ELAINE-medLLM: Lightweight English Japanese Chinese Trilingual Large Language Model for Bio-medical Domain

Ken Yano[1*], **Zheheng Luo**[2*], **Jimin Huang**[3], **Qianqian Xie**[3], **Masaki Asada**[1],
Chenhan Yuan[2], **Kailai Yang**[2], **Makoto Miwa**[4,1], **Sophia Ananiadou**[2,1], **Jun'ichi Tsujii**[1]
[1]National Institute of Advanced Industrial Science and Technology,
[2]The University of Manchester, [3]The Fin AI, [4]Toyota Technological Institute
Correspondence: yano.ken@aist.go.jp

Abstract

We propose ELAINE (EngLish-jApanese-chINesE)-medLLM, a trilingual (English, Japanese, Chinese[1]) large language model adapted for the bio-medical domain based on Llama-3-8B. The training dataset was carefully curated in terms of volume and diversity to adapt to the biomedical domain and endow trilingual capability while preserving the knowledge and abilities of the base model. The training follows 2-stage paths: continued pre-training and supervised fine-tuning (SFT). Our results demonstrate that ELAINE-medLLM exhibits superior trilingual capabilities compared to existing bilingual or multilingual medical LLMs without severely sacrificing the base model's capability.

1 Introduction

General open-source foundation LLMs such as Llama series (Touvron et al., 2023a,b) are trained by a large amount of English with limited multilingual data. Therefore, their performance often underperforms in non-English languages such as Japanese and Chinese and in specific domains such as biomedical. To address this limitation, we developed ELAINE-medLLM, a trilingual LLM that supports English, Japanese, and Chinese and is adapted for a bio-medical domain.

Although there exist many previous studies for bilingual and multilingual medLLM (Singhal et al., 2022; Li et al., 2023; Singhal et al., 2023; Chen et al., 2023c), this is the first attempt that focuses on the trilingual aspect of the medLLM. Mixed language training of medLLM is practical because it can harness language-independent biomedical knowledge during training. The resultant medLLM can be further fine-tuned for independent languages or tasks with less training data and time than training it from a bare base model. However, continued pretraining from a base model to endow non-English capability or to adapt to specific domains comes with the issue of degradation of the capabilities of the base model (Scialom et al., 2022; Fujii et al., 2024; Ankit Pal, 2024). In this paper, we demonstrate that the issue can be remedied by careful data selection and appropriate data processing for the continued pretraining, followed by instruction-following fine-tuning.

Following prior studies (Chen et al., 2023c; Lewkowycz et al., 2022), we incorporate data that closely align with the distribution of the models encountered during pre-training. Due to its public nature, Wikipedia is commonly used in pre-training, so we include the tri-lingual Wikipedia in our replay training data. Since our training scenario involves both domain adaptation and multilingual support, we increase the proportion of replay data and adjust the allocation of each language based on the size of its respective dataset in the biomedical domain.

The experimental results over various biomedical QA benchmarks show that our pertained medLLM has strong trilingual capabilities, and our finetuned medLLM also performs competitively almost equally in three languages.

2 Related work

Med-PaLM (Singhal et al., 2022) based on PaLM (Chowdhery et al., 2022) is one of the earliest LLMs adapted for the bio-medical domain from general LLMs. Since then, there are numerous medLLMs (Singhal et al., 2022; Li et al., 2023; Singhal et al., 2023; Chen et al., 2023c) developed, and their performance has kept rising as being adapted from newer base models such as Llama series (Touvron et al., 2023a,b), which have been released every few months. Along with English meddlers, there have been quite a few stud-

---

[*] These authors contributed equally to this work.

[1] The order reflects the size of the training data we used for each language.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4670–4688
January 19–24, 2025. ©2025 Association for Computational Linguistics

ies to develop bilingual or multi-lingual medLLMs which support non-English such as Chinese and other languages (Wang et al., 2023; Zhang et al., 2023; Bao et al., 2023; Wang et al., 2024).

The available biomedical datasets for languages other than English and Chinese are limited and scarce. Our medLLM is one of the initial efforts to utilize strong Japanese proficiency along with English and Chinese effectively. We aim to share our findings to support future research on low-resourced medLLM.

lang	category	size [Mbytes]	%
en	Guidelines	615	0.26
	PMC	89,105	37.15
	Papers	1,165	0.49
	Pubmed	47,401	19.76
	Textbook	1,373	0.57
	Web Text	2,228	0.93
	Wikipedia	28,417	11.85
ja	Papers	2,406	1.0
	Pubmed (translation)	43,601	18.18
	Wikipedia	13,253	5.52
zh	Guidelines	353	0.15
	Papers	1,028	0.43
	Textbook	926	0.39
	Web Text	2,931	1.22
	Wikipedia	5,074	2.12
total		239,880	100.0

Table 1: The size of pretraining text in the bio-medical domain and Wikipedia.

3 Training Dataset

3.1 Unstructured domain text for continued pretraining

For continued pretraining, we collected English, Japanese, and Chinese text in the bio-medical domain. The domain text collected is classified into six categories: 1) scientific papers, 2) medical guidelines, 3) web text related to biomedical, 4) textbook of biomedical, 5) PubMed abstracts, and 6) PubMed Central (PMC) archives. For the Japanese PubMed abstract, we used the original English PubMed abstract translated in Japanese[1].

We used only open-licensed text except for the Japanese biomedical papers from J-STAGE (the details are in Appendix A). We also included trilingual Wikipedia to prevent catastrophic forgetting of the base model (Scialom et al., 2022). Table 1 shows the details of the total file sizes in each language and the category. The ratio of text size for languages was 71.0%, 24.7%, and 4.3% for English, Japanese, and Chinese, respectively.

Data cleaning and processing

Text collected from web crawlers and books can contain low-quality and toxic content. Text automatically converted from PDF, such as academic papers, is inevitable to contain nonsensical or irrelevant text data originating from metadata such as titles, subtitles, header, footer, table content, and image captions. Low-quality data can have a detrimental impact on pretraining LLMs, so they should be adequately cleaned.

However, no standard methodology applies to all LLMs, so we also took an empirical approach. Specifically, we used Data-Juicer (Chen et al., 2023a), which offers a data processing pipeline for constructing datasets for LLM models. Initially, it supports only English and Chinese; however, by providing the ARPA language model such as by KenLM (Heafield, 2011) and the SentencePiece tokenizer model (Kudo and Richardson, 2018), it can help other languages, such as Japanese. It includes over 50 built-in operators for easy composition and extension of data mapping and filtering operations. A configuration file can fully configure the data processing pipeline. In Appendix C, we show the *yaml* configuration file to process PMC dataset. Other configuration files are very similar to this configuration. After we applied the data processing to the original dataset, we obtained the following token statistics summarized in Table 2. The total number of tokens is about 30B, much less than 160B from Chinchilla's law (Hoffmann et al., 2022), which is the model's parameter size times 20 (8B × 20). However, for the case of continued pretraining, this token count turns out to be good enough as verified in Section 5.

lang	#tokens [M]	%
en	22,825	75.91
ja	6,490	21.59
zh	752	2.50
total	30,067	100.0

Table 2: The number of tokens used for the continued pretraining

---

[1] Provided by Dr. Yoshimasa Kawazoe at The University of Tokyo Medical School Hospital

3.2 Conversational QA dataset for supervised fine-tuning

We collected various conversational QA datasets in the bio-medical domain from different data sources. Many such datasets exist for English and Chinese. For English, we used Medical Meadow in MedAlpaca (Han et al., 2023), HealthCareMagic, and iCliniq dataset used in ChatDoctor (Li et al., 2023). We adapted the augmented QA dataset from HuatuoGPT-2 (Chen et al., 2023b) for Chinese and English. For Japanese, we could not find any existing conversational QA dataset in the biomedical domain, so instead, we used existing alpaca datasets (Taori et al., 2023) in the general domain translated in Japanese. We also tried to add the part of the English conversational QA dataset by translating it into Japanese. However, our preliminary experiments revealed that noisy translated QA datasets hurt the performances of the generated LLM. This aligns with the finding (Li et al., 2024) that low-quality self-augmented instruction-following datasets do not improve performance despite scaling up data quantity. So, instead, we only used a translated alpaca dataset, which we verified kept the quality of the Japanese text at a permissible level.

For each sample in the collected QA datasets, we converted it from the original format to a uniform format, as shown in Appendix D. For instance, if the original instruction-following data consists of three parts such as (*instruction*, *input*, *output*) such as in the alpaca dataset from Stanford, we concatenate ‘*instruction*’ and ‘*input*’ parts as one continuous text that follows the “Human:” placeholder in our conversational format.

We used this instruction-following QA format for the English, Japanese, and Chinese datasets so that the trained LLM would naturally learn transferable knowledge between QA datasets in different languages.

4 Experiments

We used two nodes at ABCI 2.0[2] each holding 8 NVIDIA A100 with 40GiB to pretrain and finetune ELAINE-medLLM. For training, we used the distributed training with DeepSpeed (Rajbhandari et al., 2020; Rasley et al., 2020) stage-3 and set the training parameters as follows: the device batch size of 1; the gradient accumulation of 16; the learning rate of 1e-5; the cosine scheduler; the weight decay of 0.00001; the warm-up ratio of 0.05. It took about six days to complete the continuous pretraining for one epoch. We used the same training setup for the finetuning except for setting the warm-up ratio to 0.01 and the number of training epochs to 3. It took about two days to complete the finetuning.

lang	original source	#instructions [k]	%
en	HuatuoGPT-2	1,677	33.26
	HealthCareMagic	112	2.22
	iCliniq	7	0.15
	Medical Meadow	858	17.02
ja	alpaca translated (*)	52	1.03
zh	crawled from web	792	15.70
	HuatuoGPT-2	1,544	30.62
	total	5,043	100.00

Table 3: The counts of the conversational QA instructions. (*) general domain dataset

For evaluation, we used a wide spectrum of open-sourced multiclass and multilabel QA tasks in the biomedical domain to evaluate our trilingual ELAINE-medLLM and ELAINE-medLLM-instruct and compare the results with multiple baseline models. The details are explained below.

Bio-medical QA benchmarks

Specifically, for English, we used the following five QA datasets: MedQA, MedQA-4options (Jin et al., 2020), MMLU (Hendrycks et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). For Chinese, we used MedQA, MedQA-4options (Jin et al., 2020), and CMExam (Liu et al., 2023). For Japanese, We used JJSIMQA (Sukeda et al., 2023), IgakuQA (Kasai et al., 2023), and DenQA. DenQA is the Japanese national dentistry examination for 2024 and 2023, which is prepared by the authors. The details of each QA dataset are provided in Appendix B.

The answer to all the QA tasks is a selection from multiple choices except for Japanese QA datasets: JJSIMQA, IgakuQA, and DenQA, where the answer can be more than one selection from the choices; the instruction in the question prompt specifies the number.

We used the accuracy score to measure the performance, which is the number of correctly answered questions divided by the total number of questions. For the multilabel Japanese QA tasks, we used the exact match, which judges an an-

---

[2] https://docs.abci.ai/en/

Model	gemma -7b	Llama-2 -7b	Llama-3 -8B	Swallow -8B-v0.1	ELAINE -medLLM
Release	24-Mar	23-Jul	24-Apr	24-Jul	
			EN		
MMLU	63.65	45.02	**71.22**	65.96	67.80
MedMCQA	49.81	36.84	**56.97**	51.27	54.55
MedQA	43.38	30.13	**52.60**	45.90	50.47
MedQA-4op	48.82	36.59	**57.89**	52.92	57.73
PubMedQA	**71.52**	49.90	69.70	61.01	67.27
Average	55.44	39.70	**61.68**	55.41	59.56
			JA		
DenQA	18.60	10.63	18.88	22.24	**22.38**
IgakuQA	29.02	17.64	35.09	42.21	**44.06**
JJSIMQA	18.90	8.13	23.52	27.25	**29.45**
Average	22.17	12.13	25.83	30.57	**31.96**
			ZH		
CMExam	36.34	24.33	40.30	36.19	**46.03**
MedQA	40.54	25.02	44.96	40.89	**52.50**
MedQA-4op	43.03	29.61	51.15	48.00	**58.23**
Average	39.97	26.32	45.47	41.69	**52.25**

Table 4: Comparative analysis of the accuracy scores of various multiclass and multilabel medical QA tasks by the baseline pre-trained models without fine-tuning and our ELAINE-medLLM

swer to be correct if the selected answer exactly matches the ground truth answer without excess or shortage. Hence, they are more challenging than single-label selection QA tasks.

Baseline models

We selected 7B-class open-sourced LLMs as the baseline models for comparison. Specifically, we evaluated three foundation models: gemma-7B from Google, llama2-7B, and llama3-8B from Meta, and another six domain or language-adapted models, including our ELAINE-medLLM.

To ensure that benchmark scores reflect the true capability of LLM (Sainz et al., 2023), we excluded those models that used the training split of benchmarks or have not disclosed detailed information about their training dataset. These include MMed-Llama-3-8B (Qiu et al., 2024), Qwen1.5-7B (Bai et al., 2023), etc. However, note that it is extremely difficult to completely exclude the data leakage condition because during LLM training, the benchmark dataset may be used as training without the generator publicly revealing it.

Llama-3-Swallow-8B-v0.1 (Fujii et al., 2024) is the competitive Japanese model based on Llama3. Medalpaca-7b (Han et al., 2023) based on Llama is an expansion from alpaca and is fine-tuned for the medical domain. Meditron-7B (Chen et al., 2023c) is adapted from Llama2 to the medical domain. OpenBioLLM-8B (Ankit Pal, 2024) is built upon Llama3 and finetuned for the biomedical domain. Appolo-7B (Wang et al., 2024) based on Qwen 1.8B (Bai et al., 2023) is a multilingual biomedical LLM covering English, Chinese, French, Spanish, Arabic, and Hindi. For gemma-7B, llama2-7B, llama3-8B, Llama-3-Swallow-8B-v0.1, and ELAINE-medLLM, we evaluated both pre-trained models without instruction fine-tuning and the corresponding fine-tuned model.

5 Results

We used a consistent input format for each language to evaluate the multiple benchmarks. The samples of the input format are shown in the Appendix E. For all the experiments, we used 5-shot in-context learning, using the first five examples in each benchmark dataset as the in-context input samples, except for the cases when the input length exceeded the maximum length supported, in which case we used 3-shot. The benchmark datasets and the evaluation code are available at our Github site [3].

Table 4 shows the benchmark scores for pre-trained models without instruction fine-tuning. The values in the table show each benchmark's accuracy and the average accuracy score grouped by language. The highest scores are bold, and the second-highest scores are underlined. As we can see, the ELAINE-medLLM outperforms other pre-trained models for Japanese and Chinese benchmarks and ranks second for English benchmarks, maintaining competitive English performance. Japanese-adapted Swallow-8B-v0.1

---

[3] https://github.com/aistairc/medLLM_QA_benchmark

Model	gemma -7b-it	Llama-2 -7b-chat	Llama-3 -8B -Instruct	Swallow -8B -Instruct -v0.1	med alpaca -7b	meditron -7b	Open BioLLM -8B	Apollo -7B	ELAINE -medLLM -instruct
Release	24-Mar	23-Jul	24-Apr	24-Jul	23-Apr	23-Dec	24-May	24-Mar	
					EN				
MMLU	50.55	48.71	72.79	66.88	51.48	47.32	**73.43**	68.17	72.69
MedMCQA	41.07	35.97	**60.89**	53.85	36.02	34.35	55.03	53.85	55.07
MedQA	33.12	30.99	**57.65**	47.95	31.15	29.18	50.00	45.98	55.76
MedQA-4op	39.67	38.09	61.28	56.07	39.35	32.26	56.78	53.86	**61.36**
PubMedQA	67.07	63.64	**78.99**	64.65	55.15	39.19	65.86	75.34	75.35
Average	46.30	43.48	**66.32**	57.88	42.63	36.46	60.22	59.44	64.05
					JA				
DenQA	13.71	12.03	19.72	**23.78**	10.91	9.79	18.18	17.90	22.24
IgakuQA	25.51	20.80	40.45	**44.01**	17.74	18.20	33.03	32.28	43.36
JJSIMQA	12.09	10.55	25.93	**26.81**	10.77	8.35	21.98	20.66	24.40
Average	17.10	14.46	28.70	**31.53**	13.14	12.11	24.40	23.61	30.00
					ZH				
CMExam	30.90	25.43	**52.01**	41.11	23.58	23.85	39.07	49.99	48.85
MedQA	29.03	25.37	**62.99**	45.05	24.99	25.46	42.59	58.29	55.80
MedQA-4op	34.96	32.30	**68.40**	51.27	30.11	29.82	48.73	62.99	61.59
Average	31.63	27.70	**61.13**	45.81	26.23	26.38	43.46	57.09	55.41

Table 5: Comparative analysis of the accuracy scores of various multiclass and multilabel medical QA tasks by the baseline fine-tuned models and our ELAINE-medLLM-instruct

shows competitiveness for Japanese. However, it lags behind ELAINE-medLLM in English and Chinese.

Table 5 shows the benchmark scores for fine-tuned models. For the English benchmarks, Llama3-8B-Instruct outperformed the other fine-tuned models for most of the benchmarks. However, our ELAINE-medLLM-instruct maintains competitive performance and ranks second for the average score.

For Japanese benchmarks, the Japanese-adapted Swallow-8B-Instruct-v0.1 shows the highest performance, making our ELAINE-medLLM-instruct rank second.

Surprisingly, Llama3-8B-Instruct outperformed the other fine-tuned models for Chinese benchmarks, showing strong Chinese performance. Apollo-7B, based on Qwen-7B, is second for all Chinese benchmarks, inheriting a solid Chinese ability from the base model. However, our ELAINE-medLLM-instruct also shows the competitiveness of Chinese performance.

To summarize, our trilingual pre-trained ELAINE-medLLM and the fine-tuned ELAINE-medLLM-instruct are robust, showing almost equal competence for all three languages. It is notable that among those LLMs based on Llama-3, Swallow-8B-v0.1, OpenBioLLM-8B, and ELAINE-medLLM, only our model does not show noticeable degradation for the English benchmark scores compared with the base model. Our results indicate that careful dataset preparation and training can prevent inevitable severe degradation issues when adapting a base LLM to non-English languages or domains.

QA for healthcare

To demonstrate the conversational and trilingual capability for healthcare QA by ELAINE-medLLM-instruct, we show the answers from the model for each of the four questions translated in each language in Appendix F. It shows our model's trilingual prowess by answering short and long questions succinctly in the same language as the corresponding question. However, we noticed that answers to long questions in English are rather indecisive, and the model tried to obtain more detailed information from human. On the other hand, the answers for Japanese questions are mostly decisive. The cause of these inconsistent qualitative answers to the questions in different languages needs further analysis in future work.

6 Conclusion

We proposed a trilingual medLLM called ELAINE based on Llama-3-8B. The experimental results on bio-medical QA tasks reveal that it performs competitively with strong English medLLMs and simultaneously shows trilingual prowess on Japanese and Chinese medical benchmarks.

Limitations

The evaluation of medLLMs was conducted using QA datasets. The metrics can measure the professional knowledge of medLLM, as exemplified by the scores of medical examinations. However, these metrics may not adequately capture medLLM's performance, particularly in assessing the usefulness and safety outputs obtained from conversational scenarios. Researchers have increasingly adopted human-centric evaluation approaches to address this limitation. These approaches will provide a more realistic assessment of medLLMs in real-world applications and constitute our future work.

Acknowledgement

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. 2023a. Data-juicer: A one-stop data processing system for large language models. *Preprint*, arXiv:2309.02033.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023b. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

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. 2023c. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

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. *Preprint*, arXiv:2404.17790.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *Preprint*, arXiv:2303.18027.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation. *Preprint*, arXiv:2308.06259.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset. *Preprint*, arXiv:2306.03030.

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 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera 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. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale

Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. 2023. JMedLoRA: Medical Domain Adaptation on Japanese Large Language Models using Instruction-tuning. *arXiv preprint arXiv:2310.10083*.

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

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. *Preprint*, arXiv:2302.13971.

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. *Preprint*, arXiv:2307.09288.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *Preprint*, arXiv:2304.06975.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

A J-Stage Japanese articles in biomedical domain

J-Stage[4] is a platform for scholarly publications in Japan. We are licensed by the Japan Science and Technology Agency (JST) to download about 1.16 million Japanese PDF files and 0.41 million English PDF files. We used 570,000 Japanese PDF files categorized in the biomedical domain (300,000 for publication years from 1996 to 2000 and 270,000 for publication years after 2000).

Tika[5] was used for PDF-to-text conversion. Then, a filtering process was applied to remove unnecessary parts to obtain clean text. We first randomly selected 20 papers for the filtering process and performed a manual binary labeling annotation at the token level to determine whether each token should be retained or discarded. Next, we trained a token-level classification model on this annotated data using a released PLM checkpoint. We then applied the trained model to clean all converted texts. For the PLM, we utilized the Japanese DeBERTa V2 large model.

B QA benchmarks for bio-medical domain

MedQA, MedQA-4options (Jin et al., 2020) It is collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese. It contains train, dev, and test splits; we used only test splits. The original question consists of 5 options for answer. MedQA-4options contains four options that are reduced from the original.

MMLU (Hendrycks et al., 2021) The test covers 57 tasks, including elementary mathematics, US history, computer science, and law. We only selected tasks related to bio-medical, namely anatomy, clinical_knowledge, college_biology, college_medicine, medical_genetics, and professional_medicine.

MedMCQA (Pal et al., 2022) The examination questions are collected from the official website of

---

[4] https://www.jstage.jst.go.jp/

[5] https://tika.apache.org/

the India Institute of Medical Science, as well as from open websites and books. It consists of train, dev, and test splits, and we used only test split.

PubMedQA (Jin et al., 2019) It is collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe using the corresponding abstracts. PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances. Each PubMedQA instance is composed of (1) a question which is either an existing research article title or derived from one; (2) a context, which is the corresponding abstract without its conclusion; (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

CMExam (Liu et al., 2023) It is sourced from the Chinese National Medical Licensing Examination. CMExam consists of 60K+ multiple-choice questions for standardized and objective evaluations, as well as solution explanations for model reasoning evaluation in an open-ended manner. It consists of train, val, and test splits, and we used only test split.

JJSIMQA (Sukeda et al., 2023) It is a dataset comprising 5-choice questions extracted from the Journal of the Japanese Society of Internal Medicine (JJSIM).

IgakuQA (Kasai et al., 2023) It contains the exam problems from the Japan National Medical Practitioners Qualifying Examination (NMPQE) and their answers in the past five years (from 2018 through 2023) extracted from the official website of the Ministry of Health, Labor and Welfare in Japan [6]. We concatenate all the exams for the past five years to make the benchmark for evaluation.

DenQA It contains the exam problems from the Japan National Dentistry Examination and their answers in the past two years (from 2023 through 2024) extracted from the official website of the Ministry of Health, Labor and Welfare in Japan [6]. We concatenate all the exams for the past two years to make the benchmark for evaluation. This dataset will become available at our GitHub site.

---

[6] https://www.mhlw.go.jp/stf/english/index.html

C A sample configuration file for data-juicer

The following is the *yaml* configuration for data-juicer[7] used to process the Pubmed Central (PMC) dataset. The configurations for the other datasets are very similar to it.

```
# global parameters
project_name: 'all'                                   # project name for distinguish your configs
dataset_path: '/path/to/your/dataset'                 # path to your dataset directory or file
export_path: '/path/to/result/dataset.jsonl'          # path to processed result dataset.
np: 4                                                 # number of subprocess to process your dataset
open_tracer: true                                    # whether to open the tracer to trace the changes
                                                      # It might take more time when opening tracer


# process schedule: a list of several process operators with their arguments
process:
  # Mapper ops. Most of these ops need no arguments.
  - clean_email_mapper:                               # remove emails from text.
  - clean_links_mapper:                               # remove web links from text.
  - clean_copyright_mapper:                           # remove copyright comments.
  - expand_macro_mapper:                              # expand macro definitions in Latex text.
  - fix_unicode_mapper:                               # fix unicode errors in text.
 - punctuation_normalization_mapper:            # normalize unicode punctuations to English punctuations.
  - remove_bibliography_mapper:                       # remove bibliography from Latex text.
  - remove_comments_mapper:                           # remove comments from Latex text, code, etc.
 - whitespace_normalization_mapper:            # normalize different kinds of whitespaces to English whitespace.

  # Filter ops
  - alphanumeric_filter:                              # filter text with alphabet/numeric ratio
      tokenization: false                              # whether to count the ratio of alphanumeric
      min_ratio: 0.2787                                # the min ratio of filter range
  - average_line_length_filter:                       # filter text with the average length of lines
      max_len: 1500
  - character_repetition_filter:                    # filter text with the character repetition ratio
      rep_len: 10                                      # repetition length for char-level n-gram
      #min_ratio: 0.0                                  # the min ratio of filter range
      max_ratio: 0.1                                   # the max ratio of filter range
  - flagged_words_filter:                             # filter text with the flagged-word ratio
      lang: en                                         # consider flagged words in what language
      tokenization: true                               # whether to use model to tokenize documents
      max_ratio: 0.00195                               # the max ratio to filter text
      #flagged_words_dir: ./assets                     # directory to store flagged words dictionaries
     #use_words_aug: false                          # whether to augment words, especially for Chinese
      #words_aug_group_sizes: [2]                      # the group size of words to augment
      #words_aug_join_char: ""                         # the join char between words to augment

 - language_id_score_filter:                    # filter text in specific language with language scores
      lang: en                                         # keep text in what language
      min_score: 0.9                                   # the min language scores to filter text
  - maximum_line_length_filter:                       # filter text with the maximum length of lines
      #min_len: 10                                     # the min length of filter range
      max_len: 7328                                    # the max length of filter range
 - perplexity_filter:                           # filter text with perplexity score out of specific range
      lang: en                                         # compute perplexity in what language
      max_ppl: 1000                                    # the max perplexity score to filter text
 - special_characters_filter:                  # filter text with special-char ratio out of specific range
      #min_ratio: 0.0                                  # the min ratio of filter range
      max_ratio: 0.3                                   # the max ratio of filter range
  - text_length_filter:                               # filter text with length out of specific range
      min_len: 10                                      # the min length of filter range
      max_len: 10000                                   # the max length of filter range
 - words_num_filter:                             # filter text with number of words out of specific range
      lang: en                                         # sample in which language
      tokenization: true                               # whether to use model to tokenize documents
      min_num: 20                                      # the min number of filter range
      max_num: 23305                                   # the max number of filter range
  - word_repetition_filter:                           # filter text with the word repetition ratio
      lang: en                                         # sample in which language
```

---

[7] https://github.com/modelscope/data-juicer

```
      tokenization: true                           # whether to use model to tokenize documents
      rep_len: 10                                  # repetition length for word-level n-gram
      #min_ratio: 0.0                              # the min ratio of filter range
      max_ratio: 0.5981                            # the max ratio of filter range

  - document_simhash_deduplicator:             # deduplicate text samples using SimHash-LSH method
      tokenization: space                          # tokenization method for text.
      window_size: 6                               # window size of shingling
      num_blocks: 6                                # number of blocks in SimHash computing
      hamming_distance: 4                          # the max hamming distance to regard 2 samples
      lowercase: true                              # whether to convert text to lower case
     ignore_pattern: null                        # whether to ignore sub-strings with specific pattern
```

D A sample of conversational instruction data

The following is a sample of instruction-following data. The red part is the instruction. During training for the instruction-following fine-tuning, we only optimize the loss from the answer (the blue part).

```
Human:
If you are a doctor, please answer the medical questions based on the patient's
↪ description. I woke up this morning feeling the whole room is spinning when i
↪ was sitting down. I went to the bathroom walking unsteadily, as i tried to focus
↪ i feel nauseous. I try to vomit but it wont come out.. After taking panadol and
↪ sleep for few hours, i still feel the same.. By the way, if i lay down or sit
↪ down, my head do not spin, only when i want to move around then i feel the whole
↪ world is spinning.. And it is normal stomach discomfort at the same time?
↪ Earlier after i relieved myself, the spinning lessen so i am not sure whether
↪ its connected or coincidences.. Thank you doc!
Assistant:
Hi, Thank you for posting your query. The most likely cause for your symptoms is
↪ benign paroxysmal positional vertigo (BPPV), a type of peripheral vertigo. In
↪ this condition, the most common symptom is dizziness or giddiness, which is made
↪ worse with movements. Accompanying nausea and vomiting are common. The condition
↪ is due to problem in the ear, and improves in a few days on own. Betahistine
↪ tablets would help relieve your symptoms. Doing vestibular rehabilitation or
↪ adaptation exercises would prevent the recurrence of these symptoms. An ENT
↪ evaluation would also help. I hope it helps. Best wishes, Chat Doctor.
```

E Sample medical QA benchmark input format for 3-shot in-context learning

E.1 MedQA

```
Question: A junior orthopaedic surgery resident is completing a carpal tunnel
↪ repair with the department chairman as the attending physician. During the
↪ case, the resident inadvertently cuts a flexor tendon. The tendon is repaired
↪ without complication. The attending tells the resident that the patient will do
↪ fine, and there is no need to report this minor complication that will not harm
↪ the patient, as he does not want to make the patient worry unnecessarily. He
↪ tells the resident to leave this complication out of the operative report.
↪ Which of the following is the correct next action for the resident to take?
a: Disclose the error to the patient but leave it out of the operative report
b: Disclose the error to the patient and put it in the operative report
c: Tell the attending that he cannot fail to disclose this mistake
d: Report the physician to the ethics committee
e: Refuse to dictate the operative report
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
```

```
c
Question: A 67-year-old man with transitional cell carcinoma of the bladder comes
 ↪  to the physician because of a 2-day history of ringing sensation in his ear. He
 ↪  received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone
 ↪  audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial
 ↪  effect of the drug that caused this patient's symptoms is most likely due to
 ↪  which of the following actions?
a: Inhibition of thymidine synthesis
b: Inhibition of proteasome
c: Hyperstabilization of microtubules
d: Generation of free radicals
e: Cross-linking of DNA
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
e
Question: Two weeks after undergoing an emergency cardiac catherization with
 ↪  stenting for unstable angina pectoris, a 61-year-old man has decreased urinary
 ↪  output and malaise. He has type 2 diabetes mellitus and osteoarthritis of the
 ↪  hips. Prior to admission, his medications were insulin and naproxen. He was
 ↪  also started on aspirin, clopidogrel, and metoprolol after the coronary
 ↪  intervention. His temperature is 38řC (100.4řF), pulse is 93/min, and blood
 ↪  pressure is 125/85 mm Hg. Examination shows mottled, reticulated purplish
 ↪  discoloration of the feet. Laboratory studies show:
Hemoglobin count 14 g/dL
Leukocyte count 16,400/mm3
Segmented neutrophils 56%
Eosinophils 11%
Lymphocytes 31%
Monocytes 2%
Platelet count 260,000/mm3
Erythrocyte sedimentation rate 68 mm/h
Serum
Urea nitrogen 25 mg/dL
Creatinine 4.2 mg/dL
Renal biopsy shows intravascular spindle-shaped vacuoles. Which of the following is
 ↪  the most likely cause of this patient's symptoms?"
a: Renal papillary necrosis
b: Allergic interstitial nephritis
c: Cholesterol embolization
d: Eosinophilic granulomatosis with polyangiitis
e: Polyarteritis nodosa
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
c
Question: A 39-year-old woman is brought to the emergency department because of
 ↪  fevers, chills, and left lower quadrant pain. Her temperature is 39.1řC
 ↪  (102.3řF), pulse is 126/min, respirations are 28/min, and blood pressure is
 ↪  80/50 mm Hg. There is blood oozing around the site of a peripheral intravenous
 ↪  line. Pelvic examination shows mucopurulent discharge from the cervical os and
 ↪  left adnexal tenderness. Laboratory studies show:
Platelet count 14,200/mm3
Fibrinogen 83 mg/mL (N = 200–430 mg/dL)
```

```
D-dimer 965 ng/mL (N < 500 ng/mL)
When phenol is applied to a sample of the patient's blood at 90řC, a phosphorylated
 ↪  N-acetylglucosamine dimer with 6 fatty acids attached to a polysaccharide side
 ↪  chain is identified. A blood culture is most likely to show which of the
 ↪  following?"
a: Coagulase-positive, gram-positive cocci forming mauve-colored colonies on
 ↪  methicillin-containing agar
b: Encapsulated, gram-negative coccobacilli forming grey-colored colonies on
 ↪  charcoal blood agar
c: Spore-forming, gram-positive bacilli forming yellow colonies on casein agar
d: Lactose-fermenting, gram-negative rods forming pink colonies on MacConkey agar
e: Gamma-hemolytic, gram-positive cocci forming green colonies on vancomycin agar
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
```

E.2 PubMedQA

```
Context: Patients presenting with transient ischemic attack or stroke may have
 ↪  symptom-related lesions on acute computed tomography angiography (CTA) such as
 ↪  free-floating intraluminal thrombus (FFT). It is difficult to distinguish FFT
 ↪  from carotid plaque, but the distinction is critical as management differs. By
 ↪  contouring the shape of these vascular lesions ("virtual endarterectomy"),
 ↪  advanced morphometric analysis can be performed. The objective of our study is
 ↪  to determine whether quantitative shape analysis can accurately differentiate
 ↪  FFT from atherosclerotic plaque. We collected 23 consecutive cases of suspected
 ↪  carotid FFT seen on CTA (13 men, 65 ś 10 years; 10 women, 65.5 ś 8.8 years).
 ↪  True-positive FFT cases (FFT+) were defined as filling defects resolving with
 ↪  anticoagulant therapy versus false-positives (FFT-), which remained unchanged.
 ↪  Lesion volumes were extracted from CTA images and quantitative shape
 ↪  descriptors were computed. The five most discriminative features were used to
 ↪  construct receiver operator characteristic (ROC) curves and to generate three
 ↪  machine-learning classifiers. Average classification accuracy was determined by
 ↪  cross-validation. Follow-up imaging confirmed sixteen FFT+ and seven FFT-
 ↪  cases. Five shape descriptors delineated FFT+ from FFT- cases. The logistic
 ↪  regression model produced from combining all five shape features demonstrated a
 ↪  sensitivity of 87.5% and a specificity of 71.4% with an area under the ROC curve
 ↪  = 0.85 ś 0.09. Average accuracy for each classifier ranged from 65.2%-76.4%.
Question: Can shape analysis differentiate free-floating internal carotid artery
 ↪  thrombus from atherosclerotic plaque in patients evaluated with CTA for stroke
 ↪  or transient ischemic attack?
a: yes
b: no
c: maybe
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
a
```

```
Context: To determine the relationship between injury severity surrogates and other
↪  patient factors with the development and severity of heterotopic ossification
↪  (HO) following open reduction internal fixation of acetabular fractures treated
↪  with a posterior approach. Retrospective review. Academic level 1 trauma
↪  center. Two hundred forty-one patients who were treated through a posterior
↪  approach with a minimum of 6-month radiographic follow-up were identified from
↪  an acetabular fracture database. None. The occurrence and severity (Brooker
↪  Grade III/IV) of HO 6 months postsurgery. Length of stay (LOS) in the intensive
↪  care unit (ICU), non-ICU LOS>10 days, and HO prophylaxis with external
↪  radiation beam therapy (XRT) were significantly associated with the development
↪  of HO in a multivariate model [ 1-2 days, odds ratio (OR) = 4.33, 95% confidence
↪  interval (CI): 1.03-18.25; 3-6 days, OR = 4.1, 95% CI, 1.27-13.27;>6 days, OR =
↪  11.7, 95% CI, 3.24-42.22; non-ICU LOS>10 days (vs. 0-6 days): OR = 7.6, 95% CI,
↪  2.6-22.25; XRT HO prophylaxis: OR = 0.29, 95% CI, 0.10-0.85]. Other variables
↪  evaluated in multivariate modeling not significantly associated with
↪  development and severity of HO included age, gender, mechanism of injury,
↪  injury severity score, presence of neurologic injury, Letournel fracture type,
↪  occurrence of hip dislocation, interval from injury to surgery, operative time,
↪  and estimated blood loss.
Question: Do Surrogates of Injury Severity Influence the Occurrence of Heterotopic
↪  Ossification in Fractures of the Acetabulum?
a: yes
b: no
c: maybe
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
c
Context: To measure the dimensions of compensatory hypertrophy of the middle
↪  turbinate in patients with nasal septal deviation, before and after
↪  septoplasty. The mucosal and bony structures of the middle turbinate and the
↪  angle of the septum were measured using radiological analysis before
↪  septoplasty and at least one year after septoplasty. All pre- and
↪  post-operative measurements of the middle turbinate were compared using the
↪  paired sample t-test and Wilcoxon rank sum test. The dimensions of bony and
↪  mucosal components of the middle turbinate on concave and convex sides of the
↪  septum were not significantly changed by septoplasty. There was a significant
↪  negative correlation after septoplasty between the angle of the septum and the
↪  middle turbinate total area on the deviated side (p = 0.033).
Question: Does septoplasty change the dimensions of compensatory hypertrophy of the
↪  middle turbinate?
a: yes
b: no
c: maybe
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
b
```

```
Context: To assess whether perspective-taking, which researchers in other fields
↪  have shown to induce empathy, improves patient satisfaction in encounters
↪  between student-clinicians and standardized patients (SPs). In three studies,
↪  randomly assigned students (N = 608) received either a perspective-taking
↪  instruction or a neutral instruction prior to a clinical skills examination in
↪  2006-2007. SP satisfaction was the main outcome in all three studies. Study 1
↪  involved 245 third-year medical students from two universities. Studies 2 and 3
↪  extended Study 1 to examine generalizability across student and SP
↪  subpopulations. Study 2 (105 physician assistant students, one university)
↪  explored the effect of perspective-taking on African American SPs'
↪  satisfaction. Study 3 (258 third-year medical students, two universities)
↪  examined the intervention's effect on students with high and low baseline
↪  perspective-taking tendencies. Intervention students outscored controls in
↪  patient satisfaction in all studies: Study 1: P = .01, standardized effect size
↪  = 0.16; Study 2: P = .001, standardized effect size = 0.31; Study 3: P = .009,
↪  standardized effect size = 0.13. In Study 2, perspective-taking improved
↪  African American SPs' satisfaction. In Study 3, intervention students with high
↪  baseline perspective-taking tendencies outscored controls (P = .0004,
↪  standardized effect size = 0.25), whereas those with low perspective-taking
↪  tendencies did not (P = .72, standardized effect size = 0.00).
Question: Does perspective-taking increase patient satisfaction in medical
↪  encounters?
a: yes
b: no
c: maybe
Be sure to choose exactly 1 from a, b, c, d, e.
Answer:
```

E.3 IgakuQA

```
問題: 高アンモニア血症をきたす疾患はどれか。
a: Gaucher 病
b: von Gierke 病
c: Hurler 症候群
d: メープルシロップ尿症
e: オルニチントランスカルバミラーゼ欠損症
必ず a,b,c,d,e の中からちょうど 1 個選んでください。
答え:
e
問題: 続発性無汗症の原因とならないのはどれか。
a: 糖尿病
b: Fabry 病
c: Sjögren 症候群
d: 甲状腺機能低下症
e: 自家感作性皮膚炎
必ず a,b,c,d,e の中からちょうど 1 個選んでください。
答え:
e
問題: 高尿酸血症をきたさないのはどれか。
a: サイアザイド系利尿薬
b: Lesch-Nyhan 症候群
c: 腫瘍崩壊症候群
```

d: Fanconi 症候群
e: 慢性腎不全
必ず a,b,c,d,e の中からちょうど 1 個選んでください。
答え:
d
問題: 腸管出血性大腸菌感染症と診断された場合に正しいのはどれか。
a: 入院勧告の対象となる。
b: 届出は一週間以内に行う。
c: 届出先は市町村長である。
d: 医療費は全額公費負担となる。
e: 児童の場合は出席停止となる。
必ず a,b,c,d,e の中からちょうど 1 個選んでください。
答え:

E.4 CMExam

问题: 对医疗废物的收集, 运转, 贮存, 处置中的疾病防治工作进行定期检测监督检查的部门是
a: 市容监督机关
b: 城市规划行政部门
c: 卫生行政主管部门
d: 检验检疫行政主管部门
e: 环境保护行政部门
请确保从 a、b、c、d、e 中准确选择 1。
答话:
c
问题: 既治湿热泻痢, 又治感冒发热的药是
a: 射干
b: 白头翁
c: 马齿苋
d: 穿心莲
e: 白花蛇舌草
请确保从 a、b、c、d、e 中准确选择 1。
答话:
d
问题: 主要根据沸点高低分离物质的方法是
a: 膜分离法
b: 聚酰胺色谱法
c: 硅胶柱色谱法
d: 离子交换树脂法
e: 分馏法
请确保从 a、b、c、d、e 中准确选择 1。
答话:
e
问题: 不属于眼球屈光装置的结构是
a: 角膜
b: 房水
c: 巩膜
d: 玻璃体
e: 晶状体
请确保从 a、b、c、d、e 中准确选择 1。
答话:

F Sample healthcare QA

The following demonstrates the answers from ELAINE-medLLM-instruct for the same questions in English, Japanese, and Chinese.

F.1 QA in English

No.1

Human: At what age does hearing loss begin in adults?
Assistant: The onset of hearing loss in adults typically begins around the age of 60. However, it’s important to note that hearing loss can occur at any age due to various factors such as exposure to loud noises, certain medications, or underlying medical conditions. Regular hearing screenings and maintaining a healthy lifestyle can help prevent or manage hearing loss. Is there anything else you would like to know?

No.2

Human: Lately I have been having constant nausea and heartburn, and after eating I have a tightening abdominal pain (severe) that settles down after about 5-10 minutes each time, but it happens every time and is quite painful. Is it the solar plexus? I feel some discomfort (pressure?) in the stomach area. I also feel discomfort when I lie on my stomach. I feel discomfort when I lie on my stomach. The nausea has been going on for more than 6 months. What could be the cause? Also, what department should I see? Thank you in advance.
Assistant: Thank you for your health concern! I am sorry to hear about your symptoms. I would like to know more about your medical history. Do you have any other medical conditions? Do you take any medications? Have you had any recent surgeries? Do you have any allergies? Do you smoke? Do you drink alcohol? Do you have any family history of cancer?

No.3

Human: I tested positive for corona yesterday. I have asthma-like symptoms, so the doctor told me to take X-rays and CT as well, and as a result, he said, “There is a black shadow, and this could be lung cancer. He said, “I can’t say it’s lung cancer, but it’s possible. I had a pretty bad cough because of the corona, so I hope that’s the cause, but is it possible to have lung cancer complications from corona?
Assistant: I’m sorry to hear about your condition. I understand your concern. I would like to know more about your symptoms. Do you have any other symptoms apart from the cough? Have you had any weight loss? Do you have any other medical conditions? Have you ever smoked?

No.4

Human: I am a 50 year old woman. I have been having frequent headaches for the past half month. The pain varies from time to time. It often feels like a tightness behind my eyes or around my temples, but sometimes it hurts all over my body. I had an MRI scan 2 years ago because of the occasional pain, but it showed no abnormalities. What are the possible causes?
Assistant: Thank you for providing this information. I would like to ask a few more questions to better understand your condition. Do you have any other symptoms such as nausea, vomiting, or sensitivity to light or sound? Have you experienced any changes in your vision or hearing? Do you have any known medical conditions or take any medications regularly?

F.2 QA in Japanese

No.1

Human: 大人の場合、難聴は何歳から始まりますか?
Assistant: 大人の場合、難聴は通常50歳から始まります。ただし、難聴の発症年齢は個人によって異なります。難聴の原因は、遺伝的要因、加齢、耳の感染症、耳の損傷、または特定の薬剤の使用など、さまざまな要因によって異なります。難聴の症状には、音の聞こえにくさ、音のぼやけ、音の歪み、音の知覚の変化、音の知覚の遅れ、音の知覚の困難、音の知覚の喪失が含まれます。難聴の種類には、感音性難聴、伝音性難聴、混合性難聴があります。感音性難聴は、内耳の損傷によって引き起こされ、伝音性難聴は、耳の骨または鼓膜の損傷によって引き起こされ、混合性難聴は、内耳と耳の骨または鼓膜の両方の損傷によって引き起こされます。難聴の治療には、補聴器、人工内耳、聴覚療法、手術などがあります。

No.2

Human: 最近、吐き気と胸焼けが絶えず、食後には締め付けられるような腹痛（激しい）があり、毎回5〜10分ほどで落ち着くのですが、毎回起こるのでかなり辛いです。みぞおちでしょうか? 胃のあたりに何か違和感（圧迫感? また、うつぶせに寝ると違和感を感じる。うつぶせに寝ると違和感がある。吐き気は6ヶ月以上続いています。何が原因でしょうか? また、何科を受診したらよいでしょうか。よろしくお願いします。
Assistant: あなたの症状は胃食道逆流症の可能性があります。医師に相談する前に、食事やライフスタイルの変更を試してみることができます。医師に相談する前に、食事やライフスタイルの変更を試してみることができます。

No.3

Human: 昨日、コロナの陽性反応が出ました。喘息のような症状もあるので、レントゲンとCTも撮るように医師に言われ、その結果、「黒い影があるので、これは肺がんの可能性がある。肺がんとは言えないが、可能性はある」と言われました。コロナのせいで咳がかなりひどかったので、それが原因であればいいのですが、コロナが原因で肺がんを併発することはあるのでしょうか?
Assistant: はい、コロナウイルスが肺がんを引き起こす可能性があります。コロナウイルスは肺に直接影響を与え、肺がんを引き起こすことがあります。さらに、コロナウイルスは免疫システムを弱め、肺がんを引き起こす可能性のある他のウイルスに感染しやすくすることがあります。

No.4

Human: 50歳女性です。ここ半月くらい頭痛が頻繁に起こります。痛み方はそのときによって違います。目の奥やこめかみのあたりがぎゅーっとなる感じが多いですが、全体が痛むときもあります。市販の頭痛薬で治る時が多いですが、飲んでも違和感があるときもあります。2年前くらいにもときどき痛くてMRI検査をしましたが、異常ありませんでしたどんな原因が考えられますか?
Assistant: MRI検査で異常がなかったことから、頭痛の原因はおそらく脳の構造的な問題ではありません。もっと可能性が高いのは、ストレス、睡眠不足、栄養不足、または薬の副作用など、生活習慣に関連するものです。

F.3 QA in Chinese

No.1

Human: 成年人从几岁开始出现听力损失?
Assistant: 根据研究，成年人通常在20岁左右开始出现听力损失的迹象。然而，每个人的听力情况都不同，有些人可能在更早的年龄就开始出现听力损失。因此，如果您有任何听力问题或疑虑，建议您及时咨询专业医生进行评估和治疗。

No.2

Human: 最近，我总是恶心和烧心，吃完饭后腹部会紧缩性疼痛（剧烈），每次大约 5-10 分钟后就会缓解，但每次都会发生，而且相当痛苦。是太阳神经丛吗？我感到胃部有些不适（压力？平躺时也会感到不适。平躺时感到不适。恶心已经持续 6 个多月了。这可能是什么原因造成的？另外，我应该去哪个科室就诊？在此先表示感谢。

Assistant: 您好，根据您的描述，您可能患有胃炎或胃溃疡等胃部疾病。建议您去消化内科就诊，医生会根据您的症状和体检结果进行诊断，并给出相应的治疗方案。同时，您可以注意饮食卫生，避免辛辣，油腻等刺激性食物，多吃易消化的食物，保持良好的生活习惯和心态，有助于缓解症状。祝您早日康复!

No.3

Human: 昨天，我的电晕检测呈阳性。我有类似哮喘的症状，所以医生让我同时拍 X 光片和 CT，结果他说：" 有黑影，这可能是肺癌。他说：" 我不能说这就是肺癌，但有可能。因为电晕，我咳嗽得很厉害，所以我希望是这个原因，但电晕有可能并发肺癌吗?

Assistant: 你好，很高兴为您服务。请问您有多久出现咳嗽，咳痰，胸闷，气短等症状了? 是否有咳血，发热，盗汗，消瘦等症状? 是否有吸烟史? 是否有家族肿瘤史? 是否有过敏史? 是否有其他疾病史?

No.4

Human: 我是一名 50 岁的女性。最近半个月以来，我经常感到头痛。疼痛时有时无。我经常感到眼后和太阳穴周围闷痛，但有时也会感到全身疼痛。由于疼痛时有时无，两年前我做过一次核磁共振扫描，但没有发现异常。可能的原因是什么?

Assistant: 根据您的描述，您可能患有偏头痛。偏头痛是一种常见的神经系统疾病，通常表现为头痛，恶心，呕吐，光过敏和声音过敏等症状。偏头痛的发作通常是周期性的，可能会持续数小时或数天。偏头痛的确切原因尚不清楚，但遗传，荷尔蒙变化，饮食和环境因素等都可能与其发作有关。建议您尽快就医，进行进一步的检查和诊断，以便确定最佳的治疗方案。同时，您可以尝试一些自我管理技巧，如规律作息，避免过度劳累，避免过度使用药物等，以减轻症状。