ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Development and analysis of medical instruction-tuning for Japanese large language models

Issey Sukeda[1]*, Masahiro Suzuki[2], Hiroki Sakaji[3], and Satoshi Kodera[1]

[1]Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, Bunkyo, Tokyo, Japan

[2]Department of Systems Innovation, School of Engineering, The University of Tokyo, Bunkyo, Tokyo, Japan

[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

Abstract

In the ongoing wave of impact driven by large language models (LLMs) like ChatGPT, the adaptation of LLMs to the medical domain has emerged as a crucial research frontier. Since mainstream LLMs tend to be designed for general-purpose applications, constructing a medical LLM through domain adaptation is a huge challenge. While instruction-tuning, particularly based on low-rank adaptation (LoRA), has become a frequently employed strategy to fine-tune LLMs recently, its precise roles in domain adaptation remain unknown. Here, we investigated how LoRA-based instruction-tuning improves the performance of Japanese medical question-answering tasks by employing a multifaceted evaluation of multiple-choice questions, including scoring based on "Exact match" and "Gestalt distance" in addition to the conventional accuracy. Our findings suggest that LoRA-based instruction-tuning can partially incorporate domain-specific knowledge into LLMs, with larger models demonstrating more pronounced effects. Furthermore, our results underscore the potential of adapting English-centric models for Japanese applications in domain adaptation, while also highlighting the persisting limitations of Japanese-centric models. This initiative represents a pioneering effort in enabling medical institutions to fine-tune and operate models without relying on external services.

Keywords: Medical large language models; Llama2; Instruction-tuning; Domain adaptation; Low-rank adaptation; QLoRA

***Corresponding author:**
Issey Sukeda
(sukeda-issei006@g.ecc.u-tokyo.ac.jp)

Citation: Sukeda I, Suzuki M, Sakaji H, Kodera S. Development and analysis of medical instruction-tuning for Japanese large language models.
Artif Intell Health. 2024;1(2): 107-116.
doi: 10.36922/aih.2695

Received: January 10, 2024

Accepted: March 13, 2024

Published Online: April 8, 2024

1. Introduction

The study and development of medical large language models (LLMs) like ChatGPT have the potential to revolutionize the field of medicine and healthcare in profound ways. These models, when fine-tuned and adapted to the medical domain, can assist healthcare professionals in numerous critical tasks, such as disease diagnosis, treatment planning, and patient care. Due to their vast language comprehension capabilities, LLMs may provide up-to-date information, suggest evidence-based treatment options, and even predict disease outcomes with a high degree of accuracy.

Domain adaptation remains a crucial approach for tailoring mainstream LLMs to the practical use in clinical environments, even after the surge of ChatGPT (https://chat.openai.com/), a powerful LLM service, that has revolutionized the way we interact with text and language by its astonishing ability to generate sentences. While these general-purpose models are powerful in zero-shot inference in unseen tasks, fine-tuned models may have the potential to outperform them in domain-specific tasks. Several works on domain adaptation within the medical field in the context of powerful English-centric LLMs[1-4] exist as well, but research in this direction is largely lacking in Japanese, highlighting the need to pioneer studies in non-English contexts. The drive to develop large-scale medical LLMs in one's native language is not only prevalent in Japan but also starting to mainstream in other non-English-speaking countries. In Japan, the sole precedent in the area of Japanese medical language model is the work of Sugimoto *et al*.,[5] who developed a Japanese medical language model named JMedRoBERTa based on RoBERTa, a BERT[6] -based model. This study is the first exploration along this line using large-scale GPT-models with a focus on text generation.

Moreover, ChatGPT utilization is impeded in clinical practices due to the concerns related to data privacy and security. The potential risks associated with data breaches or misuse of confidential patient information underscore the need for robust security measures and ethical considerations, further complicating its seamless integration into clinical settings. Hence, we need to consider domain adaptation using other LLMs for incorporating medical knowledge.

Recently, several parameter- efficient fine-tuning methods have been proposed, including low-rank adaptation (LoRA) and its quantized version (QLoRA),[7,8] where only the limited parameters are chosen as the target of the fine-tuning. Performed along with instruction-tuning, LoRA has demonstrated some success in acquiring conversational abilities and improving domain-specific performances such as financial question-answering tasks.[9,10] That being said, the ability and limitation of LoRA-based instruction-tuning have not been clarified in domain adaptation. "Superficial Alignment Hypotheses," which was proposed recently, provide a conjecture that fine-tuning does not contribute significantly to the acquisition of knowledge, but this topic remains controversial.[11] Therefore, we aim to investigate whether LoRA-based instruction tuning can be effective in acquiring domain-specific knowledge, especially medical knowledge.

The primary research questions guiding our study are as follows:

i. How and how much can domain knowledge be incorporated into LLMs by LoRA-based fine-tuning?
ii. Do larger English-centric LLMs outperform smaller Japanese-centric LLMs?
iii. Does the amount of fine-tuning hold significance?

To answer these questions, we conducted a comprehensive comparison between different LLMs fine-tuned with our own Japanese medical dataset by evaluating e ach model through medical question-answering approach. This enables us to clarify the strengths and limitations of incorporating domain-specific knowledge by LoRA, setting the stage for constructing enhanced versions of various domain-specific Japanese LLMs.

2. Related works

In recent years, there has been active research in constructing pretrained language models specialized for the medical domain. Before the emergence of GPT-3[12] in 2020 and ChatGPT in 2022, the prevailing trend in research involved building BERT[6] -based language models and evaluating them in classification tasks. In English-speaking regions, models such as BioBERT,[13] Med-BERT,[14] ClinicalBERT,[15] and PubMedBERT[16] have been proposed, leveraging medical literature databases such as PubMed and clinical records databases such as MIMIC-III.[17] Also in Japan, UTH-BERT[18] and JMedRoBERTa[5] have become available online. UTH-BERT[18] is the first medical pretrained language model in Japanese, pretrained by approximately 120 million lines of clinical texts. On the other hand, JMedRoBERTa[5] utilizes 11 million lines of journal articles in medicine, with the goal of accumulating information across a diverse range of content, encompassing basic research to case studies.

In the wake of GPT-3[12] and ChatGPT emergence, the focus of research shifted toward LLMs leveraging Transformer[19] accompanied with a steady increase in the parameter size of models. The primary tasks of interest in research also transitioned from classification tasks to medical text generation or medical question-answering. For the English-centric model, BioMedLM (formerly known as PubMedGPT),[20] BioGPT,[21] and BioMedGPT[22] have been proposed, harnessing the strength of the latest general-purpose LLMs. However, the currently available models have limited sizes: BioMedLM[20] has 2.7 billion parameters, BioGPT[21] is based on the GPT-2[23] architecture with 1.3 billion parameters, and BioMedGPT[22] comprises 10 billion parameters. On the other hand, Google has pursued its own path in developing medical models, including Med-PaLM[1] and Med-PaLM2[2] with 540 billion and 340 billion parameters, respectively; nonetheless, these models are not accessible to the public. To the best of our

knowledge, there has been n research conducted to deepen the medical specialization of Japanese-centric model.

3. Data and methods

We conducted a comprehensive comparison between different LLMs fine-tuned with Japanese medical dataset, including those we have created ourselves. To determine whether one should start from a smaller Japanese model or a larger English model, we prepared OpenCALM-7B and Llama2-70B as base models. In addition, to observe the effectiveness of pretraining, we introduced a model additionally trained on medical documents. Subsequently, we applied medical instruction-tuning (LoRA, QLoRA) to each of them and evaluated performance based on the accuracy of medical question-answering tasks. The entire procedure is outlined in Figure 1. The models trained and used in our experiments are available at https://huggingface.co/AIgroup-CVM-utokyohospital.

3.1. Base model preparation

To create a Japanese-centric model, we utilized OpenCALM-7B (https://huggingface.co/cyberagent/open-calm-7b), an open-source Japanese foundation LLM with 6.5 billion parameters developed by CyberAgent, Inc. In addition, we trained a new base model MedCALM, which is based on OpenCALM-7B and continually pretrained on our own medical text dataset. Here, the training dataset consists of 2420 examples, and the evaluation dataset has 50 examples. The maximum token count is set to 768, and the batch size is set to 63. The model was trained for 2000 steps. On the other hand, we further used Llama2-70B-chat-hf (https://huggingface.co/meta-Llama/Llama-2-70b-chat-hf), a powerful English-centric LLM released by Meta Inc.[24] Hereinafter, it is referred to as Llama2-70B. The use of this model is governed by the Meta license (https://ai.meta.com/resources/models-and-libraries/llama-downloads/).

3.2. Medical instruction-tuning

Instruction-tuning refers to the process of fine-tuning or optimizing the behavior and output of the model by providing explicit instructions or guidance as a prompt during the generation of text.[25] We employed LoRA, one of the popular parameter-efficient fine-tuning methods provided in PEFT library,[7,26] since full fine-tuning, which retrains all model parameters, is unfeasible in our environment. LoRA freezes the pretrained model weights and inserts trainable rank decomposition matrices into each layer of the target model to reduce the number of trainable parameters for downstream tasks. Specifically, instead of directly updating the $d \times k$ parameter matrix of a linear layer in LLM from W_0 to $W_0+\Delta W$, LoRA updates a $d \times r$ matrix B and a $r \times k$ matrix A where BA is low-rank decomposition of ΔW, that is, $r \ll min\ (d,k)$.

Given our computational constraints, particularly the limited GPU memory, LoRA for OpenCALM-7B is feasible, but not for Llama2-70B. Instead, we opted for the quantized version, named QLoRA,[8] which is intended to trade off a slight performance drop for a significant reduction in model size, making the experiment using Llama2-70B feasible. Consequently, we applied LoRA to OpenCALM-7B and QLoRA to Llama2-70B, respectively. The hyperparameters of LoRA/QLoRA are listed in Table 1, which follow the default setting specified in PEFT library and QLoRA library, respectively.[8,26]

To perform medical instruction-tuning, we constructed a medical question-answer dataset containing 77422 records in instruction format. Initially, we reviewed two medical articles, one from the official journal of The Japanese Circulation Society (containing 3569 lines) and another from the Journal of the Japanese Society of Internal Medicine (JJSIM, containing 6120 lines), for input retrieval. Then, these texts were used as inputs for ChatGPT (gpt-3.5-turbo) to generate various question-answer pairs, resulting in 21365 records and 56057 records, respectively. Since ChatGPT is known to possess strong instruction-following ability, we utilized the following prompt template to construct instruction dataset with an overall good quality:

> ### Instructions: You are a machine designed to generate various question and answer pairs. Please create data with question (instruction) and answer (output) pairs based on the following input, considering it as prior knowledge. Format the data

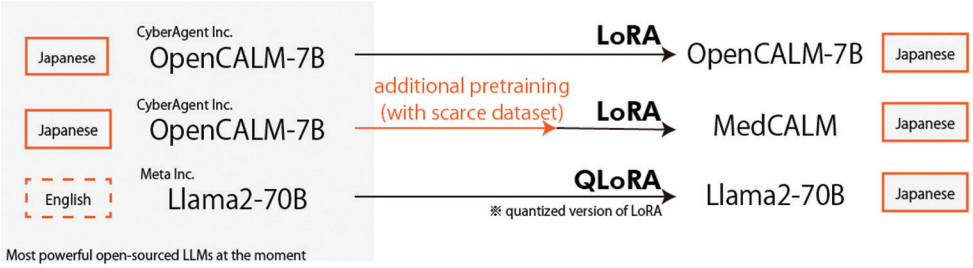

Figure 1. Overview of procedure of our medical instruction-tuning. Image created with Adobe Illustrator.

Table 1. LoRA/QLoRA parameters

	OpenCALM-7B	Llama2-70B
Fine-tuning method	LoRA	QLoRA
Learning rate	5e -5	2e-4
Input length	512	512
Target max length	512	512
Batch size	8	8
Fine-tuning steps	1k, 3k, 10k	0.9k, 3k
r of (Q) LoRA	8	64
α of (Q) LoRA	32	16
Dropout rate of (Q) LoRA	0.05	0.1
Target parameter	Query, Key, value	All linear layers

as "instruction": Question content, "output": Answer content, and do not include line breaks. Repeat this process 15 times and list one data pair per line.
Input: {input_text}

The number of epochs and steps was set to align with the overall computational time in each experiment. Using a larger model such as Llama2-70B increases the GPU memory usage per sample. To avoid this, memory usage can be reduced by decreasing the floating-point precision or by using gradient accumulation. In this study, we adopted 4-bit QLoRA on Llama2-70B. Since 4 bits is optimal in terms of the relationship between floating-point precision and model performance,[27] it is not desirable to reduce the floating-point precision any further. To experiment with less GPU memory, gradient accumulation was attempted by multiplying batch size calculation, for example, a batch size of 8 is calculated twice with four smaller mini-batch sizes. This approach allows for building larger models and reducing requirements for computing resources.

3.3. Evaluation by medical question-answering tasks

The state-of-the-art performance of English medical LLMs is typically evaluated using benchmark datasets such as MedQA (United States Medical Licensing Examination, USMLE),[28] MedMCQA,[29] and PubMedQA.[30] However, the availability of Japanese-curated medical task datasets is significantly limited, with IgakuQA (Japanese medical licensing exams)[31] being the only one available at present. Hence, in addition to IgakuQA, we prepared a new Q&A dataset JJSIMQA to assess the performance of each model in the medical domain. JJSIMQA is our own dataset comprising 5-choice questions included in JJSIM as appendices. Here are some samples from IgakuQA and JJSIMQA datasets:

An example from IgakuQA (originally in Japanese)
"problem_id": "116A1",
"problem_text": "Which of the following is incorrect regarding hypertension caused by obstructive sleep apnea?,"
"choices": {"a": "It often leads to nocturnal hypertension.", "b": "Weight reduction is recommended for obese patients.", "c": "Alpha-blockers are the first-line choice of medication.", "d": "Morning hypertension is frequently observed in home blood pressure measurements.", "e": "Continuous positive airway pressure (CPAP) therapy is expected to lower blood pressure."},
"text_only": True,
"answer": ["c"]
An example from JJSIMQA, 5-choice questions in JJSIM (originally in Japanese)
"problem_text": "Which of the following is incorrect about recent cases of hepatitis B in Japan? Choose one.",
"choices": {"a": "The HBs antigen positivity rate has significantly decreased due to the initiation of mother-to-child infection prevention programs.", "b": "HBV (hepatitis B virus) genotype Ae can become a carrier through horizontal transmission in adults.", "c": "In Japan, routine HBV vaccination began in October 2016.", "d": "HBV genotype C is more prevalent in the Tohoku and Miyako-Yaeyama regions.", "e": "Horizontal transmission of HBV during childhood is thought to be partly attributed to father-to-child transmission and communal living."},
"text_only": True,
"answer": ["d"]

The prompt template used for the evaluation follows the Alpaca-format,[32] where "problem_text" is incorporated in {instruction} and "choices" is incorporated in {input}:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
Instruction:
{Instruction}
Input:
{Input}
Response:

For evaluation in our experiments, these prompts were given in Japanese for OpenCALM-7B and in English for Llama2-70B. When generating the responses, we can specify parameters. In our experiments, *temperature* was set to 0.1, *max_new_tokens* to 256, *top_p* to 0.9, and *repetition_penalty* to 1.05. Question-answering samples that yielded null responses were excluded from the dataset.

Finally, we evaluated the output responses of each model by three different metrics: *Exact match*, *Gestalt score*, and *Accuracy*. While all these metrics aim to assess how effectively models can select the correct choice from

five alternatives, they are defined with slight variations. Let R denote the response string and C^* denote the correct answer string among the five choices. *Exact match* takes the value of 1 if R and C^* exactly match at the string level, and 0 otherwise. *Gestalt score* is defined as the Gestalt distance between the response and the correct answer, which is calculated by a string matching algorithm that is based on the longest common subsequence: let K denote the longest matched string, then *Gestalt score* is calculated as $\text{GestaltScore}(R) = 2|K|/(|R|+|C^*|)$. Finally, *Accuracy* reflects the correctness by evaluating the choice closest to the model's response when measured using *Gestalt score*. Definitions are summarized as follows:

$S = \{C_1, C_2, C_3, C_4, C_5\}$: Choices,

C^* ($\in S$): The correct choice,

R: the response of the model,

$\text{ExactMatch}(R) = 1$ if $R = C^*$ else 0,

$\text{GestaltDistance}(R,C) = 2|K|/(|R|+|C|)$, K: the longest matched string between R and C,

$\text{GestaltScore}(R) = \text{GestaltDistance}(R,C^*)$,

$\text{Accuracy}(R) = 1$ if $\text{argmax}\_\{C \in S\}\ \text{GestaltDistance}(R,C) = C^*$ else 0.

All the evaluation metrics mentioned above take the value between 0 and 1, and the larger value indicates the better performance of the model.

3.4. Experimental settings

The whole dataset used in this work is summarized in Table 2. The experiments were run on 4 NVIDIA A100 with 80GB RAM each. All codes were implemented in Python, and the software and libraries we used include Transformers[33] and PEFT[26] from Hugging Face.

4. Results

4.1. The effect of medical instruction-tuning

The average score of experiments conducted for both 0-shot inference and 1-shot inference, measured by *Exact match*, *Gestalt score*, and *Accuracy* is summarized in Table 3 and Figure 2. The 0-shot inference refers to making responses without any specific examples, while the 1-shot inference refers to when one question-answer example is included in the input prompt. In Table 3, the top 2 scores in each row are highlighted in bold.

4.2. Comparison of our string-based evaluation metrics

Evaluation of LLMs is mainly conducted via manual evaluation[1] and automated evaluation based on rules. In automated evaluation methods, likelihood-based evaluation[34] is predominant. However, this evaluation method assesses the vectors outputted by the model rather than the actual generated strings, making it unsuitable for comparison with ChatGPT. To address this issue, our evaluation metrics are based on the strings actually outputted by the model. *Exact match* is a strict criterion where a response is considered correct only if it matches the correct answer precisely. Consequently, the number of correct answers is lower because even slight deviations are not considered correct. On the other hand, *Accuracy* is a relatively lenient metric where an output is considered correct as long as it is similar to the correct answer, even if it is not an exact match. This leads to a relatively higher number of correct answers as compared to *Exact match*, as deviations are tolerated to some extent.

Table 4 is a contingency table showing the number of question-and-answer (Q&A) samples where the model produced the correct answer. As a result, 112 question-answer samples are considered correct in terms of *Accuracy* but wrong in *Exact match*, whereas the reverse is not true. Among these 112 samples, many cases that were thought to be correct were not considered correct in the \textit{Exact match} evaluation. This was due to issues such as the model' s output being corrupted by token omissions in the tokenizer, or experiencing partial misrepresentation of Japanese characters, as observed in the examples listed in Table 5. This result implies that *Accuracy* is more suitable for evaluating performance in question-answering than *Exact match*, as it is more robust against the issues that models may potentially encounter. Further discussion in this regard is given in section 5.3.

4.3. Example responses from each model

We randomly created questions that ask each model the treatment of a symptom. This type of medical question is

Table 2. Datasets used in this work

Name	Source type	Format type	Purpose	Number of records
The Japanese circulation society	Academic journal	Alpaca format[32]	Instruction-tuning	21365
The *Journal of the Japanese Society of Internal Medicine*	Academic journal	Alpaca format[32]	Instruction-tuning	56057
IgakuQA[31]	Medical license exam	5-choice question	Evaluation	2002
JJSIMQA	Review questions	5-choice question	Evaluation	460

Table 3. Performance of Japanese medical question-answering tasks

	OpenCALM-7B				MedCALM				Llama2-70B		
Steps of QLoRA	**0**	**1k**	**3k**	**10k**	**0**	**1k**	**3k**	**10k**	**0**	**0.9k**	**3k**
Exact match (1s)	0	0.042	0.059	0	0.001	0	0	0	0.097	**0.200**	**0.173**
Gestalt score (1s)	0.053	0.186	0.087	0.078	0.028	0	0.002	0.035	0.247	**0.331**	**0.314**
Accuracy (1s)	0.177	0.190	0.148	0.174	0.164	0.150	0.150	0.165	0.200	**0.258**	**0.225**
Exact match (0s)	0	0.029	0.014	0.013	0	0.018	0.019	0.014	0.001	**0.180**	**0.169**
Gestalt score (0s)	0.033	0.114	0.141	0.120	0.032	0.096	0.116	0.085	0.071	**0.276**	**0.287**
Accuracy (0s)	0.170	0.182	0.166	0.193	0.185	0.172	0.240	0.183	0.170	**0.251**	**0.244**
Training hours	-	4.6	24	37	-	8.9	23.7	58.4	-	12.7	42.4

Notes: 0s and 1s denote 0-shot inference and 1-shot inference, respectively. The top 2 scores of each row are highlighted in bold. 0 steps denote the original base model.

Table 4. Number of Q&A samples where Llama2 (0.9k steps of QLoRA) produced the correct answer

	Correct in *Exact match*	Wrong in *Exact match*
Correct in *accuracy*	384	112
Wrong in *accuracy*	0	1425

not included in the instruction dataset nor the evaluation dataset. Table 6 shows the responses of each model to the following prompt, which was originally Japanese.

> ### Instruction:
> Please provide detailed instructions for the treatment to be administered to patients with the following diseases.
> ### Input:
> deep vein thrombosis
> ### Response:

Here, we observed that the original Llama2-70B generated English responses to some questions — 81% in 0-shot prompting and 15% in 1-shot prompting — while the other models responded completely in Japanese when prompt texts were given in Japanese.

5. Discussion

5.1. Numerical evaluation of the effects of fine-tuning

We observed notable score improvements with LoRA after an appropriate number of steps, particularly with Llama2-70B showing the most significant enhancement. This suggests that utilizing a more powerful English-centric model as the base model holds promise for domain adaptation even in Japanese contexts.

Regarding instruction-tuning, it has been controversial on whether, we should repeat epochs or just once. Our results showed that a single epoch (1k steps) of instruction-tuning improves the performance but increasing the number of epochs exacerbates the model. Furthermore,

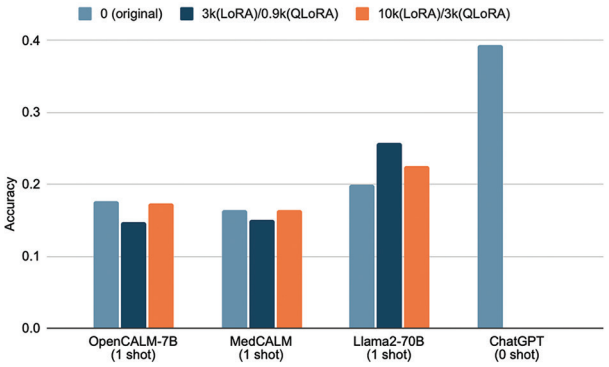

Figure 2. Comparison in accuracy of Japanese medical question-answering tasks. Image created with Google Spreadsheet.

additional pretraining did not contribute to performance improvement. Therefore, we conclude that conducting LoRA-based instruction-tuning for a single epoch without considering additional pretraining is a more practical and promising approach, especially when dealing with limited training data.

Note that in this study, we exclusively utilized medical documents closely related to the task for continual pretraining. However, we believe that the efficacy of additional pretraining could be further explored by incorporating a broader range of medical domain documents or by extracting and expanding from a general-purpose corpus. Determining the necessary amount of data for additional pretraining to improve performance in downstream tasks is a challenge, we will face in the future.

5.2. Deterioration of 1-shot performance

From Table 1, it is evident that every OpenCALM-based model except the original one experiences a decline in 1-shot inference scores rather than in 0-shot inference scores. This outcome highlights the fact that the original OpenCALM model clearly loses its capability to leverage example

Table 5. Some example outputs from Llama2-70B (0.9k steps of QLoRA)

Correct choice	Model's output	*Exact match*	*Gestalt score*	*Accuracy*
Workers' accident compensation insurance (労働者災害補償保険)	Similar but having tokenizing issue (労働者補償保険)	0	0.538	1
Functional menstrual disorders (機能性月経困難症)	Correct except third letter (機能的月経困難症)	0	0.736	1
Intra-articular hemorrhage (関節内出血)	Correct except the second letter (関箋内出血)	0	0.800	1
Monthly standard compensation for patient households (患者世帯の標準報酬月額)	Similar but having tokenizing issue (患者世帯患者世帯標準報酬月額)	0	0.666	1

Table 6. Example responses from each model

Model name	LoRA/ QLoRA	Response (exact output from LLM)
OpenCALM-7B	None	At the bedside, the nurse encouraged the patient to move their legs to assist in getting out of bed. The patient attempted to flex and extend their legs; however, they were unable to move the leg joints and needed assistance to stand up independently. The nurse approached the patient sitting on a bedside chair, provided shoulder support, lifted their arms, extended their elbows, and helped them rise slowly. The nurse then gently released their grip and returned to the chair. The nurse proceeded to call the next patient.*
	1k steps of LoRA	The procedure to be carried out for patients with the following disease involves the implementation of appropriate treatment and preventive measures.*
	3k steps of LoRA	Detailed treatment methods have not been provided.*
Llama2-70B	None	In some cases, a procedure called thrombol.
	0.9k steps of QLoRA	The treatment of deep vein thrombosis involves the use of thrombolytic agents and anticoagulants.*
	3k steps of QLoRA	The treatment for deep vein thrombosis includes pharmacological therapy to dissolve the blood clot as well as surgical interventions to remove the thrombus.*

Note: *Originally in Japanese.
Abbreviation: LLMs: Large language models.

responses provided within the context, whereas Llama2-70B retains this ability even after instruction-tuning.

5.3. Evaluation metrics

There have been some intensive arguments surrounding the evaluation of LLMs recently. Regarding the evaluation method of LLMs, there is still no unified "rule-of-thumb" method yet. While the existing metrics (*e.g.*, JGLUE[35]) or leaderboards (*e.g.*, Nejumi LLM leaderboard, http://wandb.me/nejumi) can assess the fluency of generated texts, they do not adequately evaluate the accuracy of domain-specific knowledge. It is noteworthy that three metrics used in our experiments also exhibit certain shortcomings. For example, *Exact match* cannot accurately score responses that, while conveying the correct meaning, do not match the text verbatim. *Gestalt score* is asymmetric and prone to multiple choices. Overall, our string-based metrics fall short in identifying phrases with different expressions but conveying the same meaning, and reflecting aspects such as fluency and medical accuracy. We argue that these features are not problematic in question-answering tasks where the model is required to output one or a few choices in short texts, but they become problematic when evaluating LLM for practical tasks, including medical report generation, where these aspects are crucial.

Furthermore, even the use of multiple-choice questions for evaluating LLMs has been controversial.[36,37] The development of even more superior evaluation metrics is eagerly anticipated.

5.4. Difficulty and limitations

While numerous LLM training techniques are still in the developmental stage, several shortcomings of training medical LLMs, like what we have done in this work, should be highlighted. First and foremost, the quantity and quality of data could be insufficient in our work. Preparing a medical dataset in instructional format can be expensive. In this study, we employed ChatGPT for automated generation, but this approach may become financially burdensome when preparing larger datasets. Data cleansing has also consistently posed challenges, and achieving perfect results in this work may not have been feasible.

Moreover, during the writing phase of this paper, Japanese LLMs that are considered to perform better than OpenCALM-7B, which was used in this study, have been released (see, *e.g.*, Rakuda benchmark, https://yuzuai.jp/benchmark). There is a possibility of obtaining different results when using them as the base model. Since one general

implication suggested by the results of this experiment is that "a more powerful base model is preferable to start with," an overall performance improvement by upgrading the base model is highly expected.

6. Conclusion

In this paper, we explore the capabilities and limitations of LoRA through various comparative analyses in the medical domain. LoRA-based instruction-tuning, while avoiding an excessive number of steps, can partially integrate domain-specific knowledge into LLMs, with larger models demonstrating more pronounced effects. We also observe a decrease in performance after additional pretraining on scarce training dataset. Furthermore, our results underscore the potential of adapting larger English-centric models for Japanese applications in domain adaptation, while also highlighting the persisting limitations of Japanese-centric models including the deterioration of 1-shot performance after instruction-tuning. Our findings here suggest that, at present, the most promising approach in constructing a domain-specific LLM is applying QLoRA to larger English-centric base models.

Given the current situation, the clinical translation of medical LLMs into real-life applications still falls short of our expectations. To fully harness the potential of medical LLMs in healthcare settings, addressing both the performance limitations and the associated security and privacy concerns is imperative. Further research and development efforts are needed to enhance the accuracy and reliability of these models, ensuring they meet the rigorous standards required for clinical decision.

Furthermore, the integration of medical LLMs with other AI technologies, such as those utilized in electrocardiograms and electronic medical records, has the potential to amplify their impact significantly. By collaborating and cohesively using these AI systems along with medical LLMs, physicians can achieve a more comprehensive understanding of patient data, with which they could formulate more personalized treatment plans to improve patient outcomes.

Acknowledgments

None.

Funding

This study was supported by the Japan Agency for Medical Research and Development (Grant Number: JP23hk0102078h0003).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Issey Sukeda, Satoshi Kodera
Formal analysis: Issey Sukeda
Investigation: Issey Sukeda, Satoshi Kodera
Methodology: Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji
Writing – original draft: Issey Sukeda
Writing – review & editing: Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Journal articles used in the study are available online in PDFs. ChatGPT is utilized for generating and cleansing the data. IgakuQA is available online. JJSIMQA is not made publicly available.

Further disclosure

Part of findings has been presented in *Deep Generative Models for Health* in NeurIPS 2023. In addition, a submission made to NeurIPS workshop is available on arXiv (https://doi.org/10.48550/arXiv.2310.10083).

References

1. Singhal K, Azizi S, Tu T, *et al.* Large language models encode clinical knowledge. *Nature*. 2023;620:172-180.

   doi: 10.1038/s41586-023-06291-2

2. Singhal K, Tu T, Gottweis J, *et al.* Towards expert-Level medical question answering with large language models. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2305.09617

3. Tu T, Azizi S, Driess D, *et al.* Towards generalist biomedical AI. *NEJM AI.* 2024;1(3):AIoa2300138.

   doi: 10.1056/aioa2300138

4. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2306.09968

5. Sugimoto K, Iki T, Chida Y, Kanazawa T, Aizawa A. JMedRoBERTa: A Japanese Pre-trained Language Model

on Academic Articles in Medical Sciences (in Japanese). In: *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*; 2023.

6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; 2019.
7. Hu EJ, Wallis P, Allen-Zhu Z, *et al*. LoRA: Low-rank Adaptation of Large Language Models. In: *International Conference on Learning Representations*; 2021.
8. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. In: *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*; 2023:10088-10115.
9. Suzuki M, Hirano M, Sakaji H. From base to conversational: Japanese instruction dataset and tuning large language models. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE; 2023:5684-5693.

   doi: 10.1109/bigdata59044.2023.10386605
10. Xie Q, Han W, Zhang X, *et al.* PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. In: *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*; 2023:33469-33484.
11. Zhou C, Liu P, Xu P, *et al*. Lima: Less is More for Alignment. In: *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*; 2023:55006-55021.
12. Brown T, Mann B, Ryder N, *et al.* Language models are few-shot learners. In: *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020:1877-1901.
13. Lee J, Yoon W, Kim S, *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.

   doi: 10.1093/bioinformatics/btz682
14. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. 2021;4:86.

   doi: 10.1038/s41746-021-00455-y
15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv*. Preprint posted online 2019.

   doi: 10.48550/arXiv.1904.05342
16. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23.

   doi: 10.1145/3458754
17. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.

   doi: 10.1038/sdata.2016.35
18. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One*. 2021;16(11):e0259763.

   doi: 10.1371/journal.pone.0259763
19. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017:6000-6010.
20. Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. *Stanford CRFM Introduces PubMedGPT 2.7B*; 2022. Available from: https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b [Last accessed on 2024 Apr 04].
21. Luo R, Sun L, Xia Y, *et al*. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform.* 2022;23:bbac409.

   doi: 10.1093/bib/bbac409
22. Luo Y, Zhang J, Fan S, *et al.* BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. *arXiv*. Preprint posted online 2023.

   doi: 10.48550/arXiv.2308.09442
23. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019;1:9.
24. Touvron H, Martin L, Stone K, *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv.* Preprint posted online 2023.

   doi: 10.48550/arXiv.2307.09288
25. Wei J, Bosma M, Zhao V, *et al*. Fine-tuned Language Models are Zero-shot Learners. In: *International Conference on Learning Representations*; 2022.
26. Mangrulkar S, Gugger S, Debut L, Belkada Y, Paul S. *PEFT: State-of-the-art Parameter-Efficient Fine-tuning Methods*; 2022. Available from: https://github.com/huggingface/peft [Last accessed on 2024 Apr 04].
27. Dettmers T, Zettlemoyer L. The Case for 4-bit Precision: K-bit Inference Scaling Laws. In: *ICML'23: Proceedings of the 40th International Conference on Machine Learning*. 2023:7750-7774.
28. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11(14):6421.

   doi: 10.3390/app11146421

29. Pal A, Umapathi LK, Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. In: *Proceedings of the Conference on Health, Inference, and Learning.* 2022:248-260.

30. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. 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); 2019:2567–2577.

doi: 10.18653/v1/D19-1259

31. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. *arXiv.* Preprint posted online 2023.

doi: 10.48550/arXiv.2303.18027

32. Taori R, Gulrajani I, Zhang T, *et al*. *Stanford Alpaca: An Instruction-following Llama Model*; 2023. Available from: https://github.com/tatsu-lab/stanford_alpaca [Last accessed on 2024 Apr 04].

33. Wolf T, Debut L, Sanh V, *et al*. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020:38-45.

34. Gao L, Tow J, Biderman S, *et al*. A framework for few-shot language model evaluation. *Zenodo*. 2023;v0.0.1.

doi: 10.5281/zenodo.5371629

35. Kurihara K, Kawahara D, Shibata T. JGLUE: Japanese General Language Understanding Evaluation. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*; 2022:2957-2966.

36. Pezeshkpour P, Hruschka E. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2308.11483

37. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large Language Models Are Not Robust Multiple Choice Selectors. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2309.03882