Development and bilingual evaluation of Japanese medical large language model within reasonably low computational resources

Issey Sukeda
Department of Mathematical Informatics, Graduate School of Information Science and Technology,
The University of Tokyo,
Tokyo, Japan
sukeda-issei006@g.ecc.u-tokyo.ac.jp

Abstract

The recent success of large language models (LLMs) and the scaling law has led to a widespread adoption of larger models. Particularly in the healthcare industry, there is an increasing demand for locally operated LLMs due to security concerns. However, the majority of high quality open-source LLMs have a size of 70B parameters, imposing significant financial burdens on users for GPU preparation and operation. To overcome these issues, we present a medical adaptation based on the recent 7B models, which enables the operation in low computational resources. We compare the performance on medical question-answering benchmarks in two languages (Japanese and English), demonstrating that its scores reach parity with or surpass those of currently existing medical LLMs that are ten times larger. We find that fine-tuning an English-centric base model on Japanese medical dataset improves the score in both language, supporting the effect of cross-lingual knowledge transfer. We hope that this study will alleviate financial challenges, serving as a stepping stone for clinical institutions to practically utilize LLMs locally. Our trained model and evaluation code will both be available at `https://github.com/stardust-coder/japanese-lm-med-harness`.

1 Introduction

In recent years, while the development of LLMs in the medical field has been progressing, there still remains a significant gap between their development and practical application in clinical settings. One of these gaps is the operational environment of LLMs. Current medical LLMs can broadly be divided into two types: the large models developed in closed fashion by big tech companies [38, 39, 48, 30] and the open-source models with fewer parameters. The former is said to have anywhere from tens of billions to trillions of parameters or more and is not freely accessible, or is typically accessible only via API services. This raises security concerns for clinical institutions dealing with patient's personal information, causing hesitation in implementing them in real clinical settings. On the other hand, to improve customizability and accessibility, many other medical LLMs have been released, most of which are based on the Llama series [46, 47], a series of open-source LLMs developed by Meta Inc. The number of parameters in these model developments falls into two main categories: around $7 \sim 8$B and 70B. Generally, larger models outperform smaller ones, a phenomenon known as the scaling law [22]. However, smaller models are computationally more efficient in pretraining, fine-tuning, and inference. For each clinical institute, it is challenging to allocate sufficient computational resources due to budget constraints and other factors. To enable the practical use of medical LLMs, achieving substantial performance with smaller LLMs like $7 \sim 8$B models that can operate within each institute's realistic computational environments is essential.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Since many existing medical LLMs are English-centric, it is believed that there is a strong push in non-English-speaking countries to develop similar medical LLMs in the native languages of users, such as patients and doctors, from a practical standpoint, as this would be more user-friendly. This is done by integrating medical domain adaptation and language adaptation. Particularly in Japan, although several studies have evaluated the capabilities of the commercial GPT models in handling medical queries [23], the number of local model developments lags behind compared to English-centric models [42, 43]. Details about the related works is deferred to Appendix A.

The aim of this study is to verify whether we can avoid operating LLMs as large as 70B-parameter or more; i.e., we aim at achieving enhancement or extraction of Japanese medical intelligence only using limited computational resources, thereby enabling the resulting model to be deployed across numerous clinical institutes in Japan. We conduct the evaluation based on accuracy using the only existing Japanese medical benchmark IgakuQA [23], which is essentially the NMLE (National Medical Licensing Examination in Japan). However, instead of aiming to develop models specifically to pass the medical licensing exam, our focus here is to achieve substantial performance with a 7B medical LLM.

2 Method

We follow the standard way of developing domain-specialized LLMs. First, we prepare the medical corpus from a medical journal and conducted the full-parameter training. We call this process MFPT (Medical Full-Parameter Training) in this work. Subsequently, we conduct the LoRA (Low Rank Adaptation) [17] fine-tuning using question-answering (Q&A) dataset. In this work, we refer to this process as MPEFT (Medical Parameter-Efficient Fine-Tuning).

2.1 Models

For the English-centric base model, Llama3 [12] or Qwen2 [1] have recently been two choices due to their superior performances (see e.g. [13]). We build upon the Qwen2 [1] as the base model architecture for medical adaptation since Qwen2 generally performs better than Llama3 in solving medical benchmark tasks as in Table 2 and Table 4.

We specifically focus on the 7B-parameter model of this series, which enables us to perform full-parameter training even with limited computational resources. Moreover, the Qwen2 series [1] is released under the Apache-2.0 license, which provides significant advantages for broader use.

2.2 Fine-tuning

2.2.1 Medical Full-Parameter Training (MFPT)

First, we conduct the full-parameter training using our own medical corpus, following the established approach by previous researches suggesting that continual-pretraining or additional training before instruction-tuning is effective [38]. Specifically, we used the *naika-text* corpus, which is composed of 6120 lines of Japanese sentences (3.5M letters) extracted from the Journal of the Japanese Society of Internal Medicine. The model after MFPT for 5 epochs is referred to as *Ours-MFPT* in Table 1.

2.2.2 Medical Parameter-Efficient Fine-Tuning (MPEFT)

Following MFPT, MPEFT is performed using the training split of the USMLE (United States Medical Licensing Examination), which includes 10178 training examples in a Q&A format. Since our goal is to develop a Japanese LLM, this data — originally in English — is translated into Japanese by a medical doctor manually[1]. On the other hand, we also use the original English USMLE for comparative experiments. Note that we do not include any data in training dataset from IgakuQA because the IgakuQA dataset is not large[2].

In this procedure, we apply LoRA (Low Rank Adaptation) [17], a parameter-efficient fine-tuning method that can drastically save computational resources (especially GPU memory) without signifi-

[1]This translated dataset will not be made public.

[2]Otherwise, to avoid data leakage, we need to split the dataset into training and evaluation sets, further reducing their sizes.

cant performance loss compared to full-parameter training. To compare each contribution of MFPT and MPEFT, we apply it to both the base Qwen2-7B-Instruct and *Ours-MFPT* model, respectively. *Ours-MFPT* after MPEFT for 5 epochs is referred to as *Ours-MPEFT* in Table 1.

2.3 A unified evaluation method

In medical LLM research, numerous studies have reported benchmark scores derived from question-answering tasks. These benchmarks facilitate comparative analysis. However, even when the test datasets are identical across studies, variations in the experimental settings surrounding LLM inference often preclude truly equitable comparisons. In our paper, we report unified evaluation scores measured by our own experiments, instead of quoting those from previous studies. To facilitate the evaluation method presented in this paper and in future works, and to further encourage the development of medical LLMs, we will make our evaluation codes, which can be executed with a single line of script, publicly available with customization options.

2.3.1 Benchmark dataset for evaluation

We curate four bilingual medical benchmarks in Japanese and English to assess model performance and language tendencies. IgakuQA features five-choice questions, while the other benchmarks use four-choice questions.

IgakuQA [23] is constructed based on the national medical license exam from 2018 to 2022 in Japan. Both the original Japanese dataset and the English-translated dataset are released at `https://github.com/jungokasai/IgakuQA`, including 1450 five-choice questions and answers.

MedQA [20] is composed of the medical license exam in the US, USMLE for short. We only include its evaluation split with 1273 samples in our benchmark. Since the original dataset is in English, the Japanese-translated version was prepared by hand.

MedMCQA [33] (Multi-Subject Multi-Choice Dataset for Medical domain) is a four-choice question-answering task designed to address real-world medical entrance exam questions. We only include its evaluation split with 4183 samples in our benchmark. Since the original dataset is in English, the Japanese-translated version was prepared by hand.

MMLU [15] (Massive Multi-task Language Understanding) is a four-choice Q&A dataset including 57 tasks covering various subjects. We extract five medical-related subjects, i.e., anatomy (135 samples), clinical knowledge (265 samples), college medicine (173 samples), medical genetics (100 samples), and professional medicine (272 samples).

JMMLU [58] is a Japanese-translation of a subset of MMLU, recently prepared as a counterpart of MMLU in our own language. Our evaluation covers anatomy (132 samples), clinical knowledge (150 samples), college medicine (151 samples), medical genetics (99 samples), and professional medicine (150 samples).

2.3.2 Task and evaluation metric

Experimental settings in the inference side generally include prompting, metric, and the hyperparameter of the generation process.

Prompting Prompting mainly consists of the following three factors, which are not consistent at all in the previous reports: (i) prompt template (ii) the number of few-shot examples (iii) other algorithmic prompting techniques. For the template, while Alpaca [44] has been the defacto standard, MedPaLM-2 [39] and Meditron [6] are evaluated with a slightly different prompt template, respectively. The number of few-shot examples is still controversial. Basically, the larger the better but when the prompt becomes too long, the model tends to ignore the former instructions. In addition, to improve the benchmark scores, Chain-of-thought (CoT) prompting [54] is commonly used, followed by self-consistency [52] in MedPaLM [38], and ensemble refinement in MedPaLM-2 [39]. In our experiments, the standard CoT prompt is applied; see Appendix C for the detailed prompting strategy.

model	base model	size	license	MFPT	MPEFT	en(%)	ja(%)	Ave.
Ours-LoRA(en)	Qwen2	7B	CC-BY-NC-SA-4.0	-	USMLE(en)	47.7	41.5	44.6
Ours-LoRA(ja)	Qwen2	7B	CC-BY-NC-SA-4.0	-	USMLE(ja)	51.1	48.6	49.8
Ours-MFPT	Qwen2	7B	CC-BY-NC-SA-4.0	naika-text(ja)	-	47.3	46.6	46.9
Ours-MPEFT(en)	Qwen2	7B	CC-BY-NC-SA-4.0	naika-text(ja)	USMLE(en)	46.2	44.2	45.2
Ours-MPEFT(ja)	Qwen2	7B	CC-BY-NC-SA-4.0	naika-text(ja)	USMLE(ja)	50.6	52.3	**51.4**

Table 1: Benchmark accuracy of our models evaluated with IgakuQA in English(en) and Japanese(ja).

Metric In multiple-choices Q&A tasks, evaluations based on accuracy is common, i.e., we let the model pick one choice as its response and compare it with the correct answer. The typical method involves having models select each answer from five labels ["a", "b", "c", "d", "e"] and verifying by exact match — an approach also used by Kasai et al. [23]— we instead instruct the model to output the words or sentences from the alternatives directly, as the labels themselves do not provide essential information in practice.

To calculate accuracy, *Exact match* has been the most objective and the easiest metric. Also, *Gestalt accuracy* was proposed as an alternative by Sukeda et al. [42, 43], which is a more robust metric to admit a slight mistake for LLMs. In our experiments, we employ the *Gestalt accuracy*.

Hyperparameter of the generation process Commonly, the trainer and the inference pipeline of LLMs is implemented by huggingface transformers [55], which requires to specify several hyperparameters for text generation, e.g., the sampling method, the number of beam for beam search, the temperature for sampling, repetition penalty, and so on. Practically, high temperature and high repetition penalty along with sophisticated sampling methods are recommended to achieve natural and various text generation. On the other hand, in many studies, deterministic results are more preferable for reproducability and thus beam search with one beam is typically employed but not in every case. In our following experiments, we employ the deterministic setting.

2.3.3 Miscellaneous

Although several ad-hoc techniques to improve LLMs' performance including few-shot prompting [5], mixture of experts/agents [50], and model merging [2], have been developed recently, we do not employ these techniques in our main study, as we expect their application can be independently dissociated from the core potential of the LLM that we aim to examine. In other words, these techniques can be readily integrated to one another in practical use cases.

3 Results

3.1 Our resulting model

Table 1 lists the five types we created, along with their training data and the accuracy of IgakuQA in both English and Japanese. We performed fine-tuning of the model with five different settings by varying the training steps. The Ours-LoRA models surpass MFPT (fine-tuning with low-rank adaptation) on the base Qwen2 model [1] in both English and Japanese. In contrast, Ours-MFPT involved full parameter tuning as described in Section 2.2.1. Additionally, Ours-MPEFT, which is based on Ours-MFPT, further improves upon MPEFT in the same manner.

From the accuracy of IgakuQA, it is observed that both the MFPT and MPEFT processes have steadily contributed to score improvement. Specifically, for Japanese IgakuQA, Ours-MPEFT(ja) model achieved a 10.8% increase in accuracy compared to the base model, while it improved by 2.5% in English, though the improvement was smaller. The score of this model exceeds 50% accuracy in both English and Japanese, exhibiting substantial bilingual performance as a 7B model. Hereafter, we will refer to our best model Ours-MPEFT(ja) as JMedLLM-v1-7B.

3.2 Comparison with other open-source LLMs

Setup All generation is performed in zero-shot, meaning no example input-output pairs are provided to the LLM. The 7B models are used as-is, while all 70B models are loaded using 4-bit quantization techniques by default to conserve computational resources. We did not conduct multiple runs or observe deviations; instead, we employed deterministic inference without sampling.

Models Our baseline includes all of the major medical LLMs from previous works. Each of these models are developed by continual training or fine-tuning on each base model, mostly Llama series. For English-centric models, Meditron [6] and Med42 [9] are Llama2-based while OpenBioLLM [32] and Med42-v2 [8] is Llama3-based. For Japanese-centric models, Llama3-Preferred-MedSwallow-70B is a recent Llama3-Swallow-based medical LLM reported to achieve better accuracy than GPT-4[3] in solving NMLE under their experimental settings[4]. Different from the Llama2 series, the Llama3 series is empirically known to have substantial multilingual ability [12], thus we also evaluate both English and Japanese performance for the models derivedd from the Llama3 series. In addition, we add to our baselines the base models for general purpose, which are used as the backbone of each medical LLM. Specifically, Llama3 [12] is added as the English-centric baseline while Llama3-Swallow [27] is added as the Japanese-centric baseline. Moreover, for the 7$\sim$8B scale, two Japanese-centric models — Youko [37] and Llama-3-ELYZA [16] — are added for comparison. In our experiments, we utilize the instruct version whenever available. The link to each specific model is listed in Appendix B.

Results In Table 2, we observe that at the 7$\sim$8B scale, JMedLLM-v1 outperforms other baselines including base models and medical LLMs on average in four Japanese medical benchmarks by more than 10%. It is notable that even at the 70B scale, JMedLLM-v1 outperforms other baselines except Llama3-Preferred-MedSwallow, surpassing 50% accuracy in IgakuQA, MedQA, and MedMCQA. Specifically, JMedLLM-v1 outperforms 70B-parameter OpenBioLLM by as much as 7.5%. One of the possible reasons of the improved performance is that JMedLLM-v1 is based on Qwen2 as its backbone, which achieves superior performance to Llama3 series with the same size. However, in Table 3, it is shown that additional training on medical dataset signifcantly contribute to the score improvement despite the discrepancy between training data and evaluation benchmarks in general. In fact, except the cases of solving Japanese medical benchmarks with OpenBioLLM, the among four models show substantial improvement. Especially, JMedLLM-v1 is further improved and outperforms the base Qwen2 by 13.9%.

On the other hand, in Table 4, we observe that JMedLLM-v1 outperforms other models of similar size on average across four English medical benchmarks, being the only model to surpass 50% accuracy. Despite being 10 times larger, the existing English-centric medical LLMs, Meditron [6] and Med42 [9], do not outperform JMedLLM-v1. However, two 70B-parameter models, OpenBioLLM and Llama3-Preferred-MedSwallow, achieve higher scores than JMedLLM-v1, surpassing 60% in Gestalt accuracy on average. Table 5 further exhibits the score improvement of each LLM, where we can see JMedLLM-v1 has improved by 13.3% in Gestalt accuracy on average also in English medical benchmarks.

Overall, Llama3-Preferred-MedSwallow scores the highest among Japanese medical models, followed by our JMedLLM-v1. OpenBioLLM [32] performs best in English medical tasks but performs worse in Japanese. However, Llama3-Preferred-MedSwallow and JMedLLM-v1 also show strong bilingual performance. Among 7B parameter models, our model stands out as the best performer .

3.3 Comparison with the state-of-the-art

Table 6 shows the gap between the top-3 open-source models from Table 4 and three closed models: Med-Gemini [36], Med-PaLM2 [39], and GPT-4. Although these score comparisons are for reference only, as the evaluation settings are not completely aligned and each score is taken from previous reports, the closed models generally outperform the open-source models. Med-Gemini-L 1.0 [36] is outstanding in MedQA, achieving 91.1% accuracy. Meanwhile, OpenBioLLM [32], with 70B parameters, approaches GPT-4, particularly in the MedMCQA scores.

3.4 Low computational resources

Our MFPT phase took only 7.5 hours on 8 NVIDIA A100 GPUs. Our MPEFT phase took only 28.5 hours on 4 NVIDIA V100 GPUs. All experiments were conducted using ABCI, a Japanese domestic cloud computing infrastructure. These computational burdens are significantly smaller than Meditron

[3] `https://openai.com/index/gpt-4/`
[4] `https://tech.preferred.jp/ja/blog/llama3-preferred-medswallow-70b/`

model(-size)	base	language	IgakuQA(ja)	MedQA(ja)	MedMCQA(ja)	JMMLU	Ave.(ja)
Llama3-70B[27]	Llama3	en	43.1	**40.9**	37.2	45.3	**41.6**
Llama3-Swallow-70B[27]	Llama3	ja	44.6	32.9	33.7	37.5	37.2
OpenBioLLM-70B[32]	Llama3	en	35.6	35.4	**39.9**	**54.6**	41.4
Llama3-Preferred-MedSwallow-70B	Llama3	ja	**62.6**	**55.6**	**43.4**	**58.4**	**55.0**
Llama3-8B[12]	Llama3	en	23.8	28.7	31.7	30.8	28.8
Youko-8B[37]	Llama3	ja	33.5	31.1	34.0	35.7	33.6
Llama3-Swallow-8B[27]	Llama3	ja	28.5	26.9	31.3	30.4	29.3
Llama-3-ELYZA-JP-8B[16]	Llama3	ja	38.0	33.4	34.1	42.9	37.1
MMedLlama3-8B[35]	Llama3	en	31.5	34.3	33.9	40.1	35.0
Qwen2-7B[1]	Qwen2	en	**44.6**	30.8	31.5	33.2	35.0
JMedLLM-v1-7B (Ours)	Qwen2	ja	**52.3**	**51.2**	**41.2**	**50.8**	**48.9**

Table 2: **JMedLLM-v1 against open-source baselines in Japanese medical benchmarks.** This table shows the main results of JMedLLM-v1's medical task performance in Japanese against other best-performing open-source medical LLMs measured by the Gestalt accuracy(%). Top 3 scores in each row are marked in bold.

base LLM	→	Japanese LLM	IgakuQA	MedQA	MedMCQA	JMMLU
Llama3-70B [12]	→	Llama3-Swallow-70B [27]	+1.5	−8.0	−3.5	−7.8
Llama3-8B [12]	→	Youko-8B [37]	+9.3	+2.4	+2.3	+4.9
Llama3-8B [12]	→	Llama3-Swallow-8B [27]	+4.7	−1.8	−0.4	−0.4
Llama3-8B [12]	→	Llama-3-ELYZA-JP-8B[16]	+14.2	+4.7	+2.4	+12.1

base LLM	→	medical LLM	IgakuQA	MedQA	MedMCQA	JMMLU
Llama3-70B [12]	→	OpenBioLLM-70B [32]	−7.5	−5.5	+2.7	+9.3
Llama3-Swallow-70B [27]	→	Llama3-Preferred-MedSwallow-70B	+18.0	+22.7	+9.7	+20.9
Llama3-8B [12]	→	MMedLlama3-8B[35]	+7.7	+5.6	+2.2	+9.3
Qwen2-7B[1]	→	JMedLLM-v1-7B (Ours)	+7.7	+20.4	+9.7	+17.6

Table 3: **Score improvements of LLMs compared to each base model in Japanese benchmarks.** This table shows the difference in scores between each medical LLM and its corresponding base model, as presented in Table 2.

(332 hours on 128 A100 GPUs for training) and Med42 (the Condor Galaxy 1 supercomputer for full-parameter fine-tuning). [5]

On the other hand, all the evaluation experiments for 70B models with quantization were run on 4 NVIDIA V100 GPUs, whereas for 7B models were run on 1 NVIDIA V100 GPUs.

4 Discussion

Model performance and size from a practical use perspective Since LLMs generally perform better as their size increases, a phenomenon known as the scaling law [22], it is inevitable to confront this tradeoff in practice. The aim of our work is to develop 7B-parameter medical LLMs to the fullest extent possible, so that clinical institutes do not necessarily have to rely on external API services or the computationally burdensome 70B models.

In our work, we demonstrate that the 7B-parameter model, with a solid base model and fine-tuning process using a domain-specific corpus, can potentially outperform 70B models in a medical question-answering benchmarks both in Japanese and English. Although its performances shown in Table 2 and 4 are not totally sufficient, this result highlights the potential for the practical use of medical LLMs in clinical institutions, as 7B-parameter models can operate relatively quickly with modest resources, such as a single GPU in a standard environment.

[5] For example, Amazon Web Services (`https://aws.amazon.com/?nc1=h_ls`) provides these GPU instances as cloud environment. The cost for training can be simulated as $32.77 \times 7.5 + 12.24 \times 28.5 = 595$ USD.

model(-size)	base	language	IgakuQA(en)	MedQA(en)	MedMCQA(en)	MMLU	Ave.(en)
Meditron-70B [6]	Llama2	en	29.9	44.7	32.8	49.6	39.3
Med42-70B [9]	Llama2	en	45.0	56.2	**48.2**	60.9	52.6
Llama3-70B[12]	Llama3	en	38.3	**57.7**	38.8	**63.7**	49.6
Llama3-Swallow-70B[27]	Llama3	ja	**52.8**	39.0	43.0	51.2	46.5
OpenBioLLM-70B[32]	Llama3	en	**58.5**	**70.2**	**65.0**	**80.0**	**68.4**
Llama3-Preferred-MedSwallow-70B	Llama3	ja	**55.0**	**61.3**	**52.9**	**68.1**	**59.3**
Llama3-8B[12]	Llama3	en	35.0	43.0	39.1	41.3	39.6
Youko-8B[37]	Llama3	ja	38.1	34.1	29.4	44.6	36.6
Llama3-Swallow-8B[27]	Llama3	ja	34.4	30.8	36.0	38.8	35.0
Llama-3-ELYZA-JP-8B[16]	Llama3	ja	20.7	40.6	37.3	44.7	35.8
MMedLlama3-8B[35]	Llama3	en	26.4	36.8	37.5	37.7	34.6
Qwen2-7B[1]	Qwen2	en	46.4	36.9	34.7	43.1	40.3
JMedLLM-v1-7B (Ours)	Qwen2	ja	50.6	54.6	46.1	63.0	**53.6**

Table 4: **JMedLLM-v1 against open-source baselines in English medical benchmarks.** This table shows the main results of JMedLLM-v1's medical task performance in English against other best-performing open-source medical LLMs measured by the Gestalt accuracy(%). Top 3 scores in each row are marked in bold.

base LLM	→	Japanese LLM	IgakuQA	MedQA	MedMCQA	JMMLU
Llama3-70B [12]	→	Llama3-Swallow-70B [27]	+14.5	−18.7	+4.2	−12.5
Llama3-8B [12]	→	Youko-8B [37]	+3.1	−8.9	−9.7	+3.3
Llama3-8B [12]	→	Llama3-Swallow-8B [27]	−0.6	−12.2	−2.9	−2.5
Llama3-8B [12]	→	Llama-3-ELYZA-JP-8B[16]	−14.3	−2.4	−1.8	+3.4

base LLM	→	medical LLM	IgakuQA	MedQA	MedMCQA	JMMLU
Llama3-70B [12]	→	OpenBioLLM-70B [32]	+20.2	+12.5	+26.2	+16.3
Llama3-Swallow-70B [27]	→	Llama3-Preferred-MedSwallow-70B	+2.2	+22.3	+9.9	+16.9
Llama3-8B [12]	→	MMedLlama3-8B[35]	−8.6	−6.2	−1.6	−3.6
Qwen2-7B[1]	→	JMedLLM-v1-7B (Ours)	+4.2	+17.7	+11.4	+19.9

Table 5: **Score improvements of LLMs compared to each base model in English benchmarks.** This table shows the difference in scores between each medical LLM and its corresponding base model, as presented in Table 4.

To implement medical LLMs in real clinical institutions, how far should we make progress? One way to set the performance necessary for practical use is to compare it with large, closed-source LLMs such as GPT-4, the MedPaLM series, and the recent top-performing Med-Gemini. Table 6 shows that open-source medical LLMs still have a significant gap compared to closed-source models in benchmark scores. However, for example, when compared to Med-PaLM2, JMedLLM-v1 with 7B parameters achieves from 60 to 70% of the performance with only less than 2.5% parameters. [6]

Score improvements resulting from fine-tuning in different languages By comparing Tables 3 and 5, it is observed that Japanese adaptation tends to result in only a small improvement in scores or even a decline in performance. In contrast, medical adaptation generally leads to more significant score improvements across medical benchmarks in both languages.

Individually, the fine-tuning of MMedLlama3 [35] in a multilingual context enhances performance in Japanese while causing degration in English. Conversely, OpenBioLLM is English-centric, leading to significant improvements in English benchmarks but sometimes causing a decline in performance

[6] Med-PaLM2 is a closed medical LLM based on PaLM2, which is also a closed model. According to unofficial news sources (`https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html`, accessed 2024/8/10), PaLM2 is reported to have 340B parameters based on internal documents. Official information is not yet available.

Model name	IgakuQA(ja)	MedQA(en)	MedMCQA(en)	MMLU
Med-Gemini-L 1.0	-	91.1$^{(a)}$	-	-
Med-PaLM2	-	85.4$^{(b)}$	72.3$^{(b)}$	88.4$^{(b)}$
GPT-4	78.2$^{(c)}$	78.8$^{(d)}$	69.5$^{(d)}$	86.0$^{(d)}$
Llama3-Preferred-MedSwallow	62.6	61.2	52.9	68.1
OpenBioLLM	35.6	70.2	65.0	80.0
JMedLLM-v1	52.3	54.6	46.1	63.0

Table 6: **JMedLLM-v1 against the state-of-the-art LLMs.** This table shows the scores cited from previous studies. Note that each score differs in inference settings. (a) Cited from [36]. (b) : Cited or calculated from [39], where the ensemble refinement method is used, which is computationally costly. (c) : Calculated based on [23], where 3-shot in-context learning is used. (d) Cited or calculated from zero-shot performances in [29].

in certain Japanese benchmarks. On the other hand, both Llama3-Preferred-MedSwallow and JMedLLM-v1 are finetuned with Japanese dataset, yet they improve performance in both languages.

A significant improvement in Table 3 for JMedLLM-v1 on MedQA and for Llama3-Preferred-MedSwallow on IgakuQA can be attributed to the alignment between their evaluation benchmarks and training datasets, although the details for the latter are not fully disclosed.[7] However, the significant improvement in MMLU/JMMLU is impressive, whereas IgakuQA and MedMCQA show more variable results, which may be related to task difficulty: MMLU/JMMLU covers high school to university levels, MedMCQA targets graduate school level, and IgakuQA and MedQA focus on national exam level.

Knowledge extraction by fine-tuning As mentioned in the previous paragraph, it is well-known that alignment between fine-tuning and evaluation tasks generally leads to significant improvements in model performance. However, our experiments indicate that improvements can occur even in the absence of such alignment. Although USMLE and NMLE (or other benchmark tasks) are situated within the same medical domain, their questions and answer choices are not perfectly identical. The observed enhancements in our MPEFT model suggest that the base model already possesses a foundational level of medical knowledge. Rather than introducing new medical knowledge through fine-tuning with USMLE data, this process appears to activate latent capabilities within the model. These findings imply that training on tasks that are similar but not identical to the target task can still contribute to improved model performance. To further investigate or quantify this phenomenon, it would be necessary to identify the specific information required for accurately answering a specific target question, which currently appears to be technically infeasible.

Cross-lingual knowledge transfer When adapting English-centric LLMs to local languages, is there a trade-off in performance for English while learning a new language? In our English-translated IgakuQA evaluations, as shown in Table 1, an unexpected observation is that our models demonstrated improvements of +1.4% for MFPT and +2.9% for MPEFT, despite the fine-tuning training data consisting solely of Japanese texts. This phenomenon is also evident in Table 5 for JMedLLM-v1 and Llama3-Preferred-MedSwallow. Although these improvements are smaller compared to those observed in Japanese benchmarks listed in Table 3, they are still noteworthy. A similar phenomenon is observed in our comparative LoRA experiments without the MFPT step, where accuracy improved by 1.8% in English and by 5.2% in Japanese compared to the base model (Qwen2-Instruct [1]). Overall, our results suggest that cross-lingual training on English-centric models can be effective not only for acquiring local languages but also for enhancing their performance in English.

[7] According to a tech blog by the developers of Llama3-Preferred-MedSwallow, the fine-tuning dataset includes past NMLE data up to 2017. URL: `https://tech.preferred.jp/ja/blog/llama3-preferred-medswallow-70b/`

5 Limitations

5.1 Insufficient data resources

The amount and variation of medical corpus have been insufficient for training LLMs, particularly in Japanese. In our study, we utilized medical examinations from the US as a data resource; however, this may introduce a risk of reflecting biases inherent to US medicine. Not only should the selection and preparation of the training dataset be further improved, but bias correction methods across different countries, cultures, and contexts also need to be studied further to ensure practical use.

5.2 Exploration on evaluation method

Question on multiple-choice question-answering as benchmarks This work does not explore the validity of the evaluation method in depth; instead, we prioritize unification. However, in the study of LLMs, the development of a fair evaluation method is eagerly anticipated. Evaluating the performance of medical LLMs with question-answering task, which is often based on the medical licensing exam, is questioned [28] in terms of the risk reproducing social biases in clinical decision making.

Mismatch prompt formatting in training and evaluation Some LLMs are designed to adhere to specific prompt formatting, especially when instruction-tuning [53] is involved. Empirically, 70B models are sufficiently generalized and capable of handling variations in prompt formatting, whereas 7B models tend to perform worse in this regard. Nonetheless, LLMs are expected to perform optimally when the prompt format during inference is specified correctly. For instance, Meditron [6] follows the ChatML format [31], whereas our prompting strategies differ significantly (see Appendix C). This discrepancy may contribute to the poorer performance of Meditron observed in Table 4.

Tokenizer specification We also point out that our models employ the *Byte-Pair Encoding tokenization* [49] as well as the backbone Qwen series. The use of it for LLMs has been argued [4], and may not be optimal especially for Japanese LLMs.

5.3 Data contamination

Typically, the amount of training corpus for base models is extensive and not entirely publicly available. Therefore, although many reports assert that the fine-tuning process does not explicitly allow for contamination, it is not possible to guarantee that the evaluation datasets used for benchmarking (such as IgakuQA, MedQA, MedMCQA, and MMLU/JMMLU in our case) are free from contamination. Once the contamination occurs, it causes data leakage, which artificially inflates benchmark scores. For instance, Sukeda et al.[41] demonstrate in an ablation study that fine-tuning an LLM on USMLE and then evaluating it with the same data can easily lead to accuracy surpassing 80%. Therefore, it should be noted that a significant leap in accuracy shown in Table3 or 5 might be at risk of being caused by data leakage unless the data usage is clearly specified.

5.4 Model quantization

We regret that we used 4-bit quantization for evaluating all the 70B-parameter models due to limited computational resources. While quantization is generally known to speed up inference and potentially degrade performance, it is sometimes argued that 4-bit quantization can still maintain good performance in practice. Quantitatively evaluating the degradation caused by quantization will be a focus of future work.

6 Conclusion

In this work, we develop the leading 7B medical LLM and demonstrate that it achieves performance comparable to or better than existing 4-bit-quantized 70B-parameter medical LLMs on Japanese medical Q&A benchmarks. Moreover, our model also performs well on English counterparts even without additional training on English data. Our 7B model can be trained and operated in environments with limited GPU resources, addressing financial and security concerns for clinical institutes seeking to adopt practical, medical-specific LLMs.

7 Acknowledgments and Disclosure of Funding

The Japanese-translated version of USMLE was provided by Dr. Hisahiko Sato (not made public). The Japanese-translated version of MedQA and MedMCQA were provided by Mr.Junfeng Jiang (not made public). We also thank the developer of SWIFT [45], which our implementation for training models is based on, and Dr. Jun Sese and Dr. Shinnosuke Sawano for helpful comments. This work was supported by AIST KAKUSEI project (FY2023).

References

[1] Qwen2 technical report. 2024.

[2] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary Optimization of Model Merging Recipes. *arXiv preprint arXiv:2403.13187v1*, 2024.

[3] Andreas Geert Motzfeldt Aryo Pradipta Gema Ankit Pal, Pasquale Minervini and Beatrice Alex. openlifescienceaiopen_medical_llm_leaderboard. `https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard`, 2024.

[4] Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*, 2020.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[7] Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Clément Christophe, Tathagata Raha, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Nada Saadi, Hamza Javed, Umar Salman, Svetlana Maslenkova, Marco Pimentel, Ronnie Rajan, and Shadab Khan. Med42-v2 - a suite of clinically-aligned large language models. 2024.

[9] Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. 2024.

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

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily

Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan,

Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, 2024.

[13] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

[14] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. *arXiv preprint arXiv:2404.17790*, 2024.

[15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[16] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[17] 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.

[18] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[19] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[20] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*, 2020.

[21] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 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, 2019.

[22] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[23] 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.

[24] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. *Plos one*, 16(11):e0259763, 2021.

[25] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1):170, 2023.

[26] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv preprint arXiv:2402.10373*, 2024.

[27] Swallow LLM. Llama 3 Swallow, 2024.

[28] Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E Priebe, and Eric Horvitz. Medfuzz: Exploring the robustness of large language models in medical question answering. *arXiv preprint arXiv:2406.06573*, 2024.

[29] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[30] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[31] OpenAI. ChatML. *https://github.com/openai/openai-python/blob/release-v0.28.0/chatml.md*, 2023. Accessed 2024-08-09.

[32] Ankit Pal and Malaikannan Sankarasubbu. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`, 2024.

[33] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[34] Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX: Bilingual Medical Mixture of Experts LLM. *arXiv preprint arXiv:2402.13253*, 2024.

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

[36] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of Gemini Models in Medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[37] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5 2024.

[38] 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.

[39] 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.

[40] 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 *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*, 2023.

[41] Issey Sukeda, Risa Kishikawa, and Satoshi Kodera. 70B-parameter large language models in Japanese medical question-answering. *arXiv preprint arXiv:2406.14882*, 2024.

[42] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA: Medical Domain Adaptation on Japanese Large Language Models using Instruction-tuning. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

[43] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. Development and analysis of medical instruction-tuning for Japanese large language models. *Artificial Intelligence in Health*, 1(2):107–116, 2024.

[44] 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.

[45] The ModelScope Team. SWIFT:Scalable lightWeight Infrastructure for Fine-Tuning. `https://github.com/modelscope/swift`, 2024.

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[47] 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.

[48] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[49] Changhan Wang, Kyunghyun Cho, and Jiatao Gu. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160, 2020.

[50] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[51] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*, 2023.

[52] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[53] 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.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[55] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[56] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.

[57] Xwin-LM Team. Xwin-LM, 9 2023.

[58] Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. *arXiv preprint arXiv:2402.14531v1*, 2022.

A Related works

The relevant models and their training dataset are curated in Table 7 as a reference.

A.1 Before LLMs

The Japanese language model has been developed following precedents set by English research, particularly in the medical domain. Recently, Japanese medical language model research was ignited by UTH-BERT [24], which was developed by pretraining BERT [11] with approximately 120M clinical texts stored at UTokyo Hospital as the first medical language model in Japanese. Afterwards, JMedRoBERTa [40] was developed based on RoBERTa model using the abstract and main text of the non-public medical papers.

A.2 Medical LLMs in English

In open source community, two lines of family have been developed: the Llama family [46, 47] and the Mistral family [18, 19]. Generally, the Llama family tends to be a single model while the Mistral family has evolved in the direction of the Mixture-of-Experts. Specifically in biomedical areas, several medical LLMs have been built upon 70B-parameter Llama2 or Llama3 [42, 6, 9, 32, 7], while 7B-parameter Biomistral [26] and BiMediX [34] have been derived from Mistral 7B [18] and Mixtral-8x7B [19], respectively.

A.3 Medical LLMs in Japanese

Compared to English-centric models, the Japanese medical LLMs lack its number. Instead of developing the model from scratch, these models are developed based on powerful English-centric models. The first attempt in this domain was JMedLoRA [42], which conducted the QLoRA [10] instruction-tuning on Llama2-70B [39]. After the Japanese general LLM named Swallow [14] was released, Sukeda et al. [41] performed the similar fine-tuning on Llama2 [39], Xwin [57], and Swallow [14], suggesting the potential of Japanese base model to be improved largely in medical question-answering by instruction-tuning. Furthermore, after the release of Llama3-Swallow [27], which is based on Llama3 [12], Preferred Network Inc. performed the continual training via QLoRA [10] using their non-public medical training data, which is released as Llama3-Preferred-MedSwallow.

A.4 Medical benchmarks

To develop the domain-specific LLMs, the evaluation benchmarks are of great importance. Here, we review the existing medical evaluation benchmarks here.

MultiMedBench [48] is an open source multimodal medical benchmark developed for assessing the multimodal medical model named MedPaLM-M, including three of the MultiMedQA tasks used to evaluate MedPaLM [39], and radiology report summarization.

Model name	Training dataset	#Data
Llama2 [47]	See [47].	2T tokens
Meditron [6]	Clinical Guidelines, PubMed Abstracts, PubMed Papers, Experience Replay	0.107B tokens 5.48B tokens 40.7B tokens 0.420B tokens
Med42 [9]	See [9].	411064 medical samples 295649 general domain samples
Swallow [14]	Swallow corpus, Japanese Wikipedia, the RefinedWeb The Pile	312.1B characters 104.9B tokens
MedSwallow [41]	Japanese-translated USMLE	12723 samples
Llama3 [12]	Web-curated multilingual data covering 176 languages	15T tokens
Youko [35]	Japanese CC-100, Japanese C4, Japanese OSCAR, The Pile, Wikipedia, rinna curated Japanese dataset	22B tokens
Qwen2 [1]	Multilingual data supporting 30 languages	7T tokens
OpenBioLLM [32]	Custom Medical Instruct dataset DPO dataset	unknown
Llama3-Preferred-MedSwallow-70B	own medical corpus	unknown
JMedLLM-v1 (Ours)	naika dataset, USMLE	3.5M characters + 10178 samples

Table 7: Training dataset of existing LLMs. The number of tokens is presented for each dataset if available. Otherwise, the number of samples is presented.

MIRAGE [56] is a benchmark for medical LLMs and retrieval augmented generation(RAG), which includes MedQA [20], MedMCQA [33], PubMedQA [21], MMLU Subsets (Medicine) [15] and BioASQ-QA [25].

The Open Medical-LLM Leaderboard [3] is a standarized platform that provides a setup specifically designed for the medical domain, which includes MedQA [20], MedMCQA [33], PubMedQA [21], and MMLU Subsets (Medicine and Biology) [15].

CMB [51] is a comprehensive medical benchmark in Chinese, which comprises multiple-choice questions from qualification exams, and complex clinical diagnostic questions derived from real case studies.

MMedBench [35] is a comprehensive multilingual medical benchmark including six languages, English, Japanese, Chinese, French, Spanish, and Russian.

B Instruct models

For base models in our experiments, we utilize the instruct version whenever available. Specifically for non-medical base models, we use Meta-Llama-3-70B-Instruct[8], Llama-3-Swallow-70B-Instruct-v0.1[9], Meta-Llama-3-8B-Instruct[10], llama-3-youko-8b[11], and Qwen2-7B-Instruct[12].

[8] `https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct`

[9] `https://huggingface.co/tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1`

[10] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

[11] `https://huggingface.co/rinna/llama-3-youko-8b`

[12] `https://huggingface.co/Qwen/Qwen2-7B-Instruct`

For medical models, we use Meditron-70B[13], Med42-70B[14], Llama3-OpenBioLLM-70B[15], MMedLlama3-8B[16], and Llama3-Preferred-MedSwallow-70B[17].

C Prompting strategies in inference

For the prompt template, we follow the Chain-of-Thought (CoT) prompt [54] used in Med-PaLM 2 [39] by default. Japanese templates are prepared as well through translation by ChatGPT[18]. To let the model solve the given question-answering tasks, the question sentence is input into {{instruction}} and four or five candidate choices are input into {{input}}.

Chain-of-Thought (CoT) prompt

```
### Instruction:
The following are multiple choice questions about medical knowledge. Solve
them in a step-by-step fashion, starting by summarizing the available information.
Output a single option from the five options as the final answer.
### Input:
{{instruction}}
{{input}}
### Response:
```

Although prompt template selection tends to be ad-hoc since we cannot know which one is better than the other in advance, we choose this one as our experimental setting because it seemed to be slightly superior in Sukeda et al.[41]. Few-shot inference has known to be effective strategy, however we do not apply it as our standard experimental setting since the number of shots is always arbitrary. Moreover, providing more examples in a prompt tends to lead to better performance; however, it entails a token limit issue. Therefore, we align the experimental settings in each run by evaluating zero-shot performance.

As comparative studies, we additionally evaluate our best model with different prompting strategies. First we attempt the following standard prompt:

Alpaca [44] prompt

```
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:
```

As a result, Table 8 shows that using the Alpaca prompt template led to slightly worse performances both in English and Japanese. However, in theory, there should be no superiority or inferiority since the meaning of given instruction is almost identical. In experiments by Sukeda et al. [42], the superiority of these two types of prompts reversed depending on the experimental settings, making it difficult to conclude. Hence, a difference in accuracy of a few percentage points may be considered negligible.

Subsequently, we observe a few-shot performance using the following prompt:

Few-shot inference with CoT

[13] https://huggingface.co/epfl-llm/meditron-70b
[14] https://huggingface.co/m42-health/med42-70b
[15] https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B
[16] https://huggingface.co/Henrychur/MMed-Llama-3-8B
[17] https://huggingface.co/pfnet/Llama3-Preferred-MedSwallow-70B
[18] https://chatgpt.com/

Template	en(%)	ja(%)
CoT	**50.2**	**52.5**
Alpaca	49.7	49.7

Table 8: Differences by prompt templates

	en(%)	ja(%)
0-shot with CoT	**50.2**	**52.5**
1-shot with CoT	45.8	45.7
3-shot with CoT	46.7	47.8

Table 9: Difference by the number of few-shot examples

Instruction:
The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the five options as the final answer.
Input:
Which of the following is not a mandatory explanation to be provided to participants in human genomegene analysis research?
The purpose of the research, The freedom to consent, Methods for anonymity, Disadvantages of participation, Assurance of research results sharing
Response:
Assurance of research results sharing
Input:
A 57-year-old man lost consciousness and collapsed while working to remove sludge from a manhole at a sewage treatment plant. A colleague who entered to assist also suddenly lost consciousness and collapsed. Which of the following is the most likely cause? Select two.
Oxygen deficiency, Hydrogen sulfide poisoning, Carbon monoxide poisoning, Carbon dioxide poisoning, Nitrogen dioxide poisoning
Response:
Oxygen deficiency, Hydrogen sulfide poisoning
Input:
A 28-year-old woman at 30 weeks of gestation has a fundal height of 22 cm and almost no amniotic fluid is detected on abdominal ultrasound examination. What is the most likely condition in the fetus?
Esophageal atresia, Ventricular septal defect, Renal hypoplasia, Anorectal malformation, Fetal hydrops
Response:
Renal hypoplasia
Input:
{{instruction}}
{{input}}
Response:

In 1-shot experiments, only the first example was included, while the whole prompt was applied for 3-shot experiments. As shown in Table 9, few-shot prompting technique did not contribute to the score improvement. While few-shot examples are believed to instruct the model via in-context learning, our model already has the ability to follow the instruction to choose one option from five alternatives. Thus, the given few-shot examples may function as noisy information unrelated to the target question.

D Licenses

D.1 Data

The *naika-text* corpus[19] is licensed free. IgakuQA is released without license currently. MedQA, MedMCQA, and MMLU are made public under MIT license. JMMLU (the subset we used) is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.

[19] `https://www.jstage.jst.go.jp/browse/naika/-char/ja`

D.2 Models

Llama2 and its variants follow the LLAMA 2 COMMUNITY LICENSE AGREEMENT (https://github.com/meta-llama/llama/blob/main/LICENSE). Llama3 and its variants follow the META LLAMA 3 COMMUNITY LICENSE AGREEMENT (https://llama.meta.com/llama3/license/). Qwen2 series is released under Apache License 2.0. Our developed models will be licensed later.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contribution and scope, involving the development of medical LLMs under low resources from a practical viewpoint, are reflected both in abstract and introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The trained model is open-sourced and will be disclosed after acceptance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: No, but partly yes. All the experimental settings are mentioned in the paper for reproducibility. Our training data cannot be made open access. Our evaluation data is publicly available from the previous studies and we have provided sufficient instructions in the paper. Our evaluation codes will be provided open access after acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: They are provided in Section 3.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Since the performance of the models in the field discussed in the work is not sufficient yet to be practically useful, we do not discuss any societal impacts but solely focus on the model performance based on technological perspectives. However, the motivation of this study is driven by the privacy and considerations in clinical institutes.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our trained language model will be released with a model card notifying such risks to the users.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All information about the license is listed in Appendix D. We ensure that these licenses are properly respected in our use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The trained model is the only new assets introduced in the paper. The documentation is provided in its model card alongside the publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.