Building Effective Japanese Medical LLMs with an Open Recipe for Domain Adaptation through Continued Pre-training

Akiko Aizawa[1], Yuki Arase[2], Fei Cheng[2], Jiahao Huang[3], Zhiyi Huang[4,†], Junfeng Jiang[3], Teruhito Kanazawa[1], Daisuke Kawahara[5], Kazuma Kobayashi[1,†], Takashi Kodama[2], Sadao Kurohashi[2], Yusuke Oda[1], Yuma Tsuta[1], Zhen Wan[2], Zhishen Yang[1], Rio Yokota[6]

[1]National Institute of Informatics
[2]Kyoto University
[3]University of Tokyo
[4]Institute of Science Tokyo
[5]Waseda University
[6]Institute of Integrated Research, Institute of Science Tokyo
Tokyo, Japan; Kyoto, Japan
{aizawa, tkana, kazumkob, zsyang}@nii.ac.jp, arase@c.titech.ac.jp, feicheng@i.kyoto-u.ac.jp,
jiahao-huang@g.ecc.u-tokyo.ac.jp, {huang.zhiyi, rioyokota}@rio.scrc.iir.isct.ac.jp
jiangjf@is.s.u-tokyo.ac.jp, dkw@waseda.jp, kodama@nlp.ist.i.kyoto-u.ac.jp,
kuro@i.kyoto-u.ac.jp, yusuke.oda@predicate.jp, tsuta@tkl.iis.u-tokyo.ac.jp,
zhenwan@nlp.ist.i.kyoto-u.ac.jp

Abstract

In high-stakes domains such as medicine, ensuring transparency of the training corpus is essential, with careful consideration of local healthcare landscapes; however, the majority of existing medical large language models (LLMs) have not disclosed the details of their training corpora. Here, we introduce an open recipe for domain adaptation of LLMs to the Japanese medical domain. We employed fully open-source Japanese general-domain LLMs as base models, whose pre-training datasets are also disclosed. To establish effective corpora for domain adaptation through continued pre-training, we started with small-scale medical datasets and ultimately constructed a medical corpus consisting of 79.6B tokens, incorporating local clinical guidelines, medical textbooks, and other domain-specific resources. The resulting LLM from continued pre-training, namely `SIP-med-llm-8×13B`, with an active parameter count of 22B, demonstrated favorable accuracy on benchmarks including the Japanese National Medical Examination. This performance was comparable to that of 70B-parameter open-weight models whose construction details remain non-transparent. This represents the first case in the Japanese medical field where complete corpus details have been disclosed for fully from-scratch development, providing important insights for future efforts to construct medical LLMs tailored to the specific characteristics of local contexts. The model is available publicly at this Hugging Face repository: `https://huggingface.co/SIP-med-LLM/SIP-jmed-llm-2-8x13b-OP-instruct`.

Keywords: large language models, medicine, Japanese medical domain, domain adaptation, continued pre-training, medical corpus, Mixture of Experts, transparency

1. Introduction

The integration of large language models (LLMs) into healthcare offers transformative potential in areas such as clinical decision support, patient communication, and medical knowledge discovery. General-domain LLMs, including the GPT series (OpenAI, 2023) and LLaMA (Touvron et al., 2023a), exhibit remarkable natural language understanding capabilities that can manage certain specialized knowledge. However, in the high-stakes medical domain, where accuracy and reliability are critical, domain-specific adaptations are essential to meet the unique requirements of clinical expertise, language, ethics, and regulations. For example, models like Med-PaLM (Singhal et al., 2023) and Meditron (Chen et al., 2023a) have been fine-tuned on biomedical corpora, achieving scores that exceed the human passing threshold on the MedQA benchmark of United States Medical Licensing Examination (USMLE)-style questions (Jin et al., 2020).

Despite these advancements, existing medical LLMs still encounter key challenges, notably the lack of transparency in their construction recipes. This opacity impairs reproducibility and adaptation to local medical contexts, including racial disparities, regional disease prevalence variations, healthcare accessibility differences, and specific treatment protocols. Proprietary models like Med-PaLM (Singhal et al., 2023), for instance, provide limited details on training data, hindering bias evaluations and localized customizations. Even open-

Equal contributions for all authors, listed in alphabetical order.
† Corresponding authors.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10405–10423
11–16 May 2026.

source initiatives such as Meditron (Chen et al., 2023a), while documenting curated medical resources, rely on foundational models like LLaMA-2 with undisclosed data sources and processes. Consequently, most medical LLMs lack end-to-end transparency from scratch, underscoring the need for publicly shared, clear recipes to enable effective, locality-specific adaptations.

Adaptation of medical LLMs into locality-specific clinical contexts is essential for delivering the best patient care that reflects their actual environment and societal context. For example, in Japan, such efforts might include deep knowledge of unique elements, including specialized terminology, universal health insurance coverage, hybrid medicines integrating Western and Eastern practices (e.g., Kampo herbal medicine), and cultural-ethical considerations. Recent developments, such as Preferred-MedLLM-Qwen-72B (Iwasawa, 2025), exemplify effective localization by continuing pretraining on the Qwen2.5-72B base model (Hui et al., 2024) with a Japanese medical corpus, thereby achieving high accuracy on the IgakuQA benchmark (Kasai et al., 2023)–a rigorous Japanese equivalent to the Medical Licensing Exam. However, like other adaptations from open-weight models, Qwen-based efforts still face transparency limitations, stemming from undisclosed pretraining data sources in the base model, which can impede thorough bias evaluations and complete reproducibility in diverse localized contexts.

Here, we address these transparency and localization challenges by introducing an open recipe for domain adaptation through continued pretraining for the Japanese medical domain. By adopting fully open-source Japanese general-domain base models with disclosed pretraining datasets (LLM-jp et al., 2024), we conducted a two-stage approach (see **Figure 1** for the overview of our study). First, in the *corpus expansion studies*, we began by constructing a corpus consisting primarily of publicly available medical texts on the web. Recognizing the relative scarcity of Japanese medical texts compared to English ones, which hinders effective domain adaptation, we overcame this challenge through machine translation, ultimately building a balanced English-Japanese medical corpus consisting of 79.6 billion tokens. These include local clinical guidelines, medical textbooks, and domain-specific resources to align with Japan's unique healthcare context. Model evaluation was conducted using JMedBench (Jiang et al., 2025)–a benchmark suite for comprehensively evaluating performance in the Japanese medical domain–as the primary evaluation metric.

Consequently, we built a mixture-of-experts (MoE) model with 8×13B parameters, namely `SIP-med-llm-8×13B`. Due to its computational efficiency owing to the MoE architecture, it delivers competitive performance on these benchmarks with only 22 billion active parameters, matching larger 70B-parameter models (e.g., Qwen-2.5-72B and its derivatives). This *performance comparison* constituted our second step. Our study represents the first fully disclosed end-to-end development in the Japanese medical domain, providing key insights for developing locally tailored LLMs globally.

Our contributions are threefold:

- Introducing a reproducible, open recipe for domain adaptation that prioritizes transparency in corpus composition and pretraining processes.
- Exploring domain adaptation strategies based on comprehensive evaluations using JMed-Bench, including the necessity of machine translation for adapting to the Japanese medical domain.
- Demonstrating that a transparent MoE model with 22 billion active parameters can match the accuracy of larger, opaque counterparts on rigorous medical benchmarks.

2. Related Work

2.1. Transparency in Medical LLM Construction

Transparency of model architecture, training datasets, methodologies, and update processes is essential to ensure the dependable operation of medical LLMs as components of medical systems in clinical practice (Riedemann et al., 2024). However, full disclosure in medical LLM construction, especially regarding corpus composition and data handling, continues to pose a critical challenge. Frontier models, such as GPT (OpenAI et al., 2024; Arora et al., 2025), Med-PaLM (Singhal et al., 2023, 2025), Gemini (Team, 2025; Pal and Sankarasubbu, 2024), and DeepSeek (DeepSeek-AI et al., 2025; Sandmann et al., 2025; Tordjman et al., 2025), have demonstrated high performance across medical tasks, including medical licensing examinations, differential diagnosis (McDuff et al., 2025), and patient conversations (Tu et al., 2025). Despite these achievements, limited disclosure of data filtering and versioning hinders bias audits and reproducibility, compromising trust in medical applications (Riedemann et al., 2024; Comeau et al., 2025).

In contrast, open-weight models aim to address disclosure gaps by sharing model weights and enabling community-driven improvements. The growing list includes LLaMA-based models, such as Meditron (Chen et al., 2023b) and LLaMA-3-Meditron (Sallinen et al., 2025); Mistral-based mod-

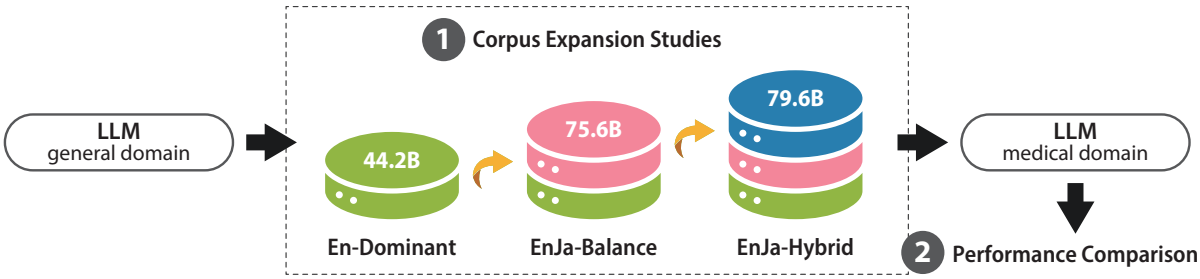

Figure 1: Study overview. This study involves two phases. (1) Corpus expansion began with En-Dominant (44.2B tokens), primarily English medical texts from Web Crawl Data, Clinical Guidelines, and English Paper Abstracts/Full Texts. EnJa-Balance (75.6B tokens) was created by adding machine-translated Japanese data from PubMed/PMC Open Access Subset and J-STAGE papers. EnJa-Hybrid (79.6B tokens) further incorporated publisher-licensed Japanese Medical Textbooks and parallel corpora from PubMed Abstracts. (2) General-domain Japanese LLMs were fine-tuned via continued pre-training on these corpora and evaluated using JMedBench, a comprehensive Japanese medical benchmark suite. Comparisons with other open-weight models validated the effectiveness of our transparent corpus construction for domain adaptation in Japanese medicine.

els, such as BioMistral (Labrak et al., 2024); Qwen-based models, such as Apollo (Wang et al., 2024) and QwQ-Med-3 (Dedhia et al., 2025); the Aloe family (Garcia-Gasulla et al., 2025) derived from either LLaMA or Qwen; and Gemma-based models, such as MedGemma (Sellergren et al., 2025) and TxGemma (Wang et al., 2025b). Nevertheless, these models are generally derived from general-purpose LLMs, such as LLaMA and Qwen, whose pre-training datasets and processes remain partially undisclosed. This limitation prevents them from fully meeting the stringent transparency demands of medical applications. Thus, achieving true transparency in medical LLMs requires not only open weights but also complete disclosure of foundational processes to ensure reliability in healthcare settings.

2.2. Localization in Non-English Medical LLMs

While most foundational LLMs are trained on English-dominant datasets, localization for non-English medical domains is crucial to address linguistic, cultural, regulatory, and clinical nuances, thus reinforcing the need to localize medical LLMs within their own contexts. In particular, the development of Chinese medical LLMs has been vigorous, with significant advancements not only in model construction but also in the curation of specialized datasets and benchmarks (Liu et al., 2024). For example, Zhongjing, a LLaMA-based model, enhances Chinese medical capabilities with a particular focus on traditional Chinese medicine (Yang et al., 2023). The Huatuo series, including HuatuoGPT (Zhang et al., 2023) and HuatuoGPT-o1 (Chen et al., 2024), focuses on complex medical reasoning and specialist disciplines, achieving over 80% accuracy on benchmarks like MedQA. Baichuan-M1, a scratch-trained medical LLM trained on 20T tokens with a range of effective training methods, not only performs strongly across general domains such as mathematics and coding but also excels in specialized medical fields (Wang et al., 2025a). Moreover, general-purpose Chinese LLMs, such as the DeepSeek and Qwen series, also excel in clinical decision support and healthcare applications (Sandmann et al., 2025; Tordjman et al., 2025), surpassing human passing scores on medical licensing exams (Zhu et al., 2025).

However, specializing LLMs in specific fields is not always straightforward for many other language communities with relatively smaller speaker populations compared to English and Chinese. One technical challenge stems from the imbalance of language resources in expert documents, where the dominant portion of circulating medical texts is written in English, surpassing the volume of those written in local languages, such as Japanese. For example, PubMed hosts over 38 million biomedical papers globally[1], while J-STAGE, a comparable Japanese database, contains only around 5 million[2]. Possibly reflecting these difficulties, the number of specialized models for Japanese medicine has been limited, including Preferred-MedLLM-Qwen-72B (Iwasawa, 2025) and LLaMA3-Preferred-MedSwallow-70B[3] built upon Qwen and LLaMA families, respectively. Therefore, a practical recipe for constructing training corpora remains crucial, regarding what types of corpora are es-

[1]Statistics of PubMed: https://pubmed.ncbi.nlm.nih.gov/about/

[2]Statistics of J-STAGE: https://www.jstage.jst.go.jp/browse/-char/en

[3]Hugging Face repository: https://huggingface.co/pfnet/Llama3-Preferred-MedSwallow-70B

sential for local medical domain adaptation and whether machine translation could effectively enhance the efficacy of LLMs working in local contexts. This study addresses this gap by providing an open recipe for medical domain adaptation in Japanese, as an example of a low-resource language. We incrementally built an effective corpus, starting with small-scale medical datasets and progressively adding subcorpora. Through this corpus expansion approach, we clarified the contributions of each subcorpus to model performance, including the effects of machine translation and the importance of various corpus types.

3. Base Model Selection

To ensure the end-to-end transparency of training corpora for Japanese medical LLMs, we adopted a family of open-source Japanese general-domain LLMs as base models. These models, namely the LLM-jp series, were developed by a cross-organizational research initiative focused on advancing Japanese language processing capabilities (LLM-jp et al., 2024). The LLM-jp series is designed to provide robust general-domain language understanding for Japanese, leveraging large-scale, publicly disclosed pre-training datasets to ensure reproducibility and transparency. These general-domain LLMs were further refined through continued pre-training on a corpus comprising 300 billion tokens, which includes a diverse set of general-domain texts supplemented with a small proportion of scientific literature to enhance knowledge coverage. The resulting models include a MoE model with 8×13B parameters, referred to as `8×13B-base`.

Of note, the `8×13B-base` model incorporates eight specialized neural network components, each with 13B parameters, known as "experts" (Shazeer et al., 2017; Nakamura et al., 2025). The component architectures, including the hidden size, number of attention heads, number of layers, and context length, are identical to those of LLaMA 2 (Touvron et al., 2023b). Additionally, a lightweight gating network determines, for each token, which subset of these eight experts is activated to process that specific input, leaving the remaining experts inactive. This selective activation allows the model to perform inference using only a fraction of its total parameters, resulting in improved computational efficiency. Indeed, a total parameter count of 73B can be reduced to an active parameter count of 22B during inference, providing lower computational cost compared to dense models with similar parameter scales. The base model is publicly available on Hugging Face,[4] and the 300B-token corpus is also disclosed on GitLab.[5] Detailed information regarding their construction is provided in the **Appendix**.

4. Medical Corpus Construction

4.1. Continued Pre-training Corpus

To enable domain-adaptive continued pre-training of the base models, we curated a comprehensive medical corpus. Our corpus construction policy was as follows: in general, medical texts can be divided into two groups: (1) widely circulated and publicly available resources, such as medical textbooks, abstracts and full texts of medical papers from scientific repositories like PubMed, clinical practice guidelines, and pharmaceutical package inserts; and (2) sensitive, confidential data containing personal information, such as electronic medical records. To ensure reproducibility in constructing our medical corpus, we exclusively used resources from the first group of publicly accessible materials.

4.1.1. Content Types

These publicly available resources were categorized based on their content types as follows:

- **Web Crawl Data**: Healthcare content collected from websites of pharmaceutical companies, medical societies, and government agencies in both English-speaking and Japanese-speaking regions, with manual curation performed at the domain level for each language region, requiring minimal filtering for broad coverage.
- **Benchmark Training Samples**: Training split of Japanese National Medical Examinations (excluding IgakuQA overlap), MedQA, and other benchmark datasets, sourced from public resources to ensure fair assessment and model robustness. Some English benchmarks were translated into Japanese for use.
- **Clinical Guidelines**: English-language guidelines from the Meditron dataset (open medical guideline data collection) (Chen et al., 2023a), plus Japanese clinical guidelines and pharmaceutical package inserts, including advanced content such as ICD manuals and guidelines for rare diseases, offering high information density for practical use.

[4] Hugging Face repository: `https://huggingface.co/llm-jp/llm-jp-3-8x13b`.

[5] GitLab repository: `https://gitlab.med-jp.nii.ac.jp/datasets/sip3-ja-general-web-corpus`.

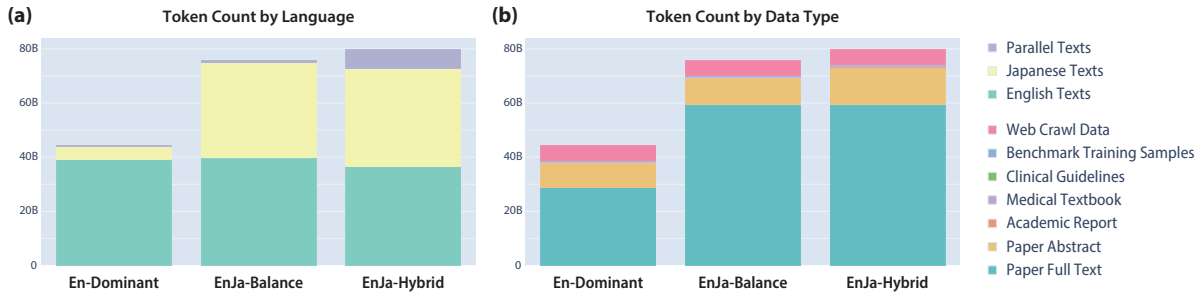

Figure 2: Language and data type composition of En-Dominant, EnJa-Balance, and EnJa-Hybrid corpora. (a) Token distribution by language, showing English, Japanese, and parallel text proportions. En-Dominant is primarily English, while EnJa-Balance and EnJa-Hybrid achieve near-balanced English-Japanese ratios through machine-translated PubMed and PMC Open Access Subset data, with EnJa-Hybrid including parallel corpora. (b) Token breakdown by data type, including Web Crawl Data, Benchmark Training Samples, Clinical Guidelines, Medical Textbooks, Academic Reports, Paper Abstracts, and Paper Full Text. The progression from En-Dominant to EnJa-Hybrid reflects increasing linguistic diversity and content specialization.

# params	Hidden size	# heads	# layers	Context window	Max/Min LR	Warm-up fraction	TP	PP	CP	EP	GB	MB
8×13B	5120	40	40	4096	$1.5 \times 10^{-4}/1.5 \times 10^{-5}$	3%	2	4	1	4	1024	1

Table 1: Model architecture and hyperparameters for continued pre-training. TP represents tensor parallelism, PP pipeline parallelism, CP context parallelism, EP expert parallelism, LR learning rate, GB global batch size, and MB micro batch size.

- **Medical Textbooks**: Web-published medical textbook-level content and publisher-licensed Japanese medical textbooks for medical school students and clinical reference, some of which require copyright permissions and PDF extraction for corpus construction.
- **Academic Reports**: Reports from KAKEN (a Japanese public database that includes information on adopted projects, assessment, and research achievements), providing native-language academic content, the majority of which are in Japanese and machine-translated into English to construct a parallel corpus.
- **Paper Abstracts**: Abstracts from PubMed, J-STAGE (an online platform for Japanese academic journals), and other scientific repositories, some of which include machine-translated texts via NICT's (National Institute of Information and Communications Technology) science translation engine to address Japanese resource scarcity (see the **Appendix** for the translation quality of this science translation engine).
- **Paper Full Text**: Full articles from PMC Open Access Subset, J-STAGE, and S2ORC (English academic paper repository by Allen Institute for AI). Similarly, some texts were translated into Japanese to augment the volume of Japanese text data.

To ensure reproducibility and transparency in corpus construction, a detailed list of data sources for each subcorpus, along with specifics on data processing and filtering, is provided in the **Appendix**.

4.1.2. Expansive Corpora Construction

These subcorpora were combined to construct En-Dominant, EnJa-Balance, and EnJa-Hybrid, with each corpus incrementally expanding on the previous one by adding or substituting distinct subcorpora, thereby increasing complexity and specialization. The differential relationships are as follows:

- **En-Dominant**: A corpus consisting of Web Crawl Data, original English Paper Abstracts, Paper Full Texts (mainly from PMC Open Access Subset and S2ORC), clinical guidelines, and academic reports, along with a dataset from Benchmark Training Samples. This was constructed as a predominantly English corpus.
- **EnJa-Balance**: A dataset that adds machine-translated Japanese data from PubMed Abstracts and PMC Open Access Subset to En-Dominant, achieving a nearly balanced ratio between English and Japanese. Paper abstracts and full text from J-STAGE were also included to enhance the medical knowledge originally written in Japanese.
- **EnJa-Hybrid**: A corpus that converts the machine-translated Japanese data from PubMed Abstracts (added in EnJa-Balance)

into a parallel corpus with the original English abstracts, and further incorporates publisher-licensed medical texts. This maintains a balanced English-Japanese ratio while adding parallel corpora.

See **Figure 2** for the language composition and detailed data types in each corpus. These differential additions culminate in EnJa-Hybrid, a robust English-Japanese medical corpus totaling 79.6 billion tokens.

4.2. Supervised Fine-tuning Corpus

We used the first version of the general-domain instruction tuning dataset published by LLM-jp[6], along with the original training datasets from MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022), as well as Japanese translations of the MedQA and PubMedQA training datasets. Additionally, we incorporated past questions from the Japanese National Medical Examination spanning 12 years, excluding any portions overlapping with IgakuQA (Kasai et al., 2023).

5. Model Training

To adapt the general-domain LLM, `8×13B-base`, for the Japanese medical domain, we implemented a two-stage training process involving continued pre-training followed by supervised fine-tuning. Note that we utilized a tokenizer from the LLM-jp series throughout the process (LLM-jp et al., 2024).

5.1. Continued Pre-training

The training was conducted on an Amazon Web Services (AWS) SageMaker cluster equipped with 32 nodes, each containing 8 NVIDIA H100 GPUs, totaling 256 GPUs. We employed Megatron-LM v0.3.0[7] for efficient parallel training. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1.0 \times 10^{-8}$, incorporating a weight decay of 0.1, gradient clipping of 1.0, and a cosine learning rate schedule, without applying dropout. To enhance memory efficiency and accelerate attention computation, FlashAttention (Dao, 2024; Dao et al., 2022) was integrated into the training process. Detailed model architecture and training hyperparameters are provided in **Table 1**.

[6] https://huggingface.co/llm-jp/llm-jp-13b-v1.0

[7] https://github.com/llm-jp/Megatron-LM/tree/v4

Task	Source	Dataset
MCQA	Original	IgakuQA (Kasai et al., 2023) JMMLU-Medical*
	Translation	MedMCQA (Pal et al., 2022) MedQA (Jin et al., 2020) USMLE-QA PubMedQA (Jin et al., 2019) MMLU-medical (Hendrycks et al., 2021b,a)
MT	Original	EJMMT (Hayakawa and Arase, 2020a)
NER	Original	MRNER-disease† MRNER-medicine† NRNER†
	Translation	BC2GM (Smith et al., 2008) BC5Chem (Pavlova and Makhlouf, 2023) BC5-Disease (Li et al., 2016) JNLPBA (Collier et al., 2004) NCBI Disease (Doğan et al., 2014)
DC	Original	CRADE† RRTNM† SMDIS†
STS	Original	JCSTS (Mutinda et al., 2021)

Table 2: Detailed information about JMedBench. Among these datasets, MRNER-disease†, MRNER-medicine†, NRNER†, CRADE†, RRTNM†, and SMDIS† originate from JMED-LLM (available at `https://github.com/sociocom/JMED-LLM`), while JMMLU-Medical* is available at `https://github.com/nlp-waseda/JMMLU`.

5.2. Supervised Fine-tuning

The supervised fine-tuning was conducted on an AWS SageMaker cluster consisting of 8 nodes, each equipped with 8 NVIDIA H100 GPUs, totaling 64 GPUs. We utilized the NeMo framework (Kuchaiev et al., 2019) for supervised fine-tuning, leveraging the AdamW optimizer (Loshchilov and Hutter, 2017) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1.0 \times 10^{-8}$. The training configuration included a learning rate of 2.0×10^{-6} with 20 warm-up steps, a weight decay of 0.1, a global batch size of 64, and a cosine learning rate schedule. No dropout was applied during fine-tuning. FlashAttention (Dao, 2024; Dao et al., 2022) was enabled based on configuration settings.

6. Evaluation Framework

6.1. JMedBench

The model's performance in the Japanese medical domain was comprehensively evaluated using the JMedBench benchmark (Jiang et al., 2025). JMedBench consists of 20 Japanese and 7 English tasks, encompassing multiple-choice question answering (MCQA), machine translation (MT), named entity recognition (NER), document classification (DC), and semantic textual similarity (STS). Notably, JMedBench includes both original Japanese medical datasets and translated portions of English medical datasets to enhance its comprehensiveness (see **Table C.1** for detailed information).

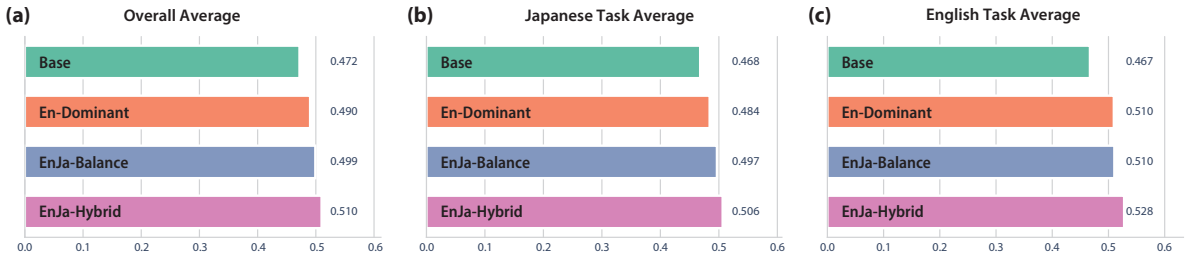

Figure 3: Performance of base and continued pre-trained models on JMedBench. This figure compares the (a) JMedBench Overall Average, (b) Japanese Task Average, and (c) English Task Average scores across the base model and models trained on En-Dominant, EnJa-Balance, and EnJa-Hybrid datasets. The base model serves as the reference, while En-Dominant, EnJa-Balance, and EnJa-Hybrid represent models enhanced through continued pre-training on their respective corpora, showing incremental improvements in performance.

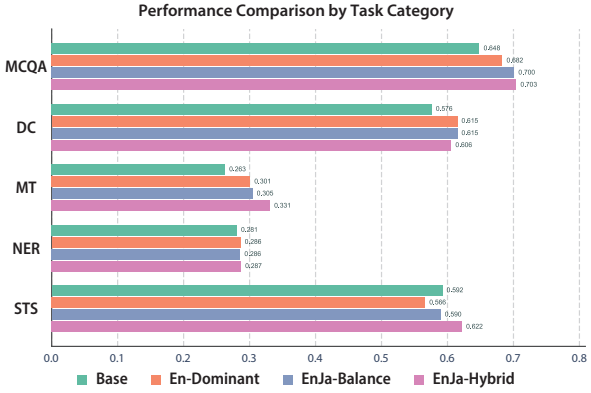

Figure 4: Task category performance across models. Scores reflect improvements across multiple-choice question answering (MCQA), machine translation (MT), named entity recognition (NER), document classification (DC), and semantic textual similarity (STS), evaluated using JMedBench. EnJa-Hybrid consistently shows the highest scores across all task categories except DC, with notably higher MT performance reflecting the enhanced parallel corpus composition.

To assess the accuracy of model outputs, JMedBench employs different calculation methods tailored to each task type. For MCQA and DC tasks, the model is required to select a single correct answer from multiple options that best matches the given question. The accuracy of these tasks is calculated by computing the likelihood of each option, with the option exhibiting the highest likelihood designated as the model's response. For other task categories, different metrics are utilized: MT performance is evaluated using the BLEU score, NER is assessed with the entity F1 score, and STS is measured by the Pearson correlation coefficient.

6.1.1. IgakuQA

IgakuQA comprises past questions and their corresponding answers from the Japanese National Medical Licensing Examination, covering the years 2018 to 2022 (Kasai et al., 2023). It encompasses a wide range of topics, including medical knowledge essential for clinical practice in Japan, legal knowledge related to medical laws, and questions emphasizing ethical considerations. As such, IgakuQA serves as the most direct and comprehensive benchmark for evaluating the knowledge of LLMs in the Japanese medical domain. Defined as a task within the MCQA category of JMedBench, IgakuQA is particularly emphasized in performance evaluations due to its critical role in assessing these competencies.

7. Experiments and Results

We aim to provide an open recipe for medical domain adaptation in Japanese, validated through corpus expansion studies (see **Figure 1** for an overview of the study). These studies demonstrate the effectiveness of machine-translated texts and reveal varying impacts depending on content sources. Subsequently, we compare the resulting model with representative 70B-parameter open-weight models to show that our recipe offers an efficient training strategy for constructing medical LLMs tailored to local contexts.

7.1. Corpus Expansion Studies

Continued pre-training, conducted after initial pre-training, utilizes additional corpora with domain-specific or target-language texts. Its effectiveness has been shown in various studies (Gupta et al., 2023; Cui et al., 2024; Pires et al., 2023; Zhu et al., 2023; Zhao et al., 2024; Fujii et al., 2024), yet detailed examinations of how incrementally adding

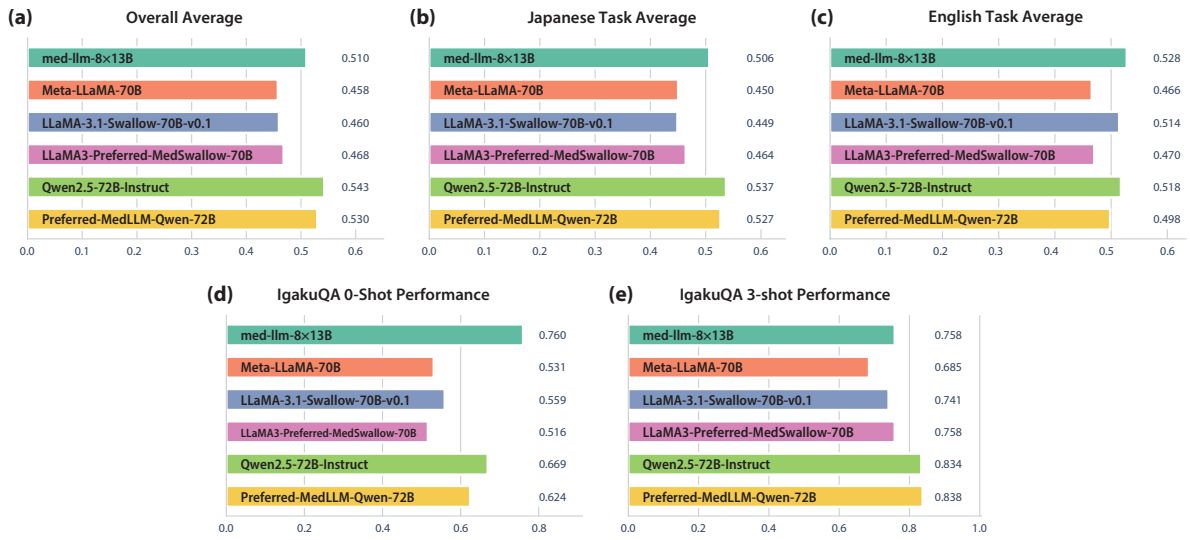

Figure 5: Performance comparison of `SIP-med-llm-8×13B` with other 70B open-weight models. Our comparison includes Meta-LLaMA-3-70B, LLaMA-3.1-Swallow-70B-v0.1, LLaMA-Preferred-MedSwallow-70B, Qwen2.5-72B-Instruct, and Preferred-MedLLM-Qwen-72B. The chart displays (a) overall average, (b) Japanese task average, (c) English task average, (d) IgakuQA 0-shot, and (e) IgakuQA 3-shot performance metrics. Our model consistently outperformed LLaMA-based models and achieved the highest performance in IgakuQA 0-shot, but it lagged behind Qwen-based models in 3-shot performance.

subcorpora of diverse document types affects specific task performance remain limited.

Accordingly, we performed continued pre-training of the `8×13B-base` model on En-Dominant, EnJa-Balance, and EnJa-Hybrid corpora, respectively, followed by supervised fine-tuning with an instruction-tuning dataset of identical composition. The results, detailed in **Figure 3**, compare the JMedBench Overall Average, Japanese Task Average, and English Task Average scores, showing incremental performance improvements as corpora expand from En-Dominant to EnJa-Hybrid. By observing the differences between En-Dominant and EnJa-Balance in **Figures 3b** and **3c**, we find that the addition of Japanese paper texts, including machine-translated content, primarily contributes to Japanese task performance.

Figure 4 further highlights task-specific performance across MCQA, MT, NER, DC, and STS, with EnJa-Hybrid achieving the highest scores except in DC. Observation of EnJa-Hybrid in **Figure 4** shows that the addition of parallel corpora contributed to improvements in MT. These results demonstrate that the performance gains achieved through continued pre-training effectively reflect the characteristics of each corpus.

7.2. Performance Comparison

We evaluated the performance of `SIP-med-llm-8×13B` against other representative 70B open-weight models. For LLaMA-based models, we included Meta-LLaMA-3-70B (base model), LLaMA-3.1-Swallow-70B-v0.1 (Japanese general-domain adapted model), and LLaMA-Preferred-MedSwallow-70B (Japanese medical-domain adapted model). For Qwen-based models, we included Qwen2.5-72B-Instruct (base model) and Preferred-MedLLM-Qwen-72B (Japanese medical-domain adapted model).

As shown in **Figure 5**, our model `SIP-med-llm-8×13B` outperformed all 70B LLaMA-based models across all metrics evaluated here, including overall JMedBench performance, Japanese average performance, English average performance, and IgakuQA. However, compared to Qwen-based models, it slightly underperformed in overall JMedBench performance and Japanese average performance. In terms of IgakuQA zero-shot performance, our model surpassed all other models, but in three-shot performance, Qwen-based models performed better. Notably, our model operates with only 22B active parameters during inference, demonstrating competitive performance against 70B models, which highlights its efficiency from a computational cost perspective.

8. Conclusions

This study presents a fully transparent, open recipe for domain adaptation of LLMs tailored to the Japanese medical domain, culminating in the development of `SIP-med-llm-8×13B`, a mixture-of-experts model with 22B active param-

eters. Through incremental corpus expansion, from En-Dominant to EnJa-Hybrid (79.6B tokens), we demonstrated the efficacy of incorporating machine-translated texts and local medical resources, achieving competitive performance on JMedBench and IgakuQA benchmarks compared to 70B open-weight models.

Future work includes advancing bias evaluation research through transparent corpus construction, developing more sophisticated reasoning models, and evaluating clinical utility beyond benchmark performance. Additionally, we aim to improve our model's ability to leverage in-context examples, particularly targeting enhanced 3-shot performance on IgakuQA tasks. Our end-to-end disclosed corpus and training pipeline provide a reproducible framework for adapting LLMs to localized medical contexts, addressing the opacity of existing medical LLMs.

9. Acknowledgements

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

10. Bibliographical References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms.

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. 2023a. Meditron-70b: Scaling medical pretraining for large language models.

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. 2023b. Meditron-70b: Scaling medical pretraining for large language models.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners.

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

Donnella S. Comeau, Danielle S. Bitterman, and Leo Anthony Celi. 2025. Preventing unrestricted and unmonitored ai experimentation in healthcare through transparency and accountability. *npj Digital Medicine 2025 8:1*, 8:1–7.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bhishma Dedhia, Yuval Kansal, and Niraj K. Jha. 2025. Bottom-up domain-specific superintelligence: A reliable knowledge graph is what we need.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan,

Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

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

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *First Conference on Language Modeling (COLM)*.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, Sergio Alvarez-Napagao, Eduard Ayguadé-Parra, and Ulises Cortés. 2025. The aloe family recipe for open and specialized healthcare llms.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual Pre-Training of Large Language Models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

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

Takeshi Hayakawa and Yuki Arase. 2020b. Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164.

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

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

Binyuan Hui et al. 2024. Qwen2.5 technical report.

Junichiro Iwasawa. 2025. Stabilizing reasoning in medical llms with continued pretraining and reasoning preference optimization.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. JMedBench: A Benchmark for Evaluating Japanese Biomedical Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hsin-Hsien Yeh, and Pranav Rajpurkar. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jungo Kasai et al. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

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

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions.

LLM-jp, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv preprint arXiv:2407.03963*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457.

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

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. 2025. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. *arXiv preprint arXiv:2502.19261*.

OpenAI. 2023. Gpt-4 technical report. ArXiv: 2303.08774v4.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek

Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to Med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174, pages 248–260.

Vera Pavlova and Mohammed Makhlouf. 2023. BIOptimus: Pre-training an optimal biomedical language model with curriculum learning for named entity recognition. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 337–349, Toronto, Canada. Association for Computational Linguistics.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese Large Language Models. In *Intelligent Systems*, pages 226–240.

Lars Riedemann, Maxime Labonne, and Stephen Gilbert. 2024. The path forward for large language models in medicine is open. *npj Digital Medicine*, 7:1–5.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-

Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, Tahseen Rabbani, Trevor Brokowski, Meditron Medical Doctor Working Group, Tim G. J. Rudner, and Mary-Anne Hartley. 2025. Llama-3-meditron: An open-weight suite of medical LLMs based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. 2025. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–4.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. Medgemma technical report.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

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

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models.

Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, Amine Geahchan, Anis Meribout, Nader Yatim, Nicole Ng, Phillip Robson, Alexander Zhou, Sara Lewis, Mingqian Huang, Timothy Deyer, Bachir Taouli, Hao Chih Lee, Zahi A. Fayad, and Xueyan Mei. 2025. Comparative benchmarking of the deepseek large

language model on medical tasks and clinical reasoning. *Nature Medicine*, pages 1–6.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, Da Pan, Fei Kou, Fei Li, Fuzhong Chen, Guosheng Dong, Han Liu, Hongda Zhang, Jin He, Jinjie Yang, Kangxi Wu, Kegeng Wu, Lei Su, Linlin Niu, Linzhuang Sun, Mang Wang, Pengcheng Fan, Qianli Shen, Rihui Xin, Shunya Dang, Songchi Zhou, Weipeng Chen, Wenjing Luo, Xin Chen, Xin Men, Xionghai Lin, Xuezhen Dong, Yan Zhang, Yifei Duan, Yuyan Zhou, Zhi Ma, and Zhiying Wu. 2025a. Baichuan-m1: Pushing the medical capability of large language models.

Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025b. Txgemma: Efficient and agentic llms for therapeutics.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language model to be a doctor.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. LLaMA Beyond English: An Empirical Study on Language Capability Transfer. *arXiv preprint arXiv:2401.01055*.

Shiben Zhu, Wanqin Hu, Zhi Yang, Jiani Yan, and Fang Zhang. 2025. Qwen-2.5 outperforms other large language models in the chinese national nursing licensing examination: Retrospective cross-sectional comparative study. *JMIR medical informatics*, 13:e63731.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating Large Language Models to Non-English by Aligning Languages. *arXiv preprint arXiv:2308.04948*.

A. Base Model Details

A.1. Architecture Specifications

The `8×13B-base` model is constructed through continued pre-training of `llm-jp/llm-jp-3-8x13b` (LLM-jp et al., 2024)[8] using a 300-

[8] https://huggingface.co/llm-jp/llm-jp-3-8x13b

billion (300B)-token corpus. It is a Mixture-of-Experts (MoE) architecture with 8×13B parameters (Shazeer et al., 2017; Nakamura et al., 2025), incorporating eight specialized neural network components ("experts"), each with 13B parameters. The component architectures, including hidden size (5120), number of attention heads (40), number of layers (40), and context length (4096), are identical to those of LLaMA 2 (Touvron et al., 2023b). A lightweight gating network determines, for each token, which subset of the eight experts is activated, leaving the others inactive. This results in a total parameter count of approximately 73B, reduced to an active parameter count of 22B during inference for improved computational efficiency.

A.2. Pre-training Corpus

The 300B-token corpus comprises a diverse set of general-domain subcorpora, with a nearly balanced language composition between Japanese and English, augmented by a small proportion of scientific literature to enhance knowledge coverage (see **Table A.1**).

- **Crawled HTML and PDF**: We crawled the web to collect recent Japanese texts from August to November 2024, amassing 440 million HTML documents and 13 million PDF documents. The HTML documents were filtered using Uzushio,[9] a corpus preprocessing tool designed for billion-token-scale web corpora. The PDF documents were processed using Surya,[10] to the extent permitted by computational resources, with the remaining documents processed using `pdftotext`.[11]
- **NDL WARP HTML**: We collected HTML documents from URLs registered in the Web Archiving Project (WARP)[12] of the National Diet Library (NDL) in Japan. WARP is Japan's national web archiving initiative, preserving web-based information of cultural and historical significance for future accessibility.
- **NINJAL Web Japanese Corpus (NWJC)**: Provided courtesy of the National Institute for Japanese Language and Linguistics (NINJAL), this subcorpus consists of HTML documents crawled from the fourth quarter of 2012 to the second quarter of 2015. We used documents from the second quarter of 2013 to the second quarter of 2015.

Subset	Language	Est. Tokens [B]
Crawled HTML	Japanese	34.57
Crawled PDF (processed by Surya)	Japanese	0.17
Crawled PDF (processed by `pdftotext`)	Japanese	57.63
NDL WARP HTML	Japanese	4.76
NINJAL Web Japanese Corpus (NWJC)	Japanese	58.89
J-GLOBAL	Japanese	2.60
J-GLOBAL	English	0.01
Dolma v1.7	English	150.22
Total		**309.15**

Table A.1: Composition of the 300B-token general-domain pre-training corpus. This table details subcorpora by source and language, with estimated token counts in billions (B). The corpus is nearly balanced between Japanese and English, primarily comprising general-content sources and supplemented by a small proportion of scientific literature from J-GLOBAL.

- **J-GLOBAL**: Japanese and English abstract texts from J-GLOBAL,[13] a comprehensive scientific and technical information database based in Japan, were provided courtesy of the Japan Science and Technology Agency (JST). This resource is widely utilized by Japanese researchers, engineers, and industry professionals for literature searches and access to scientific documents.
- **Dolma v1.7**: A significant portion of our English corpus was sourced from Dolma v1.7 (Soldaini et al., 2024), a large English dataset curated by the Allen Institute for AI (AI2). Specifically, we utilized the middle portion of Dolma's Common Crawl (CC) subset.

A.3. Training Configuration

The training was conducted on an Amazon Web Services (AWS) SageMaker cluster equipped with 32 nodes, each containing 8 NVIDIA H100 GPUs, totaling 256 GPUs. We employed Megatron-LM v0.3.0[14] for efficient parallel training. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1.0 \times 10^{-8}$, incorporating a weight decay of 0.1, gradient clipping of 1.0, and a cosine learning rate schedule, without applying dropout. To enhance memory efficiency and accelerate attention computation, FlashAttention (Dao, 2024; Dao et al., 2022) was integrated into the training process. Other hyperparameters were set as follows: the maximum learning rate was 1.5×10^{-4}, and the minimum learning rate was 1.5×10^{-5}. The warm-up fraction was 3%. For parallelism, tensor parallelism was set to 2, pipeline parallelism to 4, context parallelism

[9] https://github.com/WorksApplications/uzushio
[10] https://github.com/datalab-to/surya
[11] https://www.xpdfreader.com/pdftotext-man.html
[12] https://warp.ndl.go.jp/
[13] https://jglobal.jst.go.jp/en
[14] https://github.com/llm-jp/Megatron-LM/tree/v4

to 4, and expert parallelism to 1. The global batch size was 1024, and the micro batch size was 1.

A.4. Public Access

The `8×13B-base` model is publicly available on Hugging Face,[15] and the 300B-token corpus is also disclosed on GitLab.[16]

B. Medical Corpus Construction

B.1. Corpus Composition

The `SIP-med-llm-8×13B` model was developed through continued pre-training of the `8×13B-base` model using the EnJa-Hybrid medical corpus, tailored for domain adaptation to the Japanese medical field. The EnJa-Hybrid corpus, validated through corpus expansion studies outlined in the main text, comprises a nearly balanced bilingual dataset of approximately 79.6 billion tokens in English and Japanese. The data sources, categorized by content type, along with their respective processing details, token counts, and descriptions, are summarized in **Table B.1**.

Note that the subcorpora "Other Copyright Abstracts," "Japanese Medical Textbooks," J-STAGE-related subcorpora, and J-GLOBAL-related subcorpora, as well as certain subcorpora machine-translated using the National Institute of Information and Communications Technology (NICT) Science Translator, are utilized based on permissions from the respective copyright holders. Redistribution of these subcorpora is not permitted, and the usage scope of models trained on them is subject to restrictions as per the individual agreements.

B.2. Translation Performance of the Machine-Translation Models

The machine translation from English to Japanese, particularly for the "PMC OA Japanese" subcorpus and the "PubMed En-Ja Clinical Abstracts," was performed using the NICT Science Translator, courtesy of NICT. Given the relatively large token size of these corpora (see **Table B.1**), the quality of this translator is particularly important. We evaluated its English-to-Japanese translation performance on the EJMMT dataset, comparing it against the baseline reported by Hayakawa and Arase (2020b) and the `gpt-4o-2024-08-06` model. The results confirmed that the NICT translator achieves competitive performance compared to `gpt-4o-2024-08-06`. The performance metrics are presented in **Table B.2**. BLEU measures agreement with the ground truth using the SacreBLEU library[17] with the MeCab tokenizer,[18] while COMET-22[19] and COMET-23[20] serve as neural evaluation frameworks for machine translation.

B.3. Public Access

Among the models developed in this study, those that are free from restrictions based on corpus licenses and can be utilized without strict constraints are made publicly available on the following Hugging Face repository: [masked for anonymous submission].

C. JMedBench Benchmark Details

JMedBench comprises 20 Japanese and 7 English tasks, encompassing multiple-choice question answering (MCQA), machine translation (MT), named entity recognition (NER), document classification (DC), and semantic textual similarity (STS) (Jiang et al., 2025). Detailed information for each benchmark dataset, organized by task category, is provided below.

C.1. Multi-Choice Question Answering (MCQA)

MedMCQA/MedMCQA-Jp MedMCQA is a large-scale, MCQA dataset designed to address real-world medical entrance exam questions, covering 2.4 thousand health topics and 21 medical subjects sampled from medical entrance exams across India (Pal et al., 2022). This contains 4,183 test samples. MedMCQA-Jp is a Japanese translation of MedMCQA.

USMLEQA/USMLEQA-Jp USMLEQA is a large-scale, MCQA dataset with 1,273 test samples with 4 options, which are sampled from United States Medical Licensing Examinations (Jin et al., 2020). USMLEQA-Jp is a Japanese translation of USMLEQA, containing the same number of test samples.

MedQA/MedQA-Jp MedQA is a 5-option version of USMLEQA, known as a representative benchmark for medical large language models in the assessment of medical knowledge sufficient for

[15] Hugging Face repository: Hugging Face repository: https://huggingface.co/llm-jp/llm-jp-3-8x13b.

[16] GitLab repository: https://gitlab.med-jp.nii.ac.jp/datasets/sip3-ja-general-web-corpus.

[17] https://github.com/mjpost/sacrebleu

[18] https://pypi.org/project/mecab-python3/

[19] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[20] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

Content Type	Subcorpus Name	Tokens (M)	Description	Language
Paper Full Text	PMC OA Subset	28 755	Full-text English articles from the PubMed Central Open Access (OA) Subset (https://pmc.ncbi.nlm.nih.gov/tools/openftlist/), retrieved in August 2024, filtered for CC0 and CC-BY licenses.	English
Paper Full Text	PMC OA Japanese	27 757	Japanese translations of the PubMed Central OA Subset, generated using the National Institute of Information and Communications Technology (NICT) Science Translator, excluding failed translations.	Japanese
Paper Full Text	S2ORC bioRxiv	269	Full-text articles from the S2ORC bioRxiv collection (https://www.biorxiv.org/), filtered for CC0 and CC-BY licenses.	English
Paper Full Text	J-STAGE Full Text	2568	Japanese full-text articles from J-STAGE (https://www.jstage.jst.go.jp/browse/-char/en), collected via web crawling and extracted using Surya, retrieved in August 2024, with permission from the Japan Science and Technology Agency (JST) for machine learning model development.	Japanese
Paper Abstract	PubMed En-Ja Clinical Abstracts	6271	Parallel English-Japanese clinical abstracts from PubMed (https://pubmed.ncbi.nlm.nih.gov/download/), retrieved in August 2024, translated using the NICT Science Translator, based on Meditron's (https://github.com/epfLLM/meditron) approved journal list, with randomized language order.	English/Japanese
Paper Abstract	PubMed English Non-clinical Abstracts	3230	English abstracts from non-clinical medical journals in PubMed (https://pubmed.ncbi.nlm.nih.gov/download/), retrieved in August 2024.	English
Paper Abstract	J-STAGE English Abstracts	329	English abstracts from J-STAGE (https://www.jstage.jst.go.jp/browse/-char/en) in the medical domain, collected via web crawling, retrieved in August 2024, excluding duplicates from the "J-STAGE En-Ja Abstracts" parallel corpus, with permission from the JST.	English
Paper Abstract	J-STAGE Japanese Abstracts	116	Japanese abstracts from J-STAGE (https://www.jstage.jst.go.jp/browse/-char/en) in the medical domain, collected via web crawling, retrieved in August 2024, excluding duplicates from the parallel corpus, with permission from the JST.	Japanese
Paper Abstract	J-STAGE En-Ja Abstracts	333	Parallel English-Japanese abstracts from J-STAGE (https://www.jstage.jst.go.jp/browse/-char/en) in the medical domain, collected via web crawling, retrieved in August 2024, with permission from the JST.	English/Japanese
Paper Abstract	J-GLOBAL English Abstracts	14	English abstracts from J-GLOBAL (https://jglobal.jst.go.jp/en), deduplicated against PubMed and J-STAGE using DOI and other identifiers, provided by the JST.	English
Paper Abstract	J-GLOBAL Japanese Abstracts	2409	Japanese abstracts from J-GLOBAL (https://jglobal.jst.go.jp/en), deduplicated against PubMed and J-STAGE using DOI and other identifiers, provided by the JST.	Japanese
Paper Abstract	Other Copyright Abstracts	1026	Japanese medical abstracts, including those from Ichushi Web, collected via web crawling or with publisher permissions, compliant with Japanese copyright law.	Japanese
Academic Report	KAKEN En-Ja Reports	337	Parallel English-Japanese KAKEN reports, deduplicated by ID, sourced from the Hugging Face dataset hpprc/kaken-trans-ja-en (https://huggingface.co/datasets/hpprc/kaken-trans-ja-en).	English/Japanese
Medical Textbook	Japanese Medical Textbooks	100	Japanese medical textbook-quality texts from publishers such as Igaku-Shoin or web-crawled sources, collected via web crawling or with publisher permissions, compliant with Japanese copyright law.	Japanese
Clinical Guidelines	Meditron English Clinical Guidelines	141	English clinical guidelines from the Meditron (https://github.com/epfLLM/meditron) dataset, collected using its scraping tools.	English
Clinical Guidelines	Japanese Clinical Guidelines	173	Japanese clinical guidelines (e.g., https://www.jmsf.or.jp/en), PMDA pharmaceutical inserts (https://www.info.pmda.go.jp/psearch/html/menu_tenpu_base.html), rare disease information from the Ministry of Health, Labour and Welfare (https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000084783.html), and ICD-related data (https://www.who.int/standards/classifications/classification-of-diseases), collected via web crawling in compliance with robots.txt.	Japanese
Benchmark Training Dataset	English Benchmark Training Data	93	Training data from MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022), converted to question-option-answer text format.	English
Benchmark Training Dataset	Japanese Benchmark Training Data	14	Japanese training data from Japan's National Medical Examinations (2006–2017, excluding IgakuQA overlap), with English translations of MedQA and MedMCQA samples, converted to question-option-answer text format.	Japanese
Web Crawl Data	English Medical Web Crawl	3589	Web-crawled medical domain data, primarily in English.	English
Web Crawl Data	Japanese Medical Web Crawl	2096	Web-crawled medical domain data, primarily in Japanese.	Japanese

Table B.1: Composition of the EnJa-Hybrid medical corpus. Benchmark Training samples, including MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and Japanese translations of the MedQA and PubMedQA training datasets, were utilized for Instruction pre-training (Cheng et al., 2024). All URLs verified as accessible on August 1, 2025.

Model	BLEU	COMET-22	COMET-23
NICT Science Translator	37.71	80.78	65.64
EJMMT Baseline	26.77	77.86	64.93
gpt-4o-2024-08-06	27.23	79.86	68.16

Table B.2: Translation performance metrics for English-to-Japanese translation on the EJMMT dataset. Our model, the NICT Science Translator, outperformed both the baseline and gpt-4o-2024-08-06 in BLEU and COMET-22 metrics, though it scored slightly lower than gpt-4o-2024-08-06 in COMET-23.

Task	Source	Dataset
MCQA	Original	IgakuQA (Kasai et al., 2023) JMMLU-Medical*
	Translation	MedMCQA (Pal et al., 2022) MedQA (Jin et al., 2020) USMLE-QA (Jin et al., 2020) PubMedQA (Jin et al., 2019) MMLU-medical (Hendrycks et al., 2021b,a)
MT	Original	EJMMT (Hayakawa and Arase, 2020b)
NER	Original	MRNER-disease† MRNER-medicine† NRNER†
	Translation	BC2GM (Smith et al., 2008) BC5Chem (Pavlova and Makhlouf, 2023) BC5-Disease (Li et al., 2016) JNLPBA (Collier et al., 2004) NCBI Disease (Doğan et al., 2014)
DC	Original	CRADE† RRTNM† SMDIS†
STS	Original	JCSTS (Mutinda et al., 2021)

Table C.1: Detailed information about JMed-Bench. Among these datasets, MRNER-disease†, MRNER-medicine†, NRNER†, CRADE†, RRTNM†, and SMDIS† originate from JMED-LLM (available at https://github.com/sociocom/JMED-LLM), while JMMLU-Medical* is available at https://github.com/nlp-waseda/JMMLU.

medical licensure (Jin et al., 2020). MedQA-Jp is a Japanese translation of MedQA, containing the same number of test samples.

MMLU-Medical/MMLU-Medical-Jp MMLU-Medical contains 1,871 biomedical questions at the college level as test samples, which is extracted as a subset of a large-scale, multi-topics benchmark, MMLU (Hendrycks et al., 2021b,a). MMLU-Medical-Jp is a Japanese translation of MMLU-Medical.

JMMLU-Medical While the MMLU-Medical-Jp is a machine-translated version of MMLU-Medical, JMMLU-Medical consists of human-translated Japanese version of MMLU-Medical comprising 1,271 test samples[21].

IgakuQA/IgakuQA-En IgakuQA contains 989 Japanese questions based on Japanese medical licensing examinations from 2018 to 2022 (Kasai et al., 2023). This uniquely reflects Japanese-specific medical practices, healthcare systems, and epidemiological profiles. IgakuQA-En is an English translation of IgakuQA.

PubMedQA/PubMedQA-Jp PubMedQA contains 1,000 test samples focusing on the biomedical field collected from PubMed Abstracts (Jin et al., 2019). The task of PubMedQA is to answer research questions with yes/no/maybe. PubMedQA-JP is a Japanese translation of PubMedQA.

C.2. Machine Translation (MT)

EJMMT-Ja/EJMMT-En EJMMT is a Japanese–English medical machine-translation dataset with fine-grained annotation of error spans and error types (Hayakawa and Arase, 2020b). EJMMT-Ja indicates the translation accuracy in the direction of English to Japanese, while EJMMT-En indicates the Japanese to English direction. These include 2,400 test samples.

C.3. Named Entity Recognition (NER)

MRNER-Medicine MRNER-Medicine (Medical Report Named Entity Recognition for medicine) contains 90 test samples for extracting medication-related information from case reports in Japanese[22].

MRNER-Disease MRNER-Disease (Medical Report Named Entity Recognition for positive disease) contains 90 test samples for extracting symptoms actually observed in patients from case reports and radiology reports in Japanese[22].

NRNER NRNER (Nursing Record Named Entity Recognition) contains 90 test samples, involving extracting information about symptoms actually observed in patients and medication from simulated nursing records in Japanese[22].

BC2GM-Jp BC2GM-Jp is a Japanese translation of BC2GM (BioCreative II Gene Mention Recognition) (Smith et al., 2008), which contains 5,037 test samples to identify a gene mention in a sentence.

BC5Chem-Jp BC5Chem-Jp is a Japanese translation of BC5Chem (Pavlova and Makhlouf, 2023), which contains 4,801 test samples to identify disease, chemical entities and their relations from biomedical texts.

BC5Disease-Jp BC5Disease-Jp is a Japanese translation of BC5Disease (Li et al., 2016), which contains 4,797 test samples to identify disease,

[21] https://huggingface.co/datasets/nlp-waseda/JMMLU

[22] This benchmark is originally included in JMED-LLM (Japanese Medical Evaluation Dataset for Large Language Models): https://github.com/sociocom/jmed-llm

chemical entities and their relations from biomedical texts.

JNLPBA-Jp JNLPBA-Jp is a Japanese translation of JNLPBA (Collier et al., 2004), which features 4,260 test samples for bio-entity recognition, identifying and classifying technical terms in the domain of molecular biology.

NCBI-Disease-Jp NCBI-Disease-Jp is a Japanese translation of NCBI-Disease (Doğan et al., 2014), which contains 940 test samples to identify the disease name on the NCBI disease corpus.

C.4. Document Classification (DC)

CRADE CRADE (Case Report Adverse Drug Event) contains 92 test samples, which involves classifying the possibility of adverse events from medications and symptoms in case reports in Japanese[22].

RRTNM RRTNM (Radiology Report Tumor Nodes Metastasis) contains 89 test samples, which involves predicting TNM classification of cancer from radiology reports of lung cancer patients in Japanese[22].

SMDIS SMDIS (Social Media Disease) comprises 84 test samples, which involve classifying the presence or absence of diseases or symptoms of the poster or people around them from simulated Tweets in Japanese[22].

C.5. Semantic Text Similarity (STS)

JCSTS JCSTS (Japanese Clinical Semantic Textual Similarity) has 3,500 test samples in Japanese. This is a medical version of the semantic textual similarity task that determines the semantic similarity between two sentences, dealing with case reports (Mutinda et al., 2021).