JMedLoRA:Medical Domain Adaptation on Japanese Large Language Models using Instruction-tuning

Issey Sukeda[1] Masahiro Suzuki[2] Hiroki Sakaji[2]
Satoshi Kodera[1]

[1] Department of Cardiovascular Medicine, Graduate School of Medicine,
[2] Department of Systems Innovation, School of Engineering,
The University of Tokyo, Japan

Abstract

In the ongoing wave of impact driven by large language models (LLMs) like ChatGPT, the adaptation of LLMs to 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 is used to fine-tune some LLMs, its precise roles in domain adaptation remain unknown. Here we show the contribution of LoRA-based instruction-tuning to performance in Japanese medical question-answering tasks. In doing so, we employ a multifaceted evaluation for 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.

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 can possibly provide up-to-date information, suggest evidence-based treatment options, and even predict disease outcomes with a high degree of accuracy.

To fill the gap between mainstream of LLMs and the practical use in clinical environments, domain adaptation is a crucial technique even after the surge in ChatGPT[1], 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 instances of domain adaptation within the medical field in the context of the English-centric LLMs [1, 2, 3] exist, but research in this direction is largely lacking in Japanese, highlighting the need to pioneer studies in non-English contexts (It is worth noting that JMedRoBERTa [4] deals

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[1] https://chat.openai.com/

with BERT-based models, not GPT). The potential impact of large-scale medical LLMs provided in one's native language could be shared not only in Japan but also in other non-English-speaking countries.

Moreover, using ChatGPT 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 knowledges. Recently, several parameter efficient fine-tuning methods have been proposed including Low Rank Adaptation (LoRA) [5] and its quantized version (QLoRA) [6], 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 [7]. That being said, the ability and limitation of LoRA-based instruction-tuning has not been clarified in domain adaptation. Recently, "Superficial Alignment Hypotheses" has been proposed, insisting that fine-tuning does not contribute significantly to the acquisition of knowledge, but this topic remains controversial [8]. Therefore, we aim to investigate whether LoRA-based instruction tuning can be effective in acquiring domain-specific knowledge.

The primary research questions guiding our study are as follows:

1. How and how much can domain knowledge be incorporated into LLMs by LoRA-based fine-tuning?

2. Do larger English-centric LLMs outperform smaller Japanese-centric LLMs?

3. Does the amount of fine-tuning hold significance?

To answer these questions, we conduct a comprehensive comparison between different LLMs fine-tuned with our own Japanese medical dataset. Each model is evaluated by medical question-answering tasks. In doing so, we clarify the ability and limitation of incorporating domain-specific knowledge by LoRA so as to open the path to further constructing stronger versions of various domain-specific Japanese LLMs.

2 Methods

We conduct a comprehensive comparison between different LLMs fine-tuned with Japanese medical dataset, including those we have created ourselves. The models used in our experiments are available at `https://huggingface.co/AIgroup-CVM-utokyohospital`.

2.1 Base Models

As our main scope is developing Japanese LLMs, we use as the baseline OpenCALM-7B[2] , the open-sourced Japanese foundation LLM with 6.5 billion parameteres published by CyberAgent, Inc. Additionally, we train a new model MedCALM, which is based on OpenCALM-7B and additionally pretrained on our own medical text dataset. 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. We further use Llama2-70B-chat-hf [9][3], a powerful English-centric LLM released by Meta Inc.

2.2 Instruction-tuning

Instruction-tuning [10] 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 text during the generation of text. Recently, several implementations of instruction-tuning have been developed for GPT-based LLMs. Low Rank Adaptation (LoRA) [5] is one of the major parameter-efficient fine-tuning methods provided in PEFT library [11]. The quantized version of LoRA is provided as QLoRA [6], which enables us to tune a model with as many as 70 billion parameters within limited computation resources. In this study, we apply LoRA to OpenCALM-7B and QLoRA to Llama2-70B-chat-hf, respectively.

[2] `https://huggingface.co/cyberagent/open-calm-7b`

[3] `https://huggingface.co/meta-Llama/Llama-2-70b-chat-hf`

Table 1: Performance of Japanese medical question-answering tasks. 0s and 1s denote 0-shot inference and 1-shot inference, respectively. The top 2 scores of each row are highlighted in bold. 0 step denotes the original base model.

	OpenCALM-7B				MedCALM				Llama2-70B			ChatGPT
Steps of (Q)LoRA	0	1k	3k	10k	0	1k	3k	10k	0	0.9k	3k	-
Accuracy (0s)	.222	.219	.213	.239	.236	.216	.214	.223	.216	**.295**	.285	**.438**
Exact match (0s)	.002	.019	**.047**	.039	.028	.029	.029	.034	.017	.038	.045	**.112**
Gestalt score (0s)	.033	.114	.141	.078	.032	.096	.113	.085	.071	.276	**.287**	**.369**
Accuracy (1s)	.227	.244	.196	.223	.207	.188	.188	.204	.244	**.296**	**.267**	-
Exact match (1s)	.009	**.070**	.026	.019	.021	0	.001	.019	.045	**.057**	.056	-
Gestalt score (1s)	.053	.186	.087	.078	.028	0	.002	.035	.247	**.331**	**.314**	-
Training hours	-	4.6	24.0	37.0	-	8.9	23.7	58.4	-	12.7	42.4	-

3 Experiment

We applied each fine-tuning method to each base model and verified accuracy and medical correctness of the generated responses. All experiments were conducted on NVIDIA A100 with 4 GPUs (80GB VRAM each).

3.1 Training

To perform instruction-tuning, we constructed a medical question-answer dataset containing 77422 records in instruction format by utilising ChatGPT (gpt-3.5-turbo). Two medical articles, one from the official journal of The Japanese Circulation Society and another from the Journal of the Japanese Society of Internal Medicine (JJSIM), were used as input for ChatGPT. For further details of instruction data generation, see Appendix C. The number of epochs and steps were set so as to approximately align with the overall computational time.

3.2 Evaluation by Medical Question-answering Tasks

To assess performance in medical domain, we prepared two medical Q&A datasets, IgakuQA dataset [12] and JJSIMQA. JJSIMQA is a dataset comprising 5-choice questions in JJSIM (see Appendix A). We evaluated the output responses of each model by three metrics. *Gestalt score* is defined as the average of 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. *Accuracy* denotes the correctness rate by considering the choice closest to the model's response when measured using Gestalt distance as the model's output. Finally, *Exact match* is defined as the proportion of responses that contain the correct answer. All evaluation metrics above take the value between 0 and 1, and the larger the better.

All results are summarized in Table 1, where the performance by plain ChatGPT (gpt-3.5-turbo) is added to the last column for reference. Top 2 scores in each row are highlighted in bold.

4 Discussion

4.1 Numerical Evaluation of the Effects of Fine-tuning

We have 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.

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

4.2 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 when compared to 0-shot. This outcome highlights the fact that the original OpenCALM model clearly loses its capability to leverage example responses provided within the context, whereas Llama2-70B retains this ability even after instruction-tuning.

4.3 Comparison on Evaluation Metrics

The evaluation of LLMs have been intensively argued recently. Regarding the evaluation method of LLMs, there is still no unified "rule-of-thumb" method yet. While the existing metrics (e.g. JGLUE [13]) or leaderboards (e.g. Nejumi LLM leaderboard [4]) can assess the fluency of generated texts, they do not adequately evaluate the accuracy of domain-specific knowledge. We wish to point out that three metrics used in our experiments also exhibit certain shortcomings. For example, the use of multiple-choices question for evaluating LLMs has been controversial [14, 15]. *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. The development of even more superior evaluation metrics is eagerly anticipated.

4.4 Difficulty and Limitations

Not to mention that numerous LLM training techniques are still in the developmental stage, training medical LLMs presents several limitations. 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. Moreover, data cleansing has consistently posed challenges, and achieving perfect results in this work may not have been feasible.

Moreover, during the writing period of this paper, Japanese LLMs that are considered to perform better than OpenCALM, which was used in this study, have been released (see e.g. Rakuda benchmark[5]). There is a possibility of obtaining different results when using them as the base model.

5 Conclusion and Future Work

In this paper, we explore the capabilities and limitations of LoRA through various comparative analyses in 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 by additional pretraining under 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.

That being said, the improvement in performance is not sufficient regarding the clinical implementations. 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.

[4] `http://wandb.me/nejumi`

[5] `https://yuzuai.jp/benchmark`

Acknowledgment

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

References

[1] 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. Large language models encode clinical knowledge. *Nature*, pages 1–9, 2023.

[2] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[3] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

[4] Kaito Sugimoto, Taichi Iki, Yuki Chida, Teruhito Kanazawa, and Akiko Aizawa. 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.

[5] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv e-prints*, pages arXiv–2305, 2023.

[7] Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. *arXiv preprint arXiv:2309.03412*, 2023.

[8] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

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

[10] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[11] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

[12] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*, 2023.

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

[14] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[15] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. On large language models' selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*, 2023.

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

A Medical Q&A dataset

For performance evaluation, we prepare two medical question-answer dataset, IgakuQA and JJSIMQA.

An example from IgakuQA

"problem_id": "116A1",
"problem_text": "睡眠時無呼吸症候群による高血圧について、正しくないのはどれか。",
"choices": {"a":"夜間高血圧となることが多い。", "b":"肥満患者では減量を推奨する。",
"c":"α遮断薬が第一選択薬である。", "d":"家庭血圧では早朝に高血圧となることが多い。",
"e":"持続的気道陽圧法〈CPAP〉で降圧が期待される。"},
"text_only": true,
"answer": ["c"],

An example from IgakuQA (translation of the above)

"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)

"problem_text": "近年の日本のB型肝炎について誤っているものはどれか.1つ選べ.",
"choices":{"a":"HBs 抗原陽性率は，母子感染防止事業開始により著減した.","b":"HBV(hepatitis B virus) ゲノタイプ Ae 型は成人の水平感染でキャリア化することがある.","c":"本邦では，2016 年 10 月から HBV ワクチン定期接種が始まった.","d":"HBV ゲノタイプ C は東北地方，宮古八重山地方に多い.","e":HBV の小児期での水平感染は，父子感染や集団生活が一因と考えられる."},
"text_only": true,
"answer": ["d"],

An example from JJSIMQA(5-choice questions in JJSIM) (translation of the above)

"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"],

B Sample questions and responses

We randomly create questions that ask to each model the treatment to a symptom. Table 2 shows the responses of each model to the given input "### Instruction: Please provide detailed instructions

Table 2: Sample responses to the input "### Instruction: Please provide detailed instructions for the treatment to be administered to patients with the following diseases. ### Input: deep vein thrombosis"

model	LoRA/QLoRA	response (originally in Japanese)
OpenCALM	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.
OpenCALM	1 epoch of LoRA	The procedure to be carried out for patients with the following disease involves the implementation of appropriate treatment and preventive measures
OpenCALM	10 epoch of LoRA	Detailed treatment methods have not been provided.
Llama2-70B-chat-hf	none	In some cases, a procedure called thrombol [7]
Llama2-70B-chat-hf	900 steps of QLoRA	The treatment of deep vein thrombosis involves the use of thrombolytic agents and anticoagulants.
Llama2-70B-chat-hf	3000 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.

for the treatment to be administered to patients with the following diseases. ### Input: deep vein thrombosis"[6].

C Instruction data generation by ChatGPT

The prompt we use to create instruction dataset is the following.

Prompt to generate instruction dataset by ChatGPT[8]

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 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}

D Hyperparameters

Parameters in fine-tuning steps are listed in Table 3. All follows the default parameter setting of PEFT library and qlora library, respectively.

For the text generation part, temperature is set to 0.1, max_new_tokens is set to 256, top_p is set to 0.9, and repetition_penalty is set to 1.05.

[6]Originally in Japanese.

Table 3: LoRA parameters

base model	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

For OpenCALM-7B (Japanese), zero-shot

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。
指示:
{instruction}
入力:
{input}
応答:

For Llama2-70B-chat-hf (Alpaca-format [16]), zero-shot

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:

Figure 1: Prompt templates used in our instruction-tuning

E Prompt Templates

In order to let models answer the given questions by text generation function, we provided the following prompt texts, Japanese one for OpenCALM-based models and English one for Llama2-based models. Both of them have the same meaning in each language.

For each sample from Medical Q&A dataset, "problem_text" is incorporated in {instruction} and "choices" is incorporated in {input} in Figure 1.