Anonymous ACL submission

Abstract

1 Introduction

Recently, both closed source LLMs (ope, 2024) and open source communities (tou, 2023b,a) have made great progress and surpassed humans in a range of general areas. However, they have not performed very well in specific professional fields such as medicine, especially for the open source community (Labrak et al., 2024; Han et al., 2023; yan, 2024). This is because the complex and specialized medical domain knowledge is a great challenge to successfully develop an accurate and safe medical LLM (Singhal et al., 2022). We believe that medical LLMs have great application potential and can be valuable in diagnostic assistance, consultation, drug recommendation, etc. As of now, there are some medical LLMs in this field, but these works rely entirely on SFT training (Zhang et al., 2023a, 2024). As we all known, pre-training is a key stage in learning domain knowledge (**??**), and relying only on SFT will cause the model to only give answers in a fixed format. For dataset, most of them only concern on the data construction of the SFT stage (Yang et al., 2023; Zhang et al., 2023a) or pay attention to single-turn dialogues (Li et al., 2023b; **?**; Tian et al., 2023), ignoring the scenarios of multi-turn interactions in real doctor-patient dialogues. In addition, the training datasets are all monolingual and only contain dialogue-type QA data.

To solve the above issues, we propose a bilingual medical LLM based on Aquila[1], namely Aquila-Med, which implements the entire process from continued pre-training, SFT to RLHF. In addition, for continues pre-trained, we build a 20B large-scale Chinese and English medical dataset. A high-quality Chinese and English medical SFT dataset is also constructed, comprising about approximately 330,000 examples, covering 15+ departments and 100+ disease specialties, and we also construct 13,000 high-quality DPO pairs, which include various forms such as QA and medical multiple-choice questions. It is worth noting that we are the first one to open-source the construction process of the two datasets and the entire training process. These three high-quality datasets will also be open-sourced to help more researchers in the open source community.

Specifically, we first collect a large amount of real medical corpus, which It comes from medical data classified from massive pre-training data for Aquila, open source SFT synthetic data, and a certain proportion of general data. We then do a continued pre-training based on the Aquila to obtain a base model with a medical foundation. Secondly, we collect a large amount of open source SFT medical data, and use a variety of data selection methods to filter the quality of single-turn dialogues and multi-turn dialogues respectively. Our high-quality medical SFT dataset includes: single-turn Chinese medical dialogue data, single-turn English medical dialogue data, multi-turn Chinese medical dialogue data, and medical subject knowledge multiple-choice questions, with the aim of enhancing the model's understanding and generalization capabilities in the medical domain. It is worth noting that the dataset here is partly derived from real-world medical diagnosis dialogues and partly from the construction of GPT-3.5. We hope that the model can not only generate informative, clear and logical responses, but also give more professional and personalized consultations like doctors. In the RLHF stage, based on the results of SFT, we used GPT-4 to construct a positive-negative medical data pairs. Finally, we use the Direct Preference Optimization (DPO) (Rafailov et al., 2023) algorithm to align the output of the model with the human expression style.

[1] https://github.com/FlagAI-Open/Aquila2

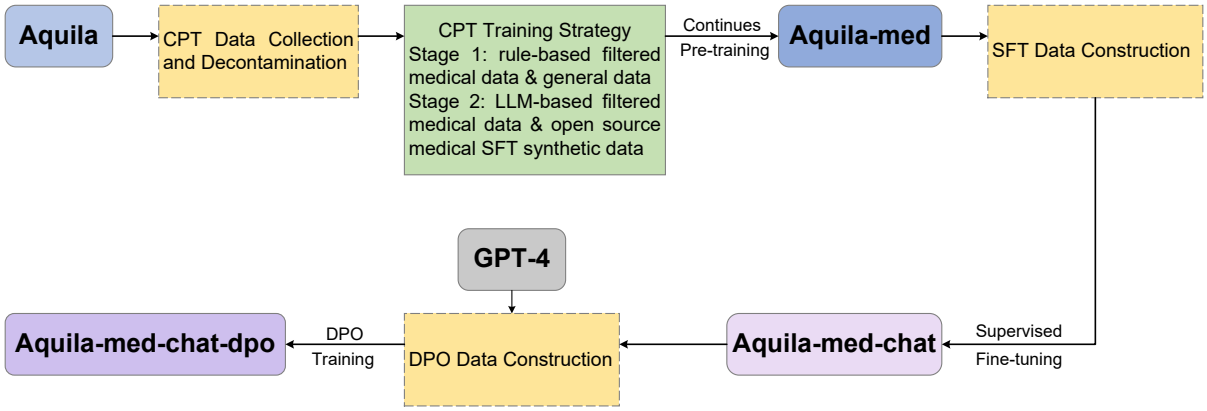


Figure 1: The overall pipline of Aquila-med-chat, which includes the continues pre-training, supervised fine-tuning, and the DPO process.

After extensive training and optimization, we successfully develope Aquila-med. We also comprehensively evaluated common benchmarks in the medical field, covering single-turn dialogue, multi-turn dialogue, and medical multiple-choice questions, involving four capability dimensions. The experimental results show that our model has achieved good results and also prove that our proposed datasets can effectively improve the model's ability to handle single-turn and multi-turn medical consultations.

The main contributions of this paper are as follows: (1) We are the first to implement a full-process from pre-training, SFT to RLHF for a Chinese and English medical LLM, Aquila-med. (2) We are the first to introduce the construction process of three datasets in the medical domain in detail: pre-training, SFT and DPO. We will make all three datasets public. (3) We conduct experiments on multiple Chinese and English benchmarks to verify the effectiveness and reliability of our proposed datasets.

2 Methodology

In this section, we instroduce three stage of model training: continuous pre-training, SFT, and RLHF. The latter two also include the data construction process. Each step is discussed sequentially to mirror the research workflow. The whole process is shown in Figure 1.

2.1 Continuous Pre-training

2.1.1 Data Collection and Decontamination

In this section, we will outline the process of building Aquila-med-cpt from a massive general pre-training database. As shown in Figure 2, we show

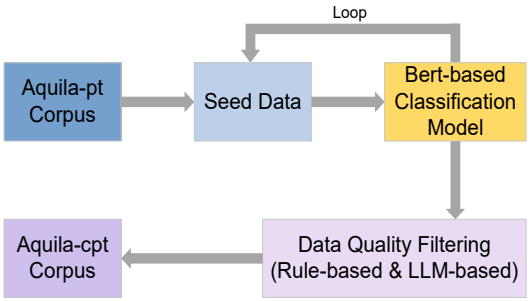


Figure 2: The overall flowchart of construction Aquila-cpt Corpus.

a pipline, including how to collect medical-related corpora from pre-training databases, rule-based quality filtering methods, and LLM-based data quality selection methods. It is worth noting that this method is also applicable to any domain.

Data Classification Since Aquila's general pre-training (Aquila-pt) corpus comes from multiple data sources, it already contains domain information. However, since there is no clear domain label, we first need to classify the data to make full use of the medical domain data in Aquila-pt. Specifically, we first randomly sample 20k data from Aquila-pt and use the upsampling method to ensure that the ratio of Chinese and English being 1:1. Based on the sampled data, GPT-4 is used to perform two rounds of domain label annotation to improve the accuracy of the labels. The data with different labels twice are removed, and finally 17k seed data is retained. Then we design a classifier using the Bert-based multilingual pre-training model. The parameter settings are as follows: batch-size is 64, learning rate is 2e-5, training epoch is 10, and the optimal checkpoint is selected according to the accuracy. The medical domain F1 of the classifier can reach 84%.

Rule-based Data Quality Filtering Since Aquila-pt mostly comes from web pages, the overall quality is not high. In order to remove the noise data, we design a rule-based data filtering solution, including rules for removing data with too few tokens, data with too many special characters, toxic data, and data containing private information.

LLM-based Data Quality Filtering By sampling and checking the data after rule filtering, we found that there exists the following problems: (1) the data contains advertising and marketing information, which will greatly affect the output preference of the trained model; (2) the data contains grammatical errors, semantic incoherence, splicing of multiple unrelated content, image and video editing information, etc. We believe that such data is not beneficial for model training because the model cannot obtain much valuable information through autoregressive learning. Therefore, we design a quality scoring regression model based on LLM to score data quality and further filter out low-quality content. Specifically, we extract 20k data from the rule-based filtered data, score them twice using the GPT4, ranging from 0 to 6, and remove the data with a difference of about 2 points between the two scores, and finally obtained 15k training data. Then we train a scoring model based on the Bert multilingual pre-trained model, using batch-size of 128, learning rate of 3e-4, and train epoch of 10. We set a threshold for high-quality data filtering.

2.1.2 Training Strategy

Our domain pre-training is divided into two stages. The Stage 1 is the training of ordinary quality domain data, and the Stage 2 is the training of high-quality domain data.

Stage 1: The aim is to prevent the model capability from being significantly degraded due to the large difference between pre-training and continue pre-training data. We use medical domian data filtered by rules and general data with a certain ratio. The data amount is about 60B tokens.

Stage 2: The aim is to further improve the capability of the medical domain model. We use medical domian data filtered by LLM quality model and open source medical SFT synthetic data. The data amount is about 20B tokens.

2.1.3 Training Details

Our model is based on Aquila-7B, which a general Chinese-English LLM with 7 billion parameters. It has been pre-trained autoregressively with 3.6T tokens. The vocabulary size of the model is 15k, the model contains 32 layers of transformers, the maximum length is 4096, the hidden layer dimension of each layer of transformer is 4096, the FFN linear layer dimension is 14336, and the GQA structure is used in attention layers, with 8 groups and 32 heads.

For the first stage of continuous pre-training, we train on 3*8 NVIDIA A100-40G GPUs, using a batch-size of 768, a learning rate of 1e-4, a maximum length of 4096, a cosine learning rate scheduler, a warmup-ratio of 0.05, and train for one epoch. For the second stage, keeping other settings unchanged, we reduce batch-size to 384, learning rate to 1e-5, and reduce warmup-ratio to 0.01. We also train for one epoch.

2.2 Supervised Fine-Tuning

To improve the ability of language models to engage in natural conversation, we firstly carry out SFT, which finetunes a pretrained LLM on chat-style data, including both queries and responses. In the following sections, we will delve into the details of data construction and training methods.

2.2.1 Data Construction

Our SFT dataset comprises a variety of question types, including medical exam multiple-choice questions, single-turn disease diagnosis, multi-turn health consultation, etc. It comes from 6 publicly available datasets, namely Chinese Medical Dialogue Data [2], Huatuo26M (Li et al., 2023a), MedDialog (Zeng et al., 2020), ChatMed Consult Dataset (Tian et al., 2023), CMB-exam[3], and ChatDoctor (Li et al., 2023b). These datasets contain not only real doctor-patient dialogues, but also dialogues generated from GPT-3.5. We believe this ensures the diversity of the dataset.

Since a relatively small high-quality dataset has been shown to be sufficient for fine-tuning LLM, we focus on how to automatically filter "good data" from massive data to ensure competitive performance with a minimal amount of data. Similar to common data cleaning operations, we first remove duplicates and data related to security issues such as violence, bias, and pornography. In the following sections, we specifically introduce the data filtering methods.

[2]https://github.com/Toyhom/Chinese-medical-dialogue-data

[3]https://github.com/FreedomIntelligence/CMB

Single-turn Medical Dialogue Data Following Liu et al. (2024); Zeng et al. (2024), we believe that "good data" should have a complex instruction and a high-quality response. Therefore, We adopt the approach from Deita (Liu et al., 2024), which employs a complexity model and a quality model to score each instance along two dimensions: instruction complexity and response quality. The complexity model assigns a complexity score c_i to each instance, while the quality model assigns a quality score q_i, reflecting the quality of the response. By multiplying c_i with q_i, we combine the complexity score and quality score to obtain a comprehensive score, that is, $s_i = c_i * q_i$. Finally, we set a score threshold to select the most effective data instances in the massive data pool.

Multi-turn Medical Dialogue Data For multi-turn dialogues, we first use Deita to calculate the score s_i of each turn separately, and average them to obtain the final score of the entire dialogue. However, we found that there are two special problems in multi-turn dialogues compared to single-turn dialogues: (1) The correlation between different turn is very low, resulting in a negative impact of the previous information on the following; (2) The correlation between different turns is too high, resulting in a large degree of context duplication and information is redundant. Therefore, we propose a Context Relevance (CR) score, which is a metric that relies on cross-entropy loss to evaluate the impact of historical information on each turn. The details are as follows:

In the instruction-tuning process, the loss of a sample pair (H, T) is calculated by continuously predicting the next tokens in the current turn T given their previous tokens and the history information H:

$$L_\theta(t_i|H) = -\frac{1}{N}\sum_{i=1}^{N} logP(w_i^j|H, w_i^1, w_i^2, ..., w_i^{j-1}; \theta) \quad (1)$$

where $H = \{t_1, t_2, ...t_{i-1}\}$, t_i is the current turn, w_i^j is the j-th token in the i-th turn, and N is the number of tokens of the current turn. We define $L_\theta(t_i|H)$ as the Conditioned Information Score, which measures the ability to generate the current turn under the guidance of corresponding historical information.

To measure the ability of LLM to generate this turn alone, we also define a Direct Information Score:

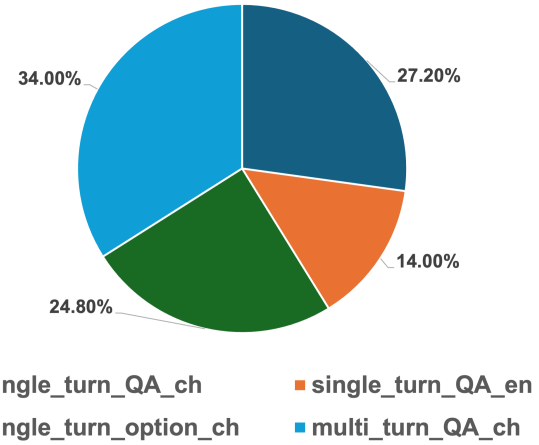


Figure 3: Statistics on the distribution of our proposed SFT dataset.

$$L_\theta(t_i) = -\frac{1}{N}\sum_{i=1}^{N} logP(w_i^j|w_i^1, w_i^2, ..., w_i^{j-1}; \theta) \quad (2)$$

We believe that the higher Direct Information Score may indicate that the turn is more challenging or complex. Finally, we try to estimate the CR score by calculating the ratio between $L_\theta(t_i)$ and $L_\theta(t_i|H)$.

$$CR_\theta(H, T) = \frac{L_\theta(t_i|H)}{L_\theta(t_i)} \quad (3)$$

Here, if $r > 1$, it means that historical information has a negative impact on current turn, that is, the correlation between contexts is very low. If $r < 1$, it means that historical information has a positive impact on current turn, that is, the correlation between contexts is high. However, too small r means that the context is highly repeated and the information is highly redundant. We also set a threshold to filter the data.

2.2.2 Training Details

Our model is based on Aquila3-7B, which a general Chinese-English LLM with 7 billion parameters. The training process has the following hyperparameters: sequence length set to 2048, batch size set to 128, and peak learning rate set to 2e-6 with cosine learning rate scheduler. To prevent overfitting, weight decay of 0.1 is applied and dropout is set to 0.1. Training is parallelized on 8 A100-40G INVIDIA GPUs using the AdamW optimizer with bf16 precision and ZeRO-3. We reserve 10% of the training set for validation and get the best checkpoint after 2 epochs.

2.2.3 Dataset Statistics

Through the above data filtering methods, we select s 320,000 high-quality SFT medical dataset from 199,000 instances, in which the ratio of Chinese and English is 86%:14%. As shown in Figure 3, it comes from single-turn Chinese medical dialogues (single_turn_QA_ch), single-turn English medical dialogues (single_turn_QA_en), multi-turn Chinese medical dialogues (multi_turn_QA_ch), and medical subject knowledge multiple-choice questions (single_turn_option_ch).

2.3 RLHF

We enhance the model's capabilities using Direct Preference Optimization (DPO) (Rafailov et al., 2023) after the SFT stage. To align the model's output with human preferences while preserving the foundational abilities gained during the Continuous Pre-training and SFT stages (Lu et al., 2024), we construct subjective preference data and objective preference data. We also provide the training details of the DPO stage.

2.3.1 Data Construction

We construct the preference pair for the DPO stage using samples that have the same distribution as the SFT dataset. This mainly includes the following two preferences.

Subjective Preference Data We aim to construct dpo pairs where the chosen response aligns closely with human preferences. For each prompt, we first ask GPT-4 to respond as a professional and helpful doctor. Then, using GPT-4, we evaluate the superiority or inferiority of the original response and this newly generated response from the prompt. The evaluation considers four aspects: Fluency, Relevance, Completeness, and Proficiency in Medicine (Zhang et al., 2023b). We select the superior response as the chosen response for the dpo pair and the inferior response as the rejection response.

Objective Preference Data While RLHF can guide LLMs to align with human expectations, numerous studies show that this method can cause LLMs to forget abilities acquired during pretraining and SFT stages (Bai et al., 2022; Dong et al., 2023a), leading to an "alignment tax" (Dong et al., 2023b; Sun et al., 2024). To mitigate this issue, we construct objective preference data. Specifically, for objective prompts with known ground truth answers, we consider the ground truth as the chosen response and randomly select incorrect answers from the remaining options as rejection responses. For instance, in multiple-choice questions, if the ground truth is option A, we randomly select from options B, C, and D to construct the rejection response.

2.3.2 Training Details

We constructed a dataset of 12,727 DPO preference pairs, consisting of 9,019 subjective and 3,708 objective data samples. We trained the model over two epochs using 8 NVIDIA Tesla A100 GPUs. The settings included a learning rate of 2e-7, a batch size of 64, and a beta of 0.03. Additionally, we employed a learning rate warmup and a cosine learning rate scheduler for optimization.

3 Evaluation

We evaluate our model's performance on several open-source Chinese and English benchmarks related to the medical domain. These benchmarks assess the model's ability to comprehend medical knowledge and engage in both single-turn and multi-turn conversations on medical topics.

3.1 Medical Knowledge Benchmark

We extract medical-related questions from the MMLU (Hendrycks et al., 2020) and C-Eval (Huang et al., 2024) benchmarks, and we utilize questions from the CMB-Exam (Wang et al., 2023), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022) and PubMedQA (Jin et al., 2019) test set to evaluate the model's proficiency in medical knowledge.

MMLU is the english multi-subject multiple-choice dataset, from which we extract medical-related tasks to evaluate the model's performance. These tasks encompass various medical domains, including anatomy, clinical knowledge, college biology, college medicine, medical genetics, and professional medicine.

C-Eval is a chinese multiple-choice dataset. We extracted tasks related to medicine from the validation set, such as basic medicine, clinical medicine, medical practice, and veterinary medicine to test the model's performance.

CMB-Exam is a collection of multiple-choice questions in Chinese, sourced from various professional mdedical qualification examinations. It encompasses questions from exams for physicians, nurses, technicians, pharmacists, undergraduate medical programs, and graduate entrance examinations. We utilize 11,200 questions from the test

set to conduct a comprehensive, multi-level assessment of the model's medical knowledge.

MedQA is a multiple-choice question dataset from the United States Medical Licensing Examination (USMLE). Its test set consists of 1,273 questions, which are used to assess a model's medical knowledge and reasoning skills required to obtain a medical license in the United States.

MedMCQA is a large-scale multiple-choice question and answer dataset, sourced from India's medical entrance exams (AIIMS/NEET). Its test set comprises 6,100 questions, enabling the evaluation of a model's general medical knowledge and reasoning abilities.

PubMedQA is a closed-domain question and answer dataset, where each question can be answered by referring to the relevant context from PubMed abstracts. We use 500 test questions from this dataset to evaluate a model's ability to understand and reason about biomedical literature.

3.2 Medical Dialogue Benchmark

We evaluate the model's capability to solve realistic patient problems by assessing its medical knowledge and complex reasoning abilities. This evaluation covers single-round dialogue scenarios, such as the Huatuo MedicalQA (Li et al., 2023a), as well as multi-round dialogue scenarios like CMtMedQA (Yang et al., 2023) and CMB-Clin (Wang et al., 2023).

Huatuo MedicalQA is a large-scale Chinese Medical Question Answering (QA) dataset, and we use its test set to evaluate the model's capability in single-round dialogues. Specifically, we sample 500 question-answer pairs from the test set and employ GPT-4 to compare the model's predicted answers with other reference answers (mainly including the ground truth answer from the dataset and the answer generated by GPT-3.5). Inspired by Zhang et al. (2023b), we use the prompt in Table 3 to judge the quality of the answers. Considering that GPT-4 may exhibit a "position bias" when judging (Zheng et al., 2024), we swap the order of the predicted answer and the reference answer. We determine a answer as winning or losing only when the judgment results are completely consistent before and after the swap.

CMtMedQA is a large-scale dataset consisting of multi-turn medical dialogues in Chinese. To evaluate the model's ability to engage in complex dialogues and initiate proactive inquiries, we utilized approximately 1,000 samples from the dataset's test set.

Model	MMLU	C-Eval	MedQA	MedMCQA	PubMedQA
Aquila	42.91	48.77	38.65	38.58	71.60
Aquila-med	49.32	48.40	41.56	38.23	72.40

Table 1: Performance on various medical knowledge benchmarks for continues pre-training. Specifically, MMLU and C-Eval represent the average scores obtained by the model on the medical-related sub-tasks within these benchmarks. Here, our setting is 3-shot.

CMB-Clin consists of 74 expertly curated medical case consultations derived from clinical diagnostic teaching materials. It evaluates the model's mastery and reasoning abilities in applying medical knowledge through multi-round diagnostic dialogues.

For multi-round dialogue datasets such as CMtMedQA and CMB-Clin, inspired by Wang et al. (2023), we employed GPT-4 to evaluate the model's responses in each round of the dialogue. The evaluation focused on four key aspects: fluency, relevance, completeness, and proficiency in medical knowledge. The specific evaluation prompt used is displayed in Table 4.

4 Experimental Results

4.1 Results for Continue Pre-training

Table 1 shows the results of our continue pre-training on five benchmarks. It can be observed that Aquila-med has improved to a certain extent compared with Aquila, especially on MMLU. This shows that even if the model uses the data which has been already learned in the pre-training stage, the professional ability of the model can be further improved by improving the quality and professional density. In general, we obtain a basic model with medical domain knowledge.

4.2 Results for Alignment

For instruct-tuning, we evaluate it from two aspects: medical subject questions and doctor-patient consultation. Table 2 shows the results on three medical knowledge benchmarks. We found that Aquila-med-chat has good command following ability, and Aquila-med-chat-dpo has made further progress, especially C-Eval. Figures 4 and 5 show the comparison of the outputs of our models with the reference and GPT-3.5 outputs in single-turn dialogues. It is observed that both Aquila-med-chat and Aquila-med-chat-dpo have achieved good results, especially Aquila-med-chat-dpo has achieved

	MMLU	C-Eval	CMB-Exam
Aquila-med-chat	56.2	50.44	47.63
Aquila-med-chat-dpo	56.4	53.10	47.12

Table 2: Performance on various medical knowledge benchmarks for supervised fine-tuning. Specifically, MMLU and C-Eval represent the average scores obtained by the model on the medical-related sub-tasks within these benchmarks.

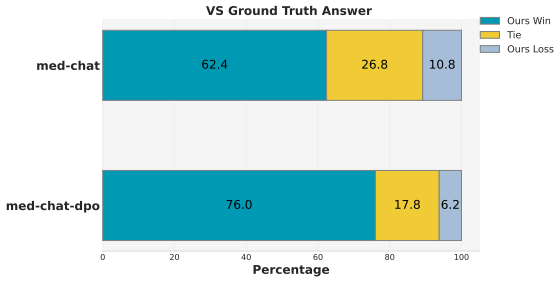


Figure 4: The comparison of our model's predicted answers and the ground truth answers from the dataset on single-round dialogues from the Huatuo MedicalQA.

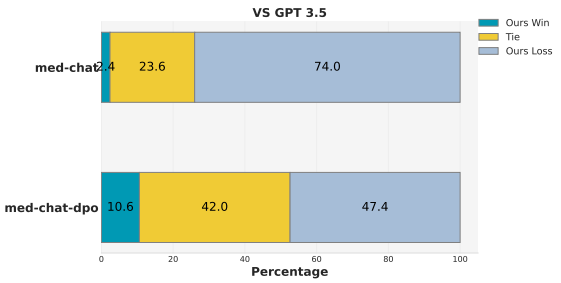


Figure 5: The comparison of our model's predicted answers and the gpt-3.5 predicted answers on single-round dialogues from the Huatuo MedicalQA.

human-style alignment. For multi-turn dialogues, we use GPT-4 to score each turn in four dimensions, and the results are shown in Tables 6 and Table 7. The evaluation results indicate that Aquila-med-chat performed well in terms of generating fluent responses. Additionally, it was observed that Aquila-med-chat-dpo significantly enhanced the model's performance in terms of relevance, completeness, and proficiency, while still maintaining a high level of fluency in the generated responses.

References

2023a. Llama 2: Open foundation and fine-tuned chat models.

2023b. Llama: Open and efficient foundation language models.

2024. Advancing multimodal medical capabilities of gemini.

2024. Gpt-4 technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023a. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023b. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

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.

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

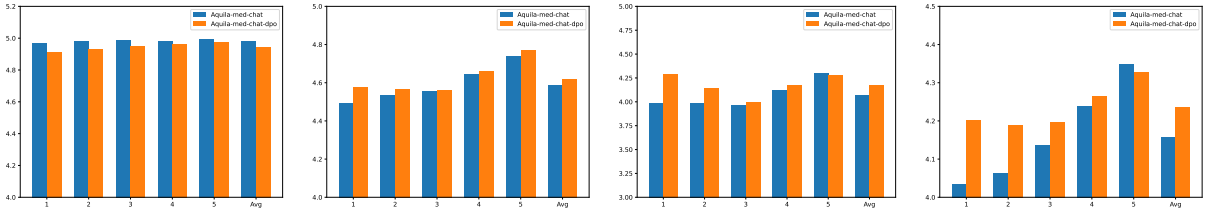
Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

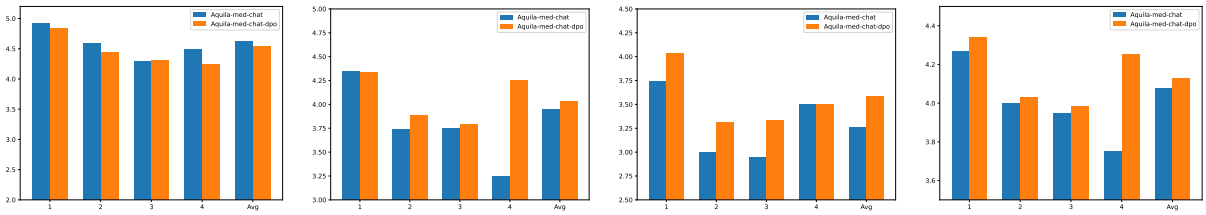
Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset.



((a)) Fluency Score ((b)) Relevance Score ((c)) Completeness Score ((d)) Proficiency Score

Figure 6: Performance on the CMTMedQA dataset in multi-round dialogues. The x-axis represents different rounds of the dialogue, while the "Avg" data point displays the average score across all rounds.



((a)) Fluency Score ((b)) Relevance Score ((c)) Completeness Score ((d)) Proficiency Score

Figure 7: Performance on the CMT-Clin dataset in multi-round dialogues. The x-axis represents different rounds of the dialogue, while the "Avg" data point displays the average score across all rounds.

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

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. 2024. Online merging optimizers for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences.

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

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.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic instruction

evolving for large language models. *arXiv preprint arXiv:2406.00770*.

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. 2023a. 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.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023b. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024. Ultramedical: Building specialized generalists in biomedicine.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

[User]
{user_query}
[End of User]
[Assistant 1]
{assistant1}
[End of Assistant 1]
[Assistant 2]
{assistant2}
[End of Assistant 2]
[System]
We would like to request your feedback on two multi-turn conversations between the AI assistant and the user displayed above. Requirements: Focus on the AI's response in the conversation. The AI assistant should act like the doctor using the tone, manner, and vocabulary the human doctor would use. It should be to the point, without unnecessary elaboration or extraneous information. The AI assistant should respond appropriately to the user in a manner that helps to progress the conversation. The description of symptoms should be comprehensive and accurate, and the provided diagnosis should be the most reasonable inference based on all relevant factors and possibilities. The treatment recommendations should be effective and reliable, taking into account the severity or stages of the illness. The prescriptions should be effective and reliable, considering indications, contraindications, and dosages. Please compare the performance of the AI assistant in each conversation. You should tell me whether Assistant 1 is 'better than', 'worse than', or 'equal to' Assistant 2. Please first compare their responses and analyze which one is more in line with the given requirements.

In the last line, please output a single line containing only a single label selecting from 'Assistant 1 is better than Assistant 2', 'Assistant 1 is worse than Assistant 2', and 'Assistant 1 is equal to Assistant 2'.

Table 3: Prompt for judging the quality of a single-round dialogue

You are an AI evaluator specializing in assessing the quality of answers provided by other language models . Your primary goal is to rate the answers based on their fluency , relevance , completeness , proficiency in medicine . Use the following scales to evaluate each criterion :
Fluency :
1: Completely broken and unreadable sentence pieces
2: Mostly broken with few readable tokens
3: Moderately fluent but with limited vocabulary
4: Mostly coherent in expressing complex subjects
5: Human - level fluency
Relevance :
1: Completely unrelated to the question
2: Some relation to the question , but mostly off - topic
3: Relevant , but lacking focus or key details
4: Highly relevant , addressing the main aspects of the question
5: Directly relevant and precisely targeted to the question
Completeness :
1: Extremely incomplete
2: Almost incomplete with limited information
3: Moderate completeness with some information
4: Mostly complete with most of the information displayed
5: Fully complete with all information presented
Proficiency in medicine :
1: Using plain languages with no medical terminology .
2: Equipped with some medical knowledge but lacking in - depth details
3: Conveying moderately complex medical information with clarity
4: Showing solid grasp of medical terminology but having some minor mistakes in detail
5: Fully correct in all presented medical knowledge
You will be provided with the following information :
- a conversation
- a question based on the conversation
- the solution to the question
- a model ' s answer to the question
[conversation]
{history}
[end of conversation]
[question]
{question}
[end of question]
[solution]
{solution}
[end of solution]
[answer]
{answer}
[end of answer]
Make sure to provide your evaluation results in JSON format and ONLY the JSON , with separate ratings for each of the mentioned criteria as in the following example :
{"fluency": 3, "relevance": 3, "completeness": 3, "proficiency": 3}

Table 4: Prompt for judging the quality of a multi-round dialogue