

FROM BEGINNER TO EXPERT: MODELING MEDICAL KNOWLEDGE INTO GENERAL LLMs

Qiang Li[†], Xiaoyan Yang[†], Haowen Wang, Qin Wang, Junjie Wang, Yang Zhang, Mingyuan Chu, Sen Hu, Yicheng Chen, Yue Shen, Cong Fan, Wangshu Zhang, Teng Xu, Jinjie Gu, Jing Zheng, Guannan Zhang
Ant Group
{mangxiao.lq, joyce.yxy, zhanying}@antgroup.com

Lei Liu^{‡,△}
Ant Group, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ)
{liulei1497}@gmail.com

ABSTRACT

Recently, large language model (LLM) based artificial intelligence (AI) systems have demonstrated remarkable capabilities in natural language understanding and generation. However, these models face a significant challenge when it comes to sensitive applications, such as reasoning over medical knowledge and answering medical questions in a physician-like manner. Prior studies attempted to overcome this challenge by increasing the model size (>100B) to learn more general medical knowledge, while there is still room for improvement in LLMs with smaller-scale model sizes (<100B). In this work, we start from a pre-trained general LLM model (AntGLM-10B) and fine-tune it from a medical beginner towards a medical expert (called AntGLM-Med-10B), which leverages a 3-stage optimization procedure, *i.e.*, general medical knowledge injection, medical domain instruction tuning, and specific medical task adaptation. Our contributions are threefold: (1) We specifically investigate how to adapt a pre-trained general LLM in medical domain, especially for a specific medical task. (2) We collect and construct large-scale medical datasets for each stage of the optimization process. These datasets encompass various data types and tasks, such as question-answering, medical reasoning, multi-choice questions, and medical conversations. (3) Specifically for multi-choice questions in the medical domain, we propose a novel Verification-of-Choice approach for prompting engineering, which significantly enhances the reasoning ability of LLMs. Remarkably, by combining the above approaches, our AntGLM-Med-10B model can outperform the most of LLMs on PubMedQA, including both general and medical LLMs, even when these LLMs have larger model size.

1 INTRODUCTION

The ability of large language models (LLMs) is truly remarkable to understand and generate text in various fields like natural language, computer code, and protein sequences. These LLMs leverage the transformer architecture (Vaswani et al., 2017), which is specifically designed for sequence modeling and trained through self-supervision (Kenton & Toutanova, 2019). By increasing the model size, dataset size, and training computation, the performance on various benchmarks are consistently improved (Liang et al., 2022). These empirical findings are in line with a theoretical analysis (Kaplan et al., 2020), which highlights the significance of scale in ensuring the reliability of inferences made by LLMs.

It is a long-standing research topic for AI in medicine to develop LLMs for solving medical problems, where accurate assessment of medical knowledge and reasoning capabilities is crucial for

[†]These authors contributed equally to this work.

[‡]Corresponding authors.

[△]Work was done during Lei Liu’s research internship in Ant Group.

informed decision-making and favorable patient outcomes. Currently, LLMs for applications in medicine usually fail to fully utilize medical domain data, due to lacking general and specific clinical knowledge (Yim et al., 2020). As indicated by (Singhal et al., 2022), there is a discordance between what AI models can do and what may be expected of them in real-world clinical workflows (Lakkaraju et al., 2022; Schaekermann et al., 2020).

There are two kinds of exploration paths to investigate the adaptation of LLMs in the medical domain. One approach is to conduct a thorough assessment of general language models like GPT-3.5 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and ChatGPT (OpenAI, 2022), without any specific fine-tuning for medical clinical issues. These models are designed for general purposes and are not specialized for medical domain. In (Nori et al., 2023), researchers evaluated performance of GPT-4 model with its predecessors in the GPT family on medical problems. Another way is to fine-tune a specialized LLM model through training or engineered to solve medical clinical tasks. For example, Singhal et al. (2023) developed a new medical LLM called Med-PaLM 2 and targeted medical domain-specific fine-tuning, which is based on a new base model (PaLM 2 Anil et al. (2023)). To evaluate how effectively LLMs encodes clinical medical knowledge, previous works (Jin et al., 2019; Singhal et al., 2023) generally considered the medical question answering task, which requires deep understanding on medical context, accurately recalling relevant medical knowledge, and reasoning with expert-level experiences.

Nevertheless, existing medical LLMs are mainly based on scaling law (Chung et al., 2022) to train a larger model over massive data, which indeed lacks a fundamental optimization paradigm to adapt a pre-trained general language model towards a medical-specific expert. We conjecture that it is mainly due to the intrinsic LLM training paradigm, *i.e.*, large-scale pre-training followed by specific fine-tuning. Concretely, as a medical beginner, a pre-trained general LLM needs to learn basic medical knowledge as background, which requires a continual pre-training over the large-scale medical data. Then, considering diverse task types on medical domain (*e.g.*, QA, Multi-choice Question, and Reasoning), a relatively large-scale instruction fine-tuning process should be applied to encode task-related knowledge into LLMs. Finally, given a medical problem with a specific task type, a careful fine-tuning step can help to quickly and accurately adjust a LLM as a medical expert.

In this study, we present a comprehensive methodology for fine-tuning a pre-trained general LLM model to transform it from a medical beginner to a medical expert. As shown in Figure 1, this process involves a 3-stage optimization procedure, namely continual pre-training for medical knowledge injection, medical domain instruction tuning, and specific medical task adaptation. To support each stage of fine-tuning, we curate and construct diverse large-scale medical datasets that encompass various data types and cover different tasks. These tasks include question-answering (QA), medical reasoning, multi-choice question, and medical conversations. Additionally, for the multi-choice question task within the medical domain, we introduce a novel Verification-of-Choice approach for prompting engineering. This approach significantly enhances the reasoning ability of LLMs, offering a valuable contribution to the field. By incorporating the afore-mentioned components, the obtained AntGLM-Med-10B can achieve an impressive accuracy on the PubMedQA. Notably, AntGLM-Med-10B can outperform many larger LLMs (>40B), demonstrating its effectiveness and potential in the medical domain.

The organization of this paper is as follows:

- (1) In Section 2, we provide a overview for the recent progress of LLMs and discuss some applications of LLMs in the medical domain.
- (2) In Section 3, we illustrate the framework preliminaries in detail, including the 3-stage optimization (continual pre-training, instruction fine-tuning, and specific-task adaptation), dataset collection and construction, and the utilized techniques.
- (3) In Section 4, we provide comprehensive experiments to indicate the effectiveness of 3-stage optimization, as well as the well analysis for each tuning strategies. The performance on PubMedQA is significantly improved.

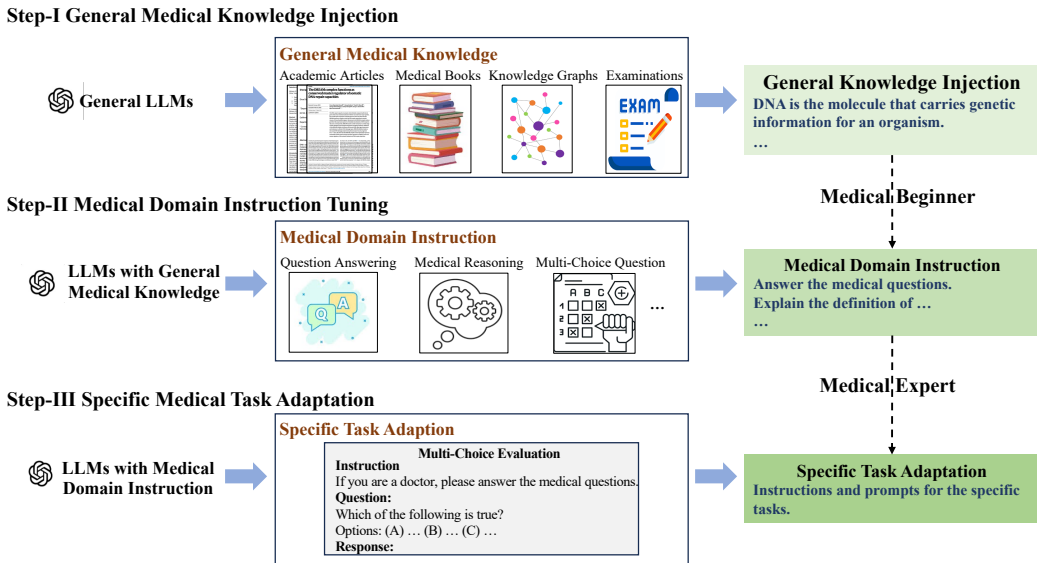


Figure 1: The 3-stage optimization procedure of AntGLM-Med-10B. We collect different data types (e.g., medical articles, books, and examinations) and medical tasks (question answering, reasoning, and multi-choice questions). In the adaptation stage, we mainly consider the optimizations for multi-choice questions, resulting in competitive performance on the PubMedQA.

2 RELATED WORK

Large Language Models. LLMs have achieved remarkable performance in various natural language processing (NLP) tasks (Brown et al., 2020; Wei et al., 2022; Liu et al., 2023), in recent years. These developments benefit from scaling up the training of both model size (typically for transformer-based models) and data scale, where scaling law Chung et al. (2022) indicates the relationship between model performance and model scale and dataset size. For example, when trained on extensive text corpora like Wikipedia and BooksCorpus, LLMs could show promising performance across various NLP tasks, including tasks requiring specialized domain knowledge and reasoning. In detail, GPT-3 (Brown et al., 2020) stands out as the pioneering language model with impressive 100 billion parameters, which showcases remarkable few-shot learning capabilities and introduces the concept of in-context learning. Following its success, a plethora of other LLMs have been proposed, such as Megatron-LM (Korthikanti et al., 2023), OPT (Zhang et al., 2022), Chinchilla (Hoffmann et al., 2022), Galactica (Taylor et al., 2022), LLaMA (Touvron et al.), PaLM (Chowdhery et al., 2022), and PaLM-2 (Singhal et al., 2023). These LLM models have further enhanced language understanding, generation, instruction following, reasoning abilities, and even possess a deep understanding of common sense knowledge (Mao et al., 2023). As a result, they have become indispensable base models across various domains, including Finance (Yang et al., 2023), Education (Milano et al., 2023), and Healthcare Arora & Arora (2023).

Medical LLMs. These advanced improvements for LLMs have also demonstrated the effectiveness on medical domain, such as HuatuoGPT (Zhang et al., 2023), Med-PaLM 2 (Singhal et al., 2023), and Visual Med-Alpaca (Gao et al., 2023). In particular, HuatuoGPT (Zhang et al., 2023) presented to actively ask questions for the patients rather than only make a respond. Visual Med-Alpaca (Gao et al., 2023) integrated visual experts with LLMs for multi-modal biomedical tasks, which can perform better over various tasks. Although these approaches utilized scientific and biomedical corpora for both discriminative and generative language modeling, they are typically small in model size compared with LLMs GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022).

Table 1: The collected datasets for different optimization stages.

Optimization	Dataset	Description
General Medical Knowledge Injection	Medical Books	General medical knowledge in medical and science textbooks
	Knowledge Graphs	Highly structured medical knowledge in open-source knowledge graphs
	Question-Answer Pairs	Real-world medical consultation information in textual form
	Exam Questions	Question, answer, and explanation for testing medical knowledge points
	Articles	Professional medical and science articles written by different doctors
Medical Domain Instruction Tuning	PromptCBLUE	A instruction-tuning dataset for multi-task and few-shot learning in Chinese
	Chinese Examination	A dataset collected from the Chinese physician examination data
	Wuma QA	A large-scale QA database covering encyclopedia, hospital, and doctor
	Huatu Wiki	A subset (the data source is Chinese Wikipedia) of Huatuo-26M
	Multiple-choice Question	Multiple-choice questions from different datasets
Specific Medical Task Adaptation	Multiple-choice Question	Multiple-choice questions from different datasets

3 FROM BEGINNER TO EXPERT

3.1 FRAMEWORK FORMULATION

In this work, our objective is to teach a medical beginner (*i.e.*, a pre-trained general LLM) and let it become a medical expert (*i.e.*, a fine-tuned medical LLM), which corresponds to a full procedure for adapting a pre-trained foundational large language model in the medical domain. The optimization process can be divided into three key steps: General Medical Knowledge Injection, Medical Domain Instruction Tuning, and Specific Medical Task Adaptation.

As shown in Figure 1, **General Medical Knowledge Injection** aims to encode the fundamental medical knowledge into the pre-trained language model. **Medical Domain Instruction Tuning** can enrich the language model with diverse medical task types. **Specific Medical Task Adaptation** can tailor the model to align with a specific clinical task.

3.2 BASE LLM: ANTGLM

The base LLM in this work is AntGLM, a general large-scale model developed by Ant Group. We conduct further pre-training and fine-tuning for its adaptations on the medical domain.

Architecture. AntGLM is based on the GLM architecture (Du et al., 2022), which combines the ideas of auto-encoding and auto-regression to enhance the learning process. AntGLM is featured by 48 transformer layers, a hidden size of 4096, and 64 attention heads, resulting in 10B parameters. AntGLM incorporates two-dimensional positional encoding and enables the pre-training task of predicting the order of blank regions, which can significantly improve the performance of blank filling during pre-training in a flexible manner. Overall, AntGLM performs well on various natural language processing tasks.

3.3 DATASETS FOR GENERAL MEDICAL KNOWLEDGE INJECTION

As introduced below, several authoritative Chinese and English datasets are collected for the stage of General Medical Knowledge Injection, which mainly include PromptCBLUE (Zhu et al., 2023), MedPaLM2 (Singhal et al., 2023) dataset, and 3 sets of Chinese examination datasets.

Medical Books. In particular, we collect some medical books as training corpora, which include medical textbooks and popular science books. The high-quality and lengthy text in the book data makes LLMs easier to learn semantic correlations among contexts. We manually remove the contents in non-text format from the books, such as tables and appendices, which may influence the model training. After data pre-processing, we used approximately 150,000 medical-related books during pre-training.

Knowledge Graphs. We use an open-source medical knowledge graph to generate pre-training data, *i.e.*, Open Medical and Healthcare Alliance (Omaha). The data in the graph is highly structured and typically stored in the form of triplets. Since it is difficult for pre-training models to directly learn

Table 4: Chinese Examination Datasets.

Chinese Examination	Size
Clinical practitioner	600
Veterinary general practice	400
Licensed pharmacist of Western medicine	480

from triplet information, we sample sub-graphs of different scales and use GPT-3.5-turbo (Brown et al., 2020; Ouyang et al., 2022) to rewrite them into natural language texts.

Table 2: The data statistics involved in general medical knowledge injection.

Data Type	Token Size
Medical Books	7.18B
Knowledge Graphs	0.15B
QA Pairs	0.24B
Exam Questions	0.11B
Articles	7.71B

However, due to specific formatting requirements such as multiple-choice and fill-in-the-blank questions, pre-training models cannot directly consume this data. We also use GPT-3.5-turbo (Brown et al., 2020; Ouyang et al., 2022) to rewrite the exam question data and convert it into medical knowledge points.

Articles. Article data consists of cutting-edge medical papers and popular science articles, which are used to enrich the latest medical knowledge to the dataset. We use PubMed Canese & Weis (2013) as the source of the paper data, which is a search engine for medical papers and includes over 25 million articles. PubMed contains both full-text and abstract data. Due to the presence of special text, such as tables in the full-text, which is difficult to process and may interfere with the model’s performance, we only use PubMed abstracts as our training data, excluding the abstracts of the expert-level labeled PQA-L (see in Table 5) in PubMedQA dataset for avoiding annotation leakage. For medical popular science articles, we purchase professional medical articles written by different doctors from websites as training corpora.

PromptCBLUE. PromptCBLUE (Zhu et al., 2023) is a dataset for Chinese medical evaluation jointly constructed by Alibaba Cloud Tianchi, East China Normal University, and other well-known enterprises and academic research institutions, with rankings being set.

3.3.1 DATASET FOR MEDICAL DOMAIN INSTRUCTION TUNING

Table 3: The data statistics involved in medical domain instruction tuning.

Data Type	Size
PromptCBLUE	68,500
Chinese Examination	151,568
Wuma QA	17,604
Huatuo Wiki	200,000
Multiple-choice Question	194,455

source of Huatuo Wiki is from the online medical encyclopedia.

Question-Answer Pairs. We collect publicly available real-world medical question-answer pairs to enrich the training, which is easily accessible and rich medical consultation information in textual form. Therefore, we directly concatenate the question and answer data and use it as training corpora.

Exam Questions. Exam questions usually consist of the question, answer, and explanation. These exam questions are designed to assess medical knowledge and require logical reasoning based on the provided knowledge. Therefore, its data quality is very high, making it an excellent source for pre-training corpora.

Chinese Examination Datasets. We collect three examination datasets from the Chinese physician examination data, including clinical practitioner, veterinary general practice, and licensed pharmacist of Western medicine. The overall dataset contains about 150k data. Table 4 illustrates the distribution of the test set, involving more than 1000 records.

Wuma QA. We additionally collect 17,600 medical question-answer pairs from the realistic environment.

Huatuo Wiki. Huatuo-26M (Li et al., 2023) is a large-scale Chinese medical QA dataset. Huatuo Wiki is a subset of Huatuo-26M (Li et al., 2023). The data

Medical Multiple-choice Question Datasets. Following (Singhal et al., 2023), as shown in Table 1, we utilize the MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019) and MMLU clinical topics (Hendrycks et al., 2020) datasets.

3.3.2 DATASET FOR SPECIFIC MEDICAL TASK ADAPTATION

Table 5: The detailed information of the PubMedQA dataset.

Statistic	PQA-L	PQA-U	PQA-A
Number of QA pairs	1.0k	61.2k	211.3k
Prop. of yes (%)	55.2	-	92.8
Prop. of no (%)	33.8	-	7.2
Prop. of maybe (%)	11.0	-	0.0
Avg. question length	14.4	15.0	16.3
Avg. context length	238.9	237.3	238.0
Avg. long answer length	43.2	45.9	41.0

PubMedQA. As shown in Table 5, the PubMedQA dataset (Jin et al., 2019) consists of 1,000 question answer pairs annotated by the medical experts and more than 60k unlabeled questions. Given a question with a PubMed abstract as the context, the task is to generate a true choice from the multiple-choice answers (*i.e.*, yes/no/maybe). Besides, PubMedQA requires answer inference from the supporting PubMed abstract context, which is a closed domain question answering task.

3.4 MEDICAL KNOWLEDGE MODELING

To take full advantage of medical data, we provide a detailed description for the training tricks for each step of 3-stage optimization. The technique tricks for different steps are shown in Figure 2.

3.4.1 CONTINUAL PRE-TRAINING FOR MEDICAL KNOWLEDGE INJECTION

Following Du et al. (2022), continual pre-training for AntGLM-Med-10B is to optimize an autoregressive blank-filling objective. Concretely, given an input text $x = [x_1, x_2, \dots, x_n]$, multiple spans $\{s_1, \dots, s_n\}$ are sampled from it as PART B. Then PART B is replaced by the special [MASK] token, resulting in a corrupted sequence $x_{corrupt}$ as PART A. Tokens in Part A can attach each other in a pair-wise way, while tokens in Part B can only attach Part A and the preceding parts in Part B. GLM randomly shuffles the spans to improve the model’s semantic understanding for capturing the relationships among different spans. Two-dimensional positional encoding is utilized to represent the positional relationships within each span and among different spans. For autoregressive generation, each span is padded with special tokens [START] and [END] for input and output, respectively. This approach allows to automatically train both a bidirectional encoder and a unidirectional decoder within a unified architecture. By controlling the span length and granularity, the model can effectively handle both natural language understanding and generation tasks. Due to the limitations of the scale and capability of the general LLM models, we further pre-train the model over the medical data, significantly improving the performance in the medical field.

Pre-Training Implementations. The training settings are basically consistent with GLM, using Adam as the optimizer with the following hyper-parameters: $\beta_1 = 0.9$, $\beta_2 = 0.95$. We employ a cosine learning rate schedule, where the final learning rate is equal to 10% of the maximum learning rate. Additionally, we apply a weight decay of 0.1 and gradient clipping of 1.0. The batch size is set to 256. Since we continue pretraining based on a general base, the learning rate is set to $7e-6$. When training a model with a parameter count of 10B, our code achieves a processing speed of approximately 2000 tokens/gpu/sec on 32 A100 GPUs with 80GB RAM.

3.4.2 FINE-TUNING FOR MEDICAL DOMAIN INSTRUCTION

Instruction Fine-tuning. We applied instruction fine-tuning for the base LLM following the protocol used by Chung et al. (2022). For this process, we use a cosine learning rate schedule with an initial learning rate of $6e-6$, a weight decay of 0.1 and gradient clipping of 1.0. We utilize 40 A100 GPUs and allocate a batch size of 9 on each device, resulting in an overall batch size of 360.

LoRA Fine-tuning. LoRA, introduced by (Hu et al., 2022), can decrease the the number of trainable parameters by optimizing pairs of rank-decomposition matrices for pre-trained LLMs, while the

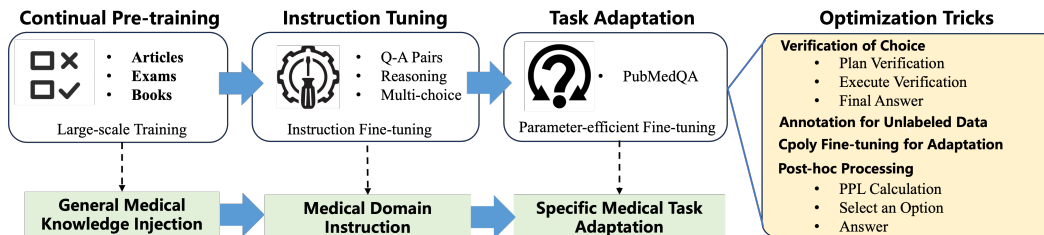


Figure 2: The detailed techniques for different optimization stages.

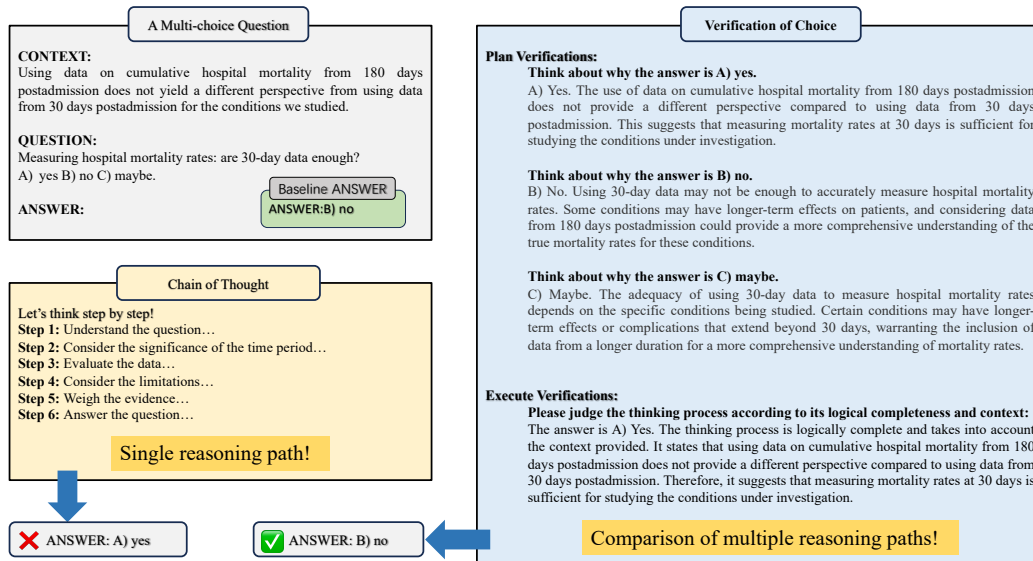


Figure 3: A comparison example for Chain-of-Thought and Verification-of-Choice.

original pre-trained model weights are frozen. The approach greatly reduces the storage requirement for LLMs, especially when adapted to some specific tasks. Thus, LoRA enables efficient task-switching without introducing inference latency. For the performance, LoRA can also outperform several other adaptation methods including adapter, prefix-tuning, and fine-tuning. We use the same SFT strategy as described by (Hu et al., 2022) for the PubMedQA benchmark.

3.4.3 FINE-TUNING FOR SPECIFIC MEDICAL TASK ADAPTATION

In this section, we provide a detailed description for the prompting strategies used for specific medical task adaptation.

Chain-of-Thought Chain-of-Thought (CoT) is firstly introduced by Wei *et al.* Wei et al. (2022), which augments few-shot examples as a enhanced prompt with a step-by-step explanation. The approach empowers an LLM to condition on multi-step outputs towards the final answer. Medical questions usually require a complex multi-step reasoning process, which is well fit for CoT prompting. According to CoT, we exploited a self-generated explanation of each choice for the given medical questions, which provides the reason why an LLM gives this answer.

Chain-of-Verification. Chain-of-Verification (COVE) Dhuliawala et al. (2023) is proposed to alleviate the hallucination issue, enabling LLMs deliberate on their responses for correcting the mistakes. Given the initial response as a draft, an LLM is required to plan verification questions for fact-checking the draft. Then the LLM should answers those questions independently, guaranteeing each response is not biased by others. Based on the above verification, the LLM can generate the final verified response with a higher confidence.

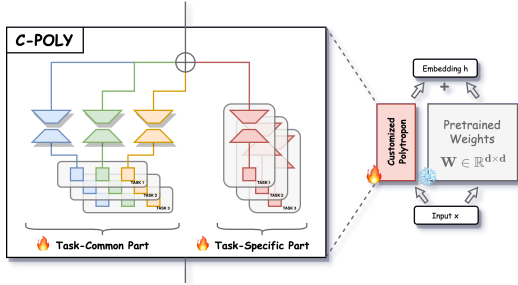


Figure 4: Overview of C-Poly framework.

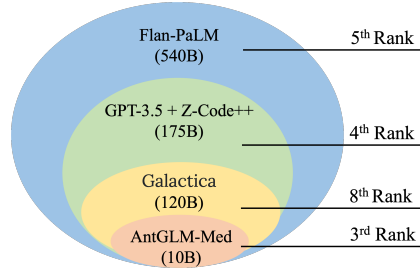


Figure 5: The PubMedQA leaderboard.

Verification-of-Choice. Building on COT Wei et al. (2022) and COVE Dhuliawala et al. (2023), we presented a simple prompting strategy named as Verification-of-Choice (VoC) for multi-choice medical questions. VoC involves conditioning an LLM on its own generations for each choice before selecting a choice as the final answer.

The overall process of VoC contains three steps: **(1) Plan Multi-Choice Verifications:** Given a query, assume that each choice is the true answer, let an LLM self-generate the corresponding explanation for each choice. Each explanation can be taken as a CoT for the choice. **(2) Execute Multi-Choice Verifications:** Taking multi-choice verifications as a context, LLM could make a comparison among them and self-analyze if there are any inconsistencies. **(3) Generate Final Response:** Given the inconsistencies (if any) among multi-choice verification, LLM could generate a final response.

Unlike CoT and COVE, VoC may be used to investigate the inconsistency between the explanations of multiple choices and questions, thus it is helpful to produce more accurate responses. For example, the explanation of a choice is different with the conditions in the question. In this work, we apply VoC only for multiple-choice question task.

Post-hoc Processing. PPL ranking is adopted as the indicator of uncertainty. For multi-choice QA datasets, the specific solution is to concatenate the original text with each option. Then a LLM is required to calculate the corresponding perplexity (Perplexity score) and select the option with the smallest PPL as the final predicted option. As for other datasets, generation and post-processing are utilized to produce the final response.

LLM-annotated PQA-U. As shown in Table 5, the PQA-U subset of the PubMedQA dataset is not well utilized due to without accurate annotations. It is noticed that long answers of the PQA-U subset are provided, which implicitly involve the correct answer for each question. To fully utilize the PQA-U subset, given a question in PQA-U, LLMs are required to generate responses for questions with the corresponding long answer, where VoC strategy is used to improve the correctness of answers. Then the generate answers are taken as pseudo annotations of PQA-U, which join the stage of specific medical task adaptation.

To evaluate the annotation accuracy for PQA-U, we fine-tuned two models and reported their performance on the test set of PubMedQA. Concretely, one model is fine-tuned on the PQA-L and PQA-U annotated using long answers. Another model is fine-tuned on the PQA-L and PQA-U annotated using both long answers and VoC. The results are reported in Section.

CPoly Fine-tuning. C-Poly, a fine-tuning method by Anonymous (2023), uses multi-task learning and adapters to differentiate shared and customized skills. It allows multi-class task samples to learn from each other in multi-adapter PEFT. The unified MTL framework C-Poly (Figure 4) enhances sample efficiency across tasks. For T tasks with data x^t , the PLE-like structure has adapter modules Φ with $|\Phi_A| + T \times |\Phi_B| = A + T \times B$ adapters, assuming $B = 1$. The C-Poly output is the sum of shared and task-specific adapters (Equation 1), with w_i as learnable weights.

$$\underbrace{\sum_{i=1}^A w_i^t \phi_i(x^t)}_{\text{Task-Common}} + \underbrace{\sum_{j=1}^B w_j^t \phi_j^t(x^t)}_{\text{Task-Specific}} = \sum_{i=1}^A w_i^t \phi_i(x^t) + w^t \phi^t(x^t) \quad (1)$$

Table 6: The leaderboard for the PubMedQA dataset. Our AntGLM-Med-10B could obtain competitive performance with a relatively small model size.

Model	Size	Accuracy
Med-PaLM 2 (Singhal et al., 2023)	NA	81.8
Palmyra-Med (Kamble & Alshikh, 2023)	40B	81.1
AntGLM-Med	10B	80.6
GPT-4-base (Nori et al., 2023)	NA	80.4
GPT-3.5 + Z-Code++ (He et al., 2023)	175B	79.6
Flan-PaLM-3-shot (Singhal et al., 2022)	540B	79.0
Codex-5-shot (Liévin et al., 2023)	175B	78.2
Human Performance (Jin et al., 2019)	NA	78.0
Galactica (Taylor et al., 2022)	120B	77.6
GatorTronGPT (Peng et al., 2023)	20B	77.6

The allocation matrix $\mathbf{W} \in \mathbb{R}^{T \times (A+T)}$ differentiates shared (\mathbf{W}_A) and customized (\mathbf{W}_B) skills. Different learning methods optimize skill acquisition, using low-rank approximations for efficiency. Shared skills use a Gumbel-sigmoid approach for differentiable sampling. Specialized skills are learned by differentiating shared and exclusive modules.

By combining the above-mentioned three optimization stages together, we fine-tune AntGLM-10B on the collected large-scale, high-quality, medical-domain corpus, resulting in our final model AntGLM-Med-10B.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

In our experiments, we mainly discuss the LLM’s performance after the final stage (Specific Medical Task Adaptation). The details are described as follows. In the case of vanilla LoRA, we set the rank of the low-rank approximation, $r = 8$. We utilize 4 parallel LoRA for task-common skills and 1 LoRA for task-custom skill in CPOLY, the rank is set to $r = 4$. This decision is made to ensure a comparable number of training parameters across all methods. We train our model for 10 epoch with a batch size of 12 on given datasets during training. The AdamW optimizer (Loshchilov & Hutter, 2017) is used with a learning rate of $5e^{-5}$. We also employ the linear decay strategy (Loshchilov & Hutter, 2016) as the learning rate scheduler with a weight decay of 0.01 and a warming up ratio of 0.06.

4.2 MODEL EVALUATION

To evaluate our AntGLM-Med-10B model, we consider multi-choice question as the adaptation task in the third optimization stage). The PubMedQA dataset is utilized as the evaluation dataset, which requires medical research comprehension skills to make reasoning over PubMed abstract context. We use 500 test samples for evaluation.

4.3 EXPERIMENTAL RESULTS

4.3.1 MAIN RESULTS

Results for PubMedQA. On PubMedQA, AntGLM-Med-10B obtained a score of 80.6%. This is below the state-of-the-art performance (81.8 from Med-PaLM 2 (Singhal et al., 2023)) and second place (81.1 from Palmyra-Med (Kamble & Alshikh, 2023)). The main reason is due to the relatively smaller model size, *e.g.*, AntGLM-Med-10B *vs.* Palmyra-Med-40B. Although that, AntGLM-Med-10B could exhibit improved performance compared to some other larger LLMs (as shown in Figure 5). Besides, AntGLM-Med-10B surpasses the models with similar model sizes (*e.g.*, GatorTronGPT (Peng et al., 2023)).

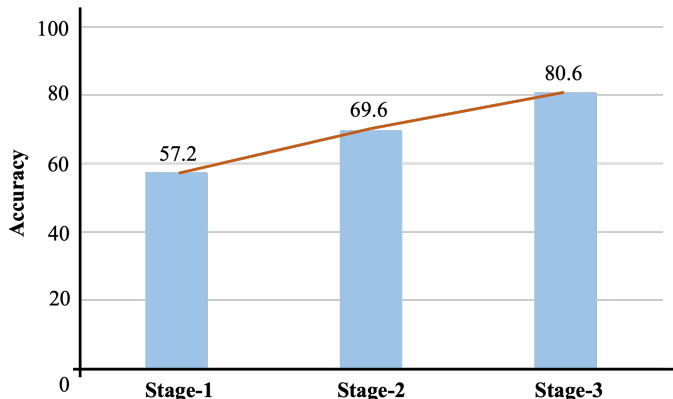


Figure 6: Accuracy results on PubMedQA at different optimization stages. Stage-1 is the accuracy based on general medical knowledge injection. Stage-2 is for medical domain instruction based on Stage-1. Stage-3 is for specific medical task adaptation based on Stage-2.

Table 7: The accuracy results for models fine-tuned on PQA-L and PQA-U. PQA-U are annotated using different strategies. Long answer and VoC can provided the best accuracy results for LoRA and full parameter fine-tuning adaptation strategies, indicating that VoC can further improve the annotation accuracy for PQA-U.

Fine-tuning	Annotation Strategy	PubMedQA
LoRA	Long Answer	86.6
	Long Answer + VoC	87.6
Full Parameter Tuning	Long Answer	87.8
	Long Answer + VoC	88.2

Table 8: The detailed performance for specific task adaptation using different datasets and tuning strategies.

Fine-tuning	Dataset	Size	PubMedQA	
Cpoly	PubMedQA	500	80.6	
	MedMCQA	500		
	MedQA	500		
	PubMedQA	500	80.4	
	MedMCQA	500		
	MedQA	500		
LoRA	PubMedQA	500	80.2	
	MedQA	500		
	MedMCQA	500	-	
	MedQA	500		
	PubMedQA	500		78.8
	MedMCQA	500		78.6
MedQA	500	78.0		

Results for Different Stages. As shown in Figure 6, we observed continuous accuracy improvements on the PubMedQA dataset throughout the entire optimization procedure. It is observed that the base model only obtained an initial accuracy of 57.2% on PubMedQA, and ultimately achieved a 80.6% accuracy by the end of the optimization procedure, which indicate the effectiveness of the 3-stage optimization.

4.3.2 MORE DISCUSSIONS

Effectiveness of VOC. To evaluate the annotation accuracy for PQA-U, we fine-tuned two LLM models and reported their performance on the test set of PubMedQA. Concretely, one model is fine-tuned on the PQA-L and PQA-U, where PQA-U is annotated only using long answers. Another model is fine-tuned on the PQA-L and PQA-U, where PQA-U is annotated using both long answers and VoC. The results are reported in Table 7. Using both long answer and VoC to annotate PQA-U, the model can exhibit higher accuracy, which indicate that VoC can help to improve the annotation accuracy for PQA-U.

Effectiveness of CPoly. We evaluated the effectiveness of CPoly on multiple datasets and conducted an ablation experiment in Table 8. We randomly selected 500 samples from the PubMedQA, MedMCQA, and MedQA datasets respectively, and combined them using different strategies. We then validated the training results of a single, two, and three dataset samples, respectively. Since the router vectors of multiple adapters trained in CPoly cannot index and effectively predict untrained unknown tasks, we only reported the dataset’s performance with the corresponding training set when using CPoly. The training results of CPoly on multiple datasets achieved significantly more improvement than those of LoRA on a single dataset. As the number of datasets involved in the CPoly training process increases, the best performance on PubMedQA increases. As the number of datasets trained with multitasking increased, the performance is significantly improved compared with training on a single dataset.

5 CONCLUSION

In this work, we explore how to adapt a pre-trained general LLM in medical domain, from a medical beginner to a medical expert. Starting from a pre-trained general LLM model (AntGLM-10B), we leverage a 3-stage optimization procedure to fine-tune it, *i.e.*, continual pre-training for medical knowledge injection, medical domain instruction tuning, and specific medical task adaptation. Different large-scale medical datasets are collected, covering various data types and different tasks, such as question-answering (QA), medical reasoning, multi-choice question, and medical conversations. Besides, for multi-choice question in medical domain, we design a novel Verification-of-Choice approach for prompting engineering, which can significantly enhance the reasoning ability of LLMs. By combining the above points together, our AntGLM-Med-10B model can exhibit competitive performance compared with both general LLMs and other LLMs pre-trained on medical knowledge on the PubMedQA. It is noticed that AntGLM-Med-10B can outperform the most of larger LLMs (>40B).

REFERENCES

- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Anonymous. Customizable combination of parameter-efficient modules for multi-task learning. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=G1Hlubz1fR>. under review.
- Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. GLM: general language model pretraining with autoregressive blank infilling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 320–335. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.26. URL <https://doi.org/10.18653/v1/2022.acl-long.26>.
- Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fujun Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*, 2023.
- Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, Wayne Xiong, Michael Zeng, Jianfeng Gao, and Xuedong Huang. Z-code++: A pre-trained language model optimized for abstractive summarization, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL <https://aclanthology.org/D19-1259>.
- Kiran Kamble and Waseem Alshikh. Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance, 2023.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5, 2023.
- Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner’s perspective. *arXiv preprint arXiv:2202.01875*, 2022.
- Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset, 2023.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory, 2023.
- Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions?, 2023.
- Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. Gptheval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*, 2023.
- Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.
- OpenAI. Chatgpt. 2022. URL <https://chat.openai.com/chat>.
- OpenAI. Gpt-4 technical report, 2023.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pp. 248–260. PMLR, 2022.

- Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A Mitchell, Naykky S Ospina, Mustafa M Ahmed, William R Hogan, Elizabeth A Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare, 2023.
- Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. Expert discussions improve comprehension of difficult cases in medical image assessment. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.
- 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. Large language models encode clinical knowledge, 2022.
- 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. Towards expert-level medical question answering with large language models, 2023.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.
- Jason Yim, Reena Chopra, Terry Spitz, Jim Winkens, Annette Obika, Christopher Kelly, Harry Askham, Marko Lukic, Josef Huemer, Katrin Fasler, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nature Medicine*, 26(6):892–899, 2020.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. Promptblue: A chinese prompt tuning benchmark for the medical domain. *arXiv preprint arXiv:2310.14151*, 2023.

Table 9: A prompt example for generating LLM-annotated PQA-U on PubMedQA dataset.

Context	<p>OBJECTIVE. We compare 30-day and 180-day postadmission hospital mortality rates for all Medicare patients and those in three categories of cardiac care: coronary artery bypass graft surgery, acute myocardial infarction, and congestive heart failure.</p> <p>DATA SOURCES/COLLECTION. Health Care Financing Administration (HCFA) hospital mortality data for FY 1989.</p> <p>STUDY DESIGN. Using hospital level public use files of actual and predicted mortality at 30 and 180 days, we constructed residual mortality measures for each hospital. We ranked hospitals and used receiver operating characteristic (ROC) curves to compare 0-30, 31-180, and 0-180-day postadmission mortality.</p> <p>PRINCIPAL FINDINGS. For the admissions we studied, we found a broad range of hospital performance when we ranked hospitals using the 30-day data; some hospitals had much lower than predicted 30-day mortality rates, while others had much higher than predicted mortality rates. Data from the time period 31-180 days postadmission yield results that corroborate the 0-30 day postadmission data. Moreover, we found evidence that hospital performance on one condition is related to performance on the other conditions, but that the correlation is much weaker in the 31-180-day interval than in the 0-30-day period. Using ROC curves, we found that the 30-day data discriminated the top and bottom fifths of the 180-day data extremely well, especially for AMI outcomes.</p> <p>QUESTION: Measuring hospital mortality rates: are 30-day enough? A) yes B) no C) maybe</p>
Verification-of-Choice	<p>Step-1: Plan Multi-Choice Verification Think about why the answer is A) yes. A) Yes. The use of data on cumulative hospital mortality from 180 days postadmission does not provide a different perspective compared to using data from 30 days postadmission. This suggests that measuring mortality rates at 30 days is sufficient for studying the conditions under investigation. Think about why the answer is B) no. B) No. Using 30-day data may not be enough to accurately measure hospital mortality rates. Some conditions may have longer-term effects on patients, and considering data from 180 days postadmission could provide a more comprehensive understanding of the true mortality rates for these conditions. Think about why the answer is C) maybe. C) Maybe. The adequacy of using 30-day data to measure hospital mortality rates depends on the specific conditions being studied. Certain conditions may have longer-term effects or complications that extend beyond 30 days, warranting the inclusion of data from a longer duration for a more comprehensive understanding of mortality rates</p> <p>Step-2: Execute Multi-Choice Verification Please judge the yes/no/maybe thinking process according to its logical completeness and context. The answer is A) Yes. The thinking process is logically complete and takes into account the context provided. It states that using data on cumulative hospital mortality from 180 days postadmission does not provide a different perspective compared to using data from 30 days postadmission. Therefore, it suggests that measuring mortality rates at 30 days is sufficient for studying the conditions under investigation.</p>
	<p>Step-3: Generate Final Response</p> <p>Answer: A</p>