

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

Akiko Aizawa¹, Yuki Arase², Fei Cheng², Jiahao Huang³, Zhiyi Huang^{4,†}, Junfeng Jiang³, Teruhito Kanazawa¹, Daisuke Kawahara⁵, Kazuma Kobayashi^{1,†}, Takashi Kodama², Sadao Kurohashi², Yusuke Oda¹, Yuma Tsuta¹, Zhen Wan², Zhishen Yang¹, Rio Yokota⁶

¹National Institute of Informatics

²Kyoto University

³University of Tokyo

⁴Institute of Science Tokyo

⁵Waseda University

⁶Institute 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.ssrc.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

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)¹ using a 300-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., 2023). 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,² a corpus preprocessing tool designed for billion-token-scale web corpora. The PDF documents were processed using Surya,³ to the extent permitted by computational resources, with the remaining documents processed using `pdftotext`.⁴
- **NDL WARP HTML:** We collected HTML documents from URLs registered in the Web Archiving Project (WARP)⁵ 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

Equal contributions for all authors, listed in alphabetical order.

† Corresponding author.

¹<https://huggingface.co/llm-jp/llm-jp-3-8x13b>

²<https://github.com/WorksApplications/uzushio>

³<https://github.com/datalab-to/surya>

⁴<https://www.xpdfreader.com/pdftotext-man.html>

⁵<https://warp.ndl.go.jp/>

Subset	Language	Est. Tokens [B]
Crawled HTML	Japanese	34.57
Crawled PDF (processed by Surya)	Japanese	0.17
Crawled PDF (processed by pdfototext)	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.

second quarter of 2015. We used documents from the second quarter of 2013 to the second quarter of 2015.

- **J-GLOBAL:** Japanese and English abstract texts from J-GLOBAL,⁶ 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 (?), 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 v0.3.0⁷ 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

⁶<https://jglobal.jst.go.jp/en>
⁷<https://github.com/llm-jp/megatron-lm/tree/v4>

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 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 $8 \times 13B$ -base model is publicly available on Hugging Face,⁸ and the 300B-token corpus is also disclosed on GitLab.⁹

B. Medical Corpus Construction

B.1. Corpus Composition

The SIP-med-llm- $8 \times 13B$ model was developed through continued pre-training of the $8 \times 13B$ -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

⁸Hugging Face repository: Hugging Face repository: <https://huggingface.co/llm-jp/llm-jp-3-8x13b>.

⁹GitLab repository: <https://gitlab.med-jp.nii.ac.jp/datasets/sip3-ja-general-web-corpus>.

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 <code>hprc/kaken-trans-ja-en</code> (https://huggingface.co/datasets/hprc/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.

baseline reported by Hayakawa and Arase (2020) and the gpt-4o-2024-08-06 model. The results

confirmed that the NICT translator achieves competitive performance compared to gpt-4o-2024-

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.

08-06. The performance metrics are presented in **Table B.2**. BLEU measures agreement with the ground truth using the SacreBLEU library¹⁰ with the MeCab tokenizer,¹¹ while COMET-22¹² and COMET-23¹³ 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.

¹⁰<https://github.com/mjpost/sacrebleu>

¹¹<https://pypi.org/project/mecab-python3/>

¹²<https://huggingface.co/Unbabel/wmt22-cometkiwi-da>

¹³<https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl>

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)
MT	Original	PubMedQA (Jin et al., 2019)
		MMLU-medical (Hendrycks et al., 2021b,a)
NER	Original	EJMMT (Hayakawa and Arase, 2020)
		MRNER-disease [†]
	Translation	MRNER-medicine [†]
		NRNER [†]
DC	Original	BC2GM (Smith et al., 2008)
		BC5Chem (Pavlova and Makhlouf, 2023)
STS	Original	BC5-Disease (Li et al., 2016)
		JNLPBA (Collier et al., 2004)
		NCBI Disease (Doğan et al., 2014)
		GRADE [†]
DC	Original	RRTNM [†]
		SMDIS [†]
STS	Original	JCSTS (Mutinda et al., 2021)

Table C.1: Detailed information about JMedBench. Among these datasets, MRNER-disease[†], MRNER-medicine[†], NRNER[†], GRADE[†], 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>.

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 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 samples¹⁴.

IgakuQA/IgakuQA-En IgakuQA contains 989 Japanese questions based on Japanese medical licensing examinations from 2018 to 2022 (Kasai

¹⁴<https://huggingface.co/datasets/nlp-waseda/JMMLU>

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, 2020). 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 Japanese¹⁵.

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 Japanese¹⁵.

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 Japanese¹⁵.

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 Makhoul, 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, 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 Japanese¹⁵.

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 Japanese¹⁵.

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 Japanese¹⁵.

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).

D. Bibliographical References

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.

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

¹⁵This benchmark is originally included in JMED-LLM (Japanese Medical Evaluation Dataset for Large Language Models): <https://github.com/sociocom/jmed-llm>

- 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)*.
- 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.
- Takeshi Hayakawa and Yuki Arase. 2020. [Fine-Grained Error Analysis on English-to-Japanese Machine Translation in the Medical Domain](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164.
- 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)*.
- 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 chat-gpt on japanese medical licensing examinations](#).
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.
- 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*.
- 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*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan 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 Makhoul. 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.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, 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*.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucu-rull, 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. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *arXiv preprint arXiv:2307.09288*.