u>51.33</u>	<u>40.18</u>
Few-shot Evaluation									
Llama2-7B	11.10	21.14	20.41	46.76	72.95	55.50	47.85	52.90	55.22
Llama3-8B	<u>15.83</u>	<u>37.26</u>	25.15	**51.98**	<u>79.42</u>	<u>63.40</u>	**53.47**	**62.05**	**61.69**
Qwen2-7B	11.65	22.31	24.93	<u>50.59</u>	76.96	55.23	49.54	57.55	57.69
Mistral-7B	15.39	32.50	<u>26.31</u>	48.15	73.06	56.12	48.11	51.33	55.83
Meditron-7B	10.70	18.73	19.13	45.12	68.36	52.05	46.02	52.49	52.47
llm-jp-13B	14.74	22.23	24.64	45.25	76.60	59.79	<u>51.77</u>	56.14	57.76
SwallowLM-7B	12.05	25.58	20.55	44.41	74.74	59.26	46.60	51.03	55.62
MMed-Llama3-8B	**17.27**	**39.47**	**29.09**	49.19	**80.34**	**65.27**	51.05	<u>61.21</u>	<u>61.14</u>

Table 3: Benchmark results on Japanese biomedical NER tasks, including MRNER-Disease (**MRD**), MRNER-Medicine (**MRM**) and NRNER (**NRN**), as well as the translated versions of BC2GM (**B2G**), BC5Chem (**B5C**), BC5Disease (**B5D**), JNLPBA (**JNL**), and NCBI-Disease (**NCB**). We report the highest F1-entity score among four prompt templates as discussed in Section 3.3. The best and second-best performances are highlighted in bold and underlined, respectively.

sue during continual pre-training. How to improve an LLM safely without losing any other ability should be considered in future research. Besides, since the SwallowLM and MMed-Llama3 were continually pre-trained with additional Japanese texts from Llama2 and Llama3, respectively, they are improved by approximately $1\% \sim 5\%$ average accuracy, indicating the importance of local-language adaptation.

4.2.2 Named Entity Recognition

Table 3 shows the results on Japanese biomedical NER datasets. In the few-shot evaluation of BC2GM, BC5Chem, BC5Disease, JNLPBA, and NCBI-Disease datasets, we use three shots of examples. However, for MRNER-Disease, MRNER-Medicine, and NRNER, we only use one shot of example because texts in these datasets are so long that multiple shots will exceed the input token limit of several models.

According to the results, we find that Llama3-8B outperforms other LLMs in both zero-shot and few-shot evaluations, with average F1-entity score of 40.69% and 61.69% respectively. The Japanese biomedical LLM, MMed-Llama3, has the second-best performance in both settings. Few-shot examples can significantly improve the performance of models on the NER tasks, ranging from 19.36% to 33.33% F1-entity improvement. Similar to the observations on MCQA tasks, we believe these examples help LLMs better understand the entity types' definition and output format. Besides, we find that LLMs perform generally worse on datasets including MRNER-Disease, MRNER-Medicine, and NRNER which are derived from JMED-LLM. Note that the average text lengths of datasets from these two sources are 69.82 and 247.81 Japanese characters, while the numbers of entities are 1.33 and 2.66, respectively. Considering the longer input text, larger number of entities and sparser entity distribution, we believe these are the main reasons why the datasets derived from JMED-LLM are more challenging.

4.2.3 Machine Translation

Table 4 shows the BLEU scores for involved comparison methods on EJMMT. MMed-Llama3-8B and Llama3-8B achieve the best and second-best performance in our benchmark under the zero-shot setting. Interestingly, we find that the English-centric models (e.g., Llama2, Mistral) tend to perform better on translating Japanese texts into English, while the Japanese-centric models (e.g., SwallowLM) perform much better in translating English texts into Japanese. We believe the main reason is the text generation ability in different

Metric	EJMMT (en->ja)	EJMMT (ja->en)	Aver	CRADE	RRTNM	SMDIS	Aver (Micro)	JCSTS
	BLUE			Accuracy (%)				Pearson
Zero-shot Evaluation								
Llama2-7B	11.13	14.18	12.65	27.17	37.08	54.76	39.67	-0.005
Llama3-8B	16.79	**23.66**	<u>20.23</u>	25.00	44.94	51.19	40.38	0.422
Qwen2-7B	15.24	19.59	17.41	**35.87**	**59.55**	**58.33**	**51.25**	**0.636**
Mistral-7B	10.93	18.24	14.59	25.00	48.31	54.76	42.69	0.110
Meditron-7B	8.39	7.22	7.81	<u>30.43</u>	52.81	54.76	<u>46.00</u>	0.072
llm-jp-13B	15.14	<u>23.13</u>	19.13	28.26	37.08	51.19	38.84	0.014
SwallowLM-7B	<u>19.32</u>	1.15	10.24	25.00	41.57	50.00	38.86	0.056
MMed-Llama3-8B	**23.00**	17.50	**20.25**	26.09	<u>55.06</u>	<u>55.95</u>	45.70	<u>0.553</u>
Few-shot Evaluation								
Llama2-7B	12.89	20.18	16.54	29.35	44.94	59.52	44.61	0.099
Llama3-8B	20.22	28.50	24.36	34.78	53.93	63.10	50.60	0.483
Qwen2-7B	18.33	25.41	21.87	**44.57**	<u>56.18</u>	**86.90**	**62.55**	**0.625**
Mistral-7B	12.76	23.05	17.91	30.43	<u>56.18</u>	66.67	51.09	0.378
Meditron-7B	11.79	21.67	16.73	26.09	35.96	54.76	38.93	0.067
llm-jp-13B	**27.93**	**28.96**	**28.45**	<u>36.96</u>	46.07	<u>67.86</u>	50.29	0.144
SwallowLM-7B	23.23	23.07	23.15	30.43	44.94	59.52	44.97	0.039
MMed-Llama3-8B	<u>25.56</u>	<u>28.73</u>	<u>27.14</u>	34.78	**57.30**	<u>67.86</u>	<u>53.31</u>	<u>0.515</u>

Table 4: Benchmark results on the rest of other tasks in JMedBench, including Machine Translation (**EJMMT**), Document Classification (**CRADE**, **RRTNM**, **SMDIS**), and Semantic Text Similarity (**JCSTS**). The best and second-best performances are highlighted in bold and underlined, respectively.

languages. Therefore, when applying LLMs to the MT task, we should consider more on the language generation ability instead of the language understanding ability. Although the llm-jp is also a Japanese-centric LLM, according to Aizawa et al. (2024), it was pre-trained with 50-50 Japanese-English mixed data. Therefore, it has a balanced bilingual NLU and NLG ability. Furthermore, with few-shot demonstrations displaying the task format, llm-jp achieves the best performance in the MT task, which shows the prospect of developing Japanese LLMs from scratch instead of continually pre-training from checkpoints in other languages. Besides, comparing Llama2 and the continually pre-trained Meditron and SwallowLM, we find that continually pre-training with texts in biomedical domains or Japanese texts only will lead to forgetting issues. Continual Learning (Wang et al., 2024) is a potential solution,but it is still challenging to continually improve the existing LLMs while maintaining their original ability.

4.2.4 Document Classification

Performances of the DC task are also shown in Table 4. We find that Qwen2 achieves the best performance again. In the zero-shot setting, Meditron achieves the second-best performance, while MMed-Llama3 achieves the second-best performance. Most of the comparison methods achieve better performance when few-shot demonstrations are given. We believe it is because of the provided task format as we discuss in Section 4.2.1. Moreover, LLMs can also recognize the fine-grained differences between different classes given few-shot demonstrations, making better decisions in classification. Especially, we notice that Meditron performs badly under the few-shot evaluation. We attribute it to the language degradation issue since it accepts a few long documents in the context, amplifying the noise when understanding Japanese.

4.2.5 Semantic Text Similairity

The performances on the STS task are varied dramatically. Qwen2 achieves excellent performance on this task, while the prediction of other models like Llama2-based models (i.e., Llama2, Meditron, SwallowLM) is close to random guess. One possible reason is that the distribution of generating numbers is close to a uniform distribution for these models. Recent works also show the shortage of LLMs from this aspect (Shah et al., 2023; Avnat et al., 2024). However, understanding and generating numbers accurately is essential in the biomedical domains (e.g., on blood test reports). Therefore, it is also a promising search direction in the field of biomedical NLP.

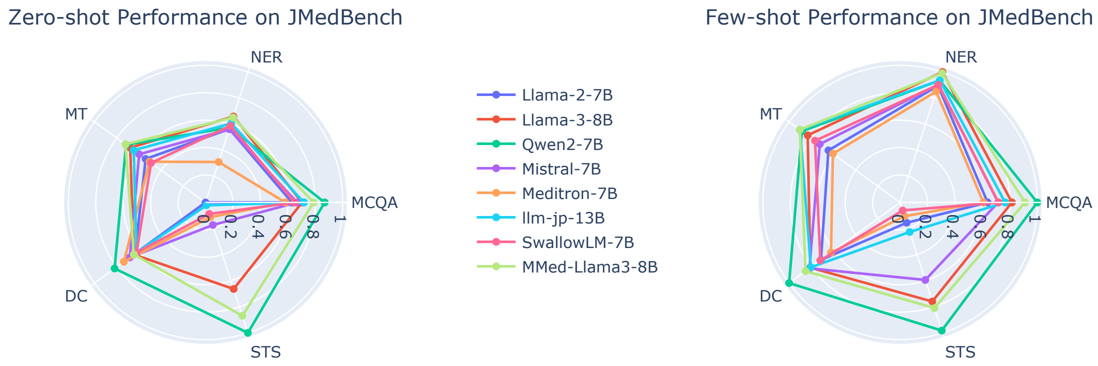

Figure 2: Zero-shot and few-shot performances on different tasks in JMedBench.

4.3 Discussions

In this section, we will conduct an integral and in-depth analysis of the experimental results.

4.3.1 Comparison of Model Performances

Figure 2 includes two radar charts that demonstrate models' zero-shot and few-shot performance on different tasks. Besides, we also rank the model performance and visualize the rankings in Figure 6 as shown in the Appendix. A larger distance from the center represents a higher ranking and better performance. From the radar charts, we can find out that basically, MMed-Llama3, Qwen2, and Llama3 are the most outstanding LLMs on various tasks. Few-shot examples also significantly improve the model performances in all tasks.

4.3.2 Effect of Prompt Templates

We also hope to understand the performance of prompt templates across different tasks and models. In zero-shot evaluation, Figure 3 illustrates that the performance of Standard, English-centric, and Instructed prompt templates do not differ significantly, but using English-centric templates usually achieves a slightly better performance. This phenomenon is even more evident in English-centric LLMs. We believe it is because these models have a greater advantage in understanding English instructions, even when facing cross-lingual contexts. Moreover, Figure 4 shows that few-shot demonstrations reduce the differences between prompt templates to a certain extent, with a particularly noticeable enhancement for minimal prompt templates. We believe it is because the output relies less on the instructions and can instead understand the task format from the few-shot examples.

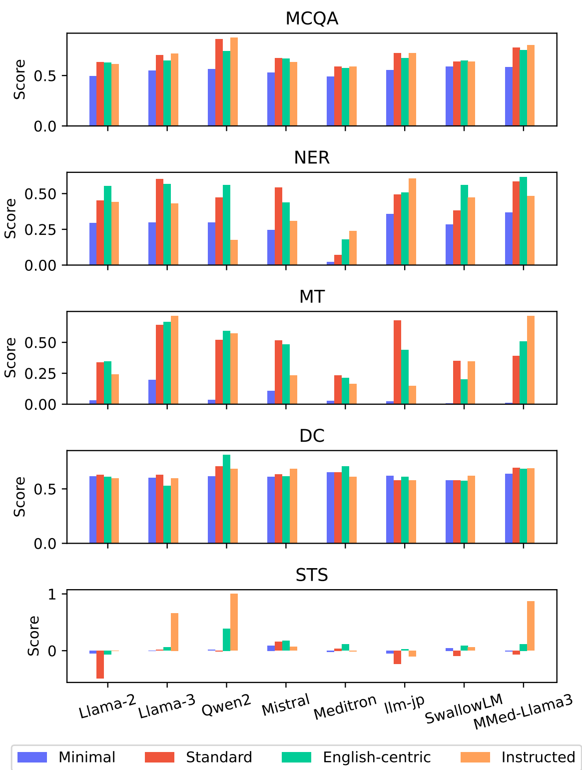

Figure 3: Zero-shot performance under different prompt templates.

5 Conclusions

In this paper, we discuss an urgent need for the field of Japanese biomedical LLMs that requires a solid benchmark for evaluation and comparison. We collect a large collection of Japanese datasets in diverse biomedical tasks, including MCQA, MT, NER, DC, and STS. Considering the scale of the human-manufactured datasets, we translate several large-scale datasets with high quality in English to ensure robust benchmarking results.

Based on the constructed dataset collection, we conduct an evaluation of four types of models, including Japanese biomedical LLMs, Japanese Gen-

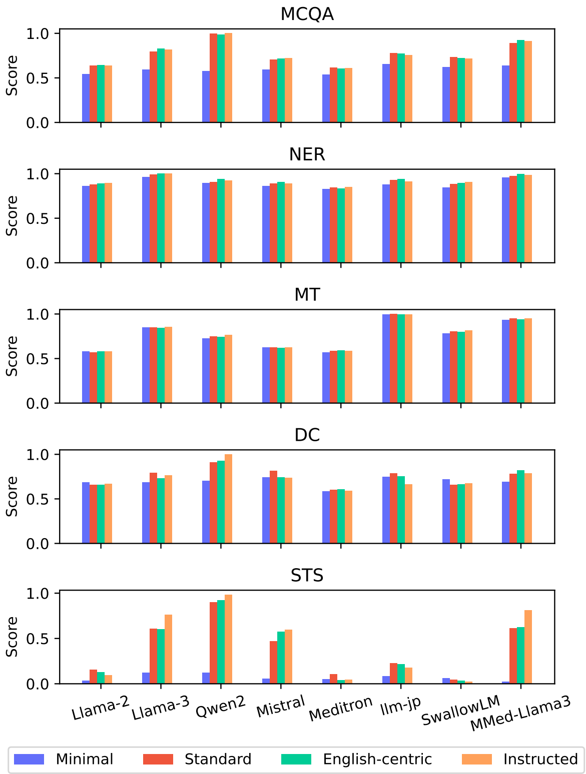

Figure 4: Few-shot performance under different prompt templates.

eral LLMs, biomedical LLMs in other languages, and general LLMs in other languages. Reported performances reveal some insights for improving existing Japanese LLMs in the biomedical domain. Furthermore, our datasets and evaluation tools are publicly available for future research.

Limitations

Considering the difficulty of evaluating natural language generation (NLG) tasks that usually require human evaluation, we only include natural language understanding (NLU) tasks or reformulate NLU tasks into NLU tasks. However, NLG tasks are also widely used in real-world applications. In the future, we consider introducing LLM-based evaluation methods to unlock an easy evaluation of NLG tasks, enriching our benchmark for a further comprehensive evaluation.

With the help of superior modern large language models, we can construct a large-scale benchmarking dataset with less human effort, but the quality of the translation is concerning. During our manual correction of 223 invalid NER samples, we realized the quality was high enough for model comparison.

Due to the limitation of our budgets, we only translate several datasets of MCQA and NER. We only perform evaluation on models with 7B/8B model parameters. For a comprehensive evaluation, we should also perform comparison in a larger scale. We leave it as a future work to include more translated large-scale datasets in other tasks and evaluation results of larger models. Moreover, though we evaluate these models with four categories of prompt templates, each category only contains one template, which may introduce some fluctuation. To further improve the robustness of our benchmark, we consider including more diverse prompt templates in each prompt category in the future.

Evaluation results on Japanese general domains and biomedical domains in other languages are also valuable for comparison, providing some insights into developing Japanese biomedical LLMs. Such multilingual biomedical benchmark containing diverse tasks is a promising research direction in the future. However, it is out of our scope in this paper.

Ethics Statement

We follow the licenses of the involved datasets, which are mainly MIT or CC-BY-4.0[9]. However, we should note that the NRNER and JCSTS datasets are distributed under the Non-Commercial CC-BY-NC-SA-4.0 license[10]. In principle, the whole JMedBench should be distributed under a non-commercial license, whereas if it is used for the commercial scenario, these two datasets (i.e., NRNER and JCSTS) should be excluded.

Besides, considering the scale of the existing human-manufactured evaluation datasets, we adopt machine translation systems (i.e., GPT-4) to translate some large-scale and high-quality English biomedical datasets into Japanese to fulfill a robust evaluation. However, machine translation systems will inevitably generate unfaithful content. Therefore, those who want to use our datasets to develop faithful biomedical LLMs for real-world applications should be aware of this limitation.

Acknowledgments

This work was supported by JST SPRING, Grant Number JPMJSP2108 and by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

[9] https://creativecommons.org/licenses/by/4.0/deed.en

[10] https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Mohammad Awad AlAfnan. 2024. Large language models as computational linguistics tools: A comparative analysis of chatgpt and google machine translations. *Journal of Artificial Intelligence and Technology*.

Eden Avnat, Michal Levy, Daniel Herstain, Elia Yanko, Daniel Ben Joya, Michal Tzuchman Katz, Dafna Eshel, Sahar Laros, Yael Dagan, Shahar Barami, et al. 2024. Performance of large language models in numerical vs. semantic medical knowledge: Benchmarking on evidence-based q&as. *arXiv preprint arXiv:2406.03855*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rintaro Enomoto, Arseny Tolmachev, Takuro Niitsuma, Shuhei Kurita, and Daisuke Kawahara. 2024. Investigating web corpus filtering methods for language model development in japanese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 154–160.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Chengguang Gan and Tatsunori Mori. 2023. Sensitivity and robustness of large language models to prompt template in Japanese text classification tasks. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Takeshi Hayakawa and Yuki Arase. 2020. Fine-grained error analysis on english-to-japanese machine translation in the medical domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models.

Transactions of the Association for Computational Linguistics, 12:229–246.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. Jglue: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickael Rouvier, Pacome Constant Dit Beaufils, Natalia Grabar, Béatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-Antoine Gourraud, and Richard Dufour. 2024. DrBenchmark: A large language understanding evaluation benchmark for French biomedical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5376–5390, Torino, Italia. ELRA and ICCL.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Faith Wavinya Mutinda, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. Semantic textual similarity in japanese clinical domain texts using bert. *Methods of Information in Medicine*, 60(S 01):e56–e64.

Yuji Naraki, Ryosuke Yamaki, Yoshikazu Ikeda, Takafumi Horie, and Hiroki Naganuma. 2024. Augmenting ner datasets with llms: Towards automated and refined annotation. *arXiv preprint arXiv:2404.01334*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolò Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *CoRR*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Lucía Sanz-Valdivieso and Belén López-Arroyo. 2023. Google translate vs. chatgpt: Can non-language professionals trust them for specialized translation? In *Proceedings of the International Conference HiT-IT*, pages 97–107.

Raj Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma. 2023. Numeric magnitude comparison effects in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6147–6161, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9:1–19.

Mingxing Tan. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Chihiro Yano, Akihiko Fukuchi, Shoko Fukasawa, Hideyuki Tachibana, and Yotaro Watanabe. 2024. Multilingual sentence-t5: Scalable sentence encoders for multilingual applications. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11849–11858, Torino, Italia. ELRA and ICCL.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

A Benchmark Construction Details

A.1 Further details of datasets in the JMedBench

Table 5 shows the statistics of involved datasets in the JMedBench. IgakuQA does not have an official training set, while its genre is similar to MedQA. Therefore, we share the training set of MedQA with IgakuQA for a few-shot evaluation. JMMLU-medical only contains the translated testing set, and we also share the training set of translated MMLU-medical-JP with JMMLU-medical. Considering our limited budgets, we only translated 1,000 training samples randomly selected from the original training set of the PubMedQA. As for the datasets derived from JMED-LLM, including EJMMT, MRNER-Medicine, MRNER-Disease, NRNER, CRADE, RRTNM, and SMDIS, we randomly split a small subset from the original dataset for few-shot evaluation. The size of the training set can be found in Table 5. As for the JCSTS, we also randomly split a small subset to be the training set. For the rest of the datasets, we strictly follow the origin setting of the split and use the training set or development set for a few-show evaluation.

Task	Dataset	Train	Test	Creator
MCQA	IgakuQA	10,178	989	Human
	JMMLU-medical	45	1,271	Human
	MedMCQA-JP	182,822	4,183	MT
	USMLE-QA-JP	10,178	1,273	MT
	MedQA-JP	10,178	1,273	MT
	MMLU-medical-JP	45	1,871	MT
	PubMedQA-JP	1,000	1,000	MT
MT	EJMMT	80	2,400	Human
NER	MRNER-Medicine	10	90	Human
	MRNER-Disease	10	90	Human
	NRNER	10	90	Human
	BC2GM-JP	12,572	5,037	MT
	BC5Chem-JP	4,562	4,801	MT
	BC5Disease-JP	4,560	4,797	MT
	JNLPBA-JP	18,607	4,260	MT
	NCBI-Disease-JP	5,424	940	MT
DC	CRADE	8	92	Human
	RRTNM	11	89	Human
	SMDIS	16	84	Human
STS	JCSTS	170	3,500	Human

Table 5: Statistics of involved datasets in JMedBench.

A.2 Prompt Templates for Data Augmentation

Table 6 shows the prompt template we used when using OpenAI's APIs for translating biomedical MCQA datasets.

Besides, Table 7 is the prompt template for translating biomedical NER datasets.

A.3 Bad Cases during NER Dataset Translation

We summarized three main failure types during machine translation: (1) ambiguity of a single word, for example, 'depression' can be considered as a mental illness (うつ病) or pressing down (抑制); (2) multiple possible expressions of a single word, for example, 'glucose' can be translated into either グルコース or 血糖; (3) differences in grammar between English and Japanese. Table 8 shows one bad case for each typical failure type during translating NER datasets. The parts underlined indicate an inconsistency between the entity and the text translation. Although there is a small number of failure cases during the machine translation phase, we still realize that the quality of the translation for both the entities and the text is very high during the manual modification process, which can prove the reliability and the scalability of our data augmentation method.

B Experimental Details

B.1 Development in chronological order

We sorted the various models according to their release dates. In chronological order, they are: Llama2-7B (Jul. 2023), SwallowLM-7B (Nov. 2023), Meditron-7B (Dec. 2023), Mistral-7B (May 2024), MMed-Llama3-8B (May 2024), Qwen2-7B (Jun. 2024), Llama3-8B (Jul. 2024), llm-jp-13B (Sep. 2024). Figure 5 illustrates the relationship between model performance and release date. The color of the points represents the corresponding tasks, and the shape represents their models. Colored lines reflect the trend of model performance on each task over time. The figure shows that as time progresses, the performance of models on various tasks is consistently improving, especially for the STS task. Moreover, the improvement in the in-context learning (ICL) capabilities of the models is even more pronounced.

B.2 Ranking of Models

Figure 6 shows the zero-shot and few-shot performance rankings on JMedBench tasks among all involved LLMs.

B.3 Comparison Methods

Detailed information for involved comparison methods is listed in Table 9.

B.4 Prompts for Each Task

Detailed prompt templates for each task are shown in Table 10, 11, 12, 13, and 14.

Prompt template for translating MCQA datasets
#System Message You are an excellent machine translation system for the biomedical domain. Translate Japanese to English. Input and output should be in the same JSON format.

{
  "question": {question}
  "options": [
    {option_a},
    {option_b},
    {option_c},
    {option_d},
  ],
  "context": {context} #Optional
}

Table 6: Prompt templates for translating biomedical MCQA tasks.

Prompt template for translating NER datasets
#System Message You are an excellent machine translation system for the biomedical domain. Translate Japanese to English. Input and output should be in the same JSON format. Please keep the original key without any changes. Please promise the consistency of translation. For same English words, you should use the same Japanese translation. Please remove unnecessary spaces while translating.

{
  "entities": {entities}
  "text": {question}
}

Table 7: Prompt templates for translating biomedical NER tasks.

Ambiguity of words
Original Text: Depression is a major clinical feature of Parkinson' s disease. **Original Entity**: depression **Translated Text**: うつ病はパーキンソン病の主要な臨床的特徴です。 **Translated Entity**: 抑制 **Explanation**: According to Cambridge English Dictionary, "depression" has multiple meanings: a mental illness (うつ病), or pressing down (抑制).
Multiple Expressions of a Single Word
Original Text: After recovery from hyperglycaemia, he remained polyuric despite normal blood glucose concentrations; water deprivation testing indicated nephrogenic diabetes insipidus, likely to be lithium-induced. **Original Entity**: glucose **Translated Text**: 高血糖からの回復後、彼は正常な血糖濃度にもかかわらず多尿であり続けました。水制限テストは、リチウム誘発性である可能性のある尿崩症を示しました **Translated Entity**: グルコース **Explanation**: "Glucose" can be translated into either "グルコース" or "血糖".
Difference in Grammar
Original Text: Molecular cloning and characterization of two genes encoding gp138, a cell surface glycoprotein involved in the sexual cell fusion of Dictyostelium discoideum. **Original Entity**: genes encoding gp138 **Translated Text**: "Dictyostelium discoideumの性的細胞融合に関与する細胞表面糖タンパク質であるgp138をコードする2つの遺伝子の分子クローニングと特性評価。 **Translated Entity**: gp138をコードする遺伝子 **Explanation**: Due to grammatical differences, the quantifier "2" is inserted between "genes" and "encoding gp138" when translating the text.

Table 8: Typical bad cases during NER dataset translation

Category	Model	#Params	Checkpoint
General LLMs in other languages	Llama2-7B	7B	`meta-llama/Llama2-7b-hf`
	Llama3-8B	8B	`meta-llama/Meta-Llama3-8B`
	Qwen2-7B	7B	`Qwen/Qwen2-7B`
	Mistral-7B	7B	`mistralai/Mistral-7B-v0.3`
Biomedical LLMs in other languages	Meditron-7B	7B	`epfl-llm/meditron-7b`
General Japanese LLMs	llm-jp-13B	13B	-
	SwallowLM-7B	7B	`tokyotech-llm/Swallow-7b-NVE-hf`
Biomedical Japanese LLMs	MMed-Llama3-8B	8B	`Henrychur/MMed-Llama3-8B`

Table 9: Detailed information of involved comparison methods. We contacted the LLM-JP team and used the provided version 3 of the llm-jp model for evaluation.

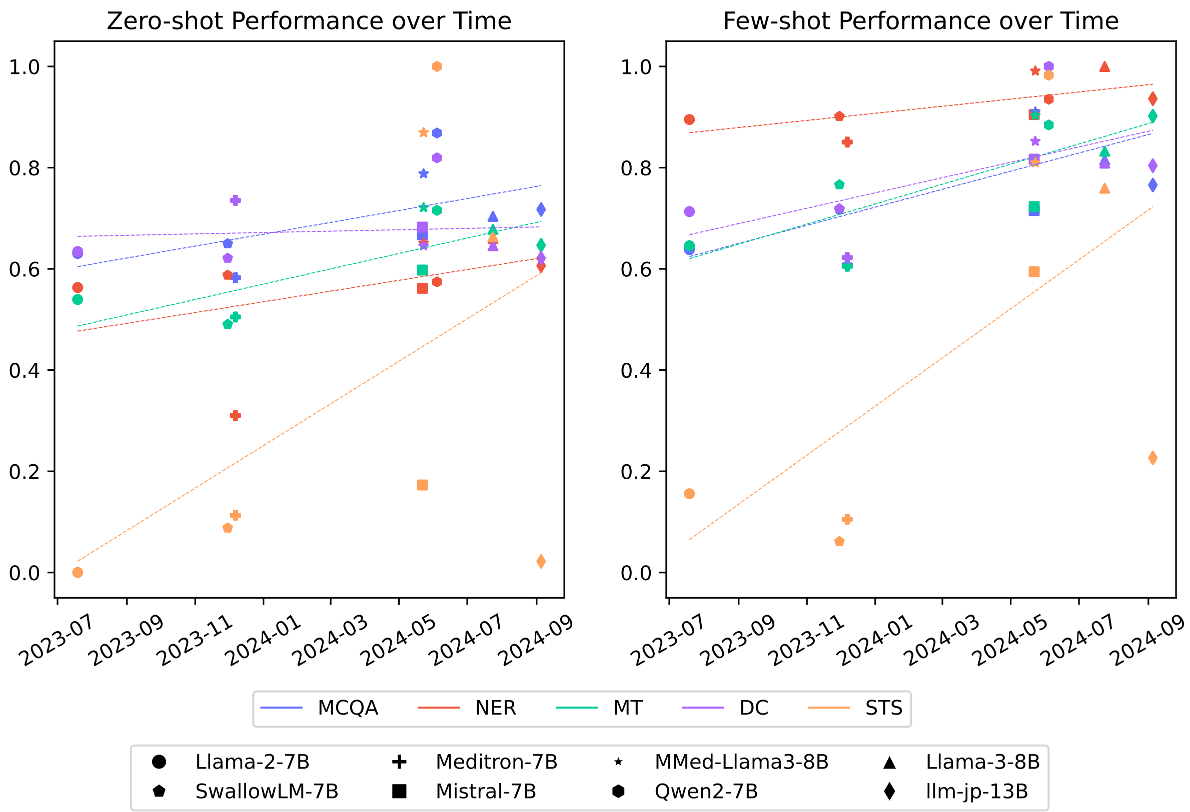

Figure 5: Zero-shot and few-shot performance over time of all involved LLMs.

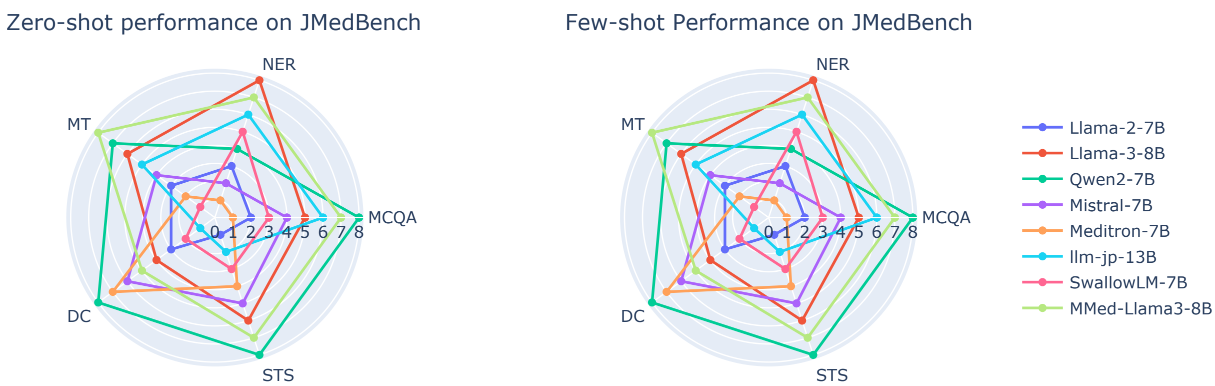

Figure 6: Zero-shot and few-shot performance rankings on JMedBench of all involved LLMs.

Prompt templates for MCQA task		
	w/o Context	**w/ Context**
Minimal	{question}	{context} {question}
Standard	質問：{question} {options} 答え：	要旨：{context} 質問：{question} 答え：
English-centric	Question: {question} {options} Answer:	Abstract: {context} Question: {question} Answer:
Instructed	あなたは医学博士です。基礎科学、臨床科学、医学知識、健康、病気、患者ケア、治療法の基礎となるメカニズムについて理解した上で、以下の選択式問題に答えなさい。以下の選択肢から正しいものを1つ選びなさい。医療ガイドラインに記載されている、現在行われている標準的な治療法に基づいて答えなさい。 質問：{question} 選択肢： {options} 答え：	臨床科学と医学知識の専門家である医師として、次の文が正しいかどうか教えてください。「はい/いいえ/たぶん」のいずれかでお答えください。 要旨：{context} 質問：{question} 答え：

Table 10: Prompt templates for the MCQA task.

Prompt templates for NER task	
Minimal	段落：{text} => {entity_type}：
Standard	以下の段落において、{entity_type}は？ 段落：{text} => {entity_type}：
English-centric	Please extract all {entity_type}s mentioned in the paragraph. Paragraph: {text} => {entity_type}:
Instructed	あなたは医療分野の専門家です。 あなたは{entity_type}のフレーズを含む段落を与えられます。 あなたのタスクは段落からこれらすべてのフレーズを抽出することです。 抽出されたフレーズのみを返し、それらを英語のカンマ（,）で区切る必要があります。 段落：{text} => {entity_type}：

Table 11: Prompt templates for the NER task.

Prompt templates for MT task		
	English→Japanese	**Japanese→English**
Minimal	{source_text} =>	{source_text} =>
Standard	翻訳（English => 日本語）：{source_text} =>	Translation (日本語 => English): {source_text} =>
English-centric	Translation (Japanese => English): source_text =>	Translation (English => Japanese): source_text =>
Instructed	あなたは生物医学文書を翻訳する医学博士です。基礎科学、臨床科学、医学知識、健康、病気、患者ケア、治療法の基礎となるメカニズムを理解した上で、以下の英文を和訳しなさい。 {source_text} =>	あなたは生物医学文書を翻訳する医学博士です。基礎科学、臨床科学、医学知識、健康、病気、患者ケア、治療法の基礎となるメカニズムを理解した上で、以下の和文を英訳しなさい。 {source_text} =>

Table 12: Prompt templates for the MT task.

Prompt templates for DC task	
Minimal	{document} {question}
Standard	文脈：{document} 質問：{question} {classes} 答え：
English-centric	Context: {document} Question: {question} {classes} Answer:
Instructed	あなたは医学博士です。基礎科学、臨床科学、医学知識、健康、病気、患者ケア、治療法の基礎となるメカニズムについて理解した上で、以下の選択式問題に答えなさい。以下の選択肢から正しいものを1つ選びなさい。 文脈：{document} 質問：{question} 選択肢：{classes} 答え：

Table 13: Prompt templates for the DC task.

Prompt templates for STS task	
Minimal	{text_1} {text_2}
Standard	テキスト1：{text_1} テキスト2：{text_2} 類似度 (0-5)：
English-centric	Text 1: {text_1} Text 2: {text_2} Semantic Text Similarity (0-5):
Instructed	あなたは医学博士です。基礎科学、臨床科学、医学知識、健康、病気、患者ケア、治療法の基礎となるメカニズムについて理解した上で、次の2つの文の意味的類似度を0から5の範囲で判断してください。 0：二つの文は完全に似ていない。 5：二つの文は完全に同等で、意味が同じである。 テキスト1：{text_1} テキスト2：{text_2} 類似度 (0-5)：

Table 14: Prompt templates for the STS task.