MedCare: Advancing Medical LLMs through Decoupling Clinical Alignment and Knowledge Aggregation

Yusheng Liao[*,♠,◇], **Shuyang Jiang**[*,♣,◇], **Zhe Chen**[♠,◇], **Yanfeng Wang**[♠,◇], **Yu Wang**[†,♠,◇]
[♠]Shanghai Jiao Tong University
[♣]Fudan University
[◇]Shanghai Artificial Intelligence Laboratory
{liao20160907,chenzhe2018,wangyanfeng622,yuwangsjtu}@sjtu.edu.cn
shuyangjiang23@m.fudan.edu.cn

Abstract

Large language models (LLMs) have shown substantial progress in natural language understanding and generation, proving valuable especially in the medical field. Despite advancements, challenges persist due to the complexity and diversity inherent in medical tasks, which can be categorized as knowledge-intensive tasks and alignment-required tasks. Previous approaches either ignore the latter task or focus on a minority of tasks and hence lose generalization. To address these drawbacks, we propose a progressive fine-tuning pipeline. This pipeline employs a Knowledge Aggregator and a Noise Aggregator to encode diverse knowledge in the first stage and filter out detrimental information. In the second stage, we drop the Noise Aggregator to avoid the interference of suboptimal representation and leverage an additional alignment module optimized towards an orthogonal direction to the knowledge space to mitigate knowledge forgetting. Based on this two-stage paradigm, we proposed a **Med**ical LLM through decoupling **C**linical **A**lignment and Knowledge Agg**re**gation (MedCare), which is designed to achieve promising performance on over 20 medical tasks, as well as results on specific medical alignment tasks. Various model sizes of MedCare (1.8B, 7B, 14B) all demonstrate significant improvements over existing models with similar model sizes. Our code and datasets are available at https://github.com/BlueZeros/MedCare.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2022; AI, 2024; Singhal et al., 2023a) have made significant strides in the realms of natural language generation and thereby finding extensive applications across a broad spectrum of disciplines (Liu et al., 2023a; Deng et al., 2024). Among these, the medical field has attracted considerable attention due to its importance and demand. Much effort has been dedicated to researching and developing specialized LLMs for this domain (Li et al., 2023b; Wang et al., 2023a; Singhal et al., 2023a).

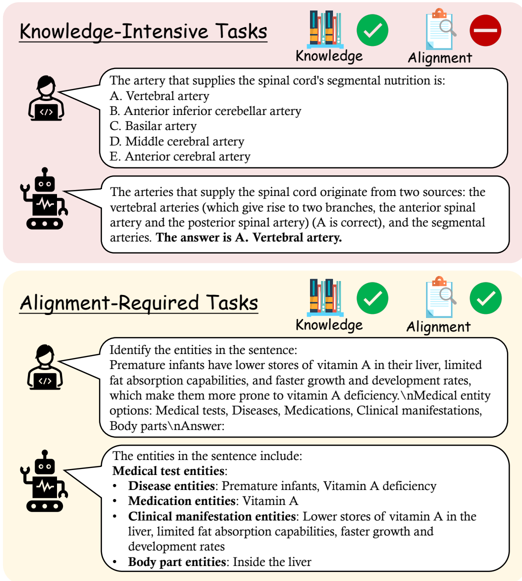


Figure 1: Examples of two types of medical tasks. Knowledge-intensive tasks require models to possess sufficient knowledge, whereas alignment-required tasks additionally necessitate the model to meet specific requirements criteria.

Although progress has been made, LLMs still face challenges due to the complexity and diversity of medical tasks. This is particularly evident when data is intentionally structured to represent a diverse set of tasks (Chung et al., 2022; Wei et al., 2021; Sanh et al., 2021). We categorize medical tasks into two types: knowledge-intensive tasks and alignment-required tasks. Knowledge-intensive tasks, such as medical question answering (Zhang et al., 2018) and medical dialogues (Liu et al., 2022a; Zhao et al., 2022), primarily test the model's understanding of internal medical knowledge. Conversely, alignment-required tasks, like

[*]Equal contribution.
[†]Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10562–10581
November 12-16, 2024

clinical terminology standardization (Zhang et al., 2022b) and medical entity recognition (Hongying et al., 2021), demand not only medical knowledge but also strict adherence to output formats. Figure 1 illustrates these two categories. In general scenarios, like medical question-answering or dialogue, the LLMs only need to answer the user's questions correctly. However, tasks in clinical applications like report generation or situations usually require the model to produce a correct response and adhere to specific formats.

Previous medical LLMs, such as HuatuoGPT (Wang et al., 2023a), HuatuoGPT-II (Chen et al., 2023a), Med-PaLM (Singhal et al., 2023a) and BioMistral (Labrak et al., 2024), have primarily focused on enhancing the encoding of medical knowledge in LLMs, neglecting the requirements of alignment-required tasks (Van Veen et al., 2024). On the other hand, alignment-oriented methods (Liu et al., 2023b; Cai et al., 2023; Wang et al., 2023a) that perform alignment through fine-tuning often suffer from an "alignment tax". This results in knowledge forgetting and performance drops in knowledge-intensive tasks (Lin et al., 2023; Gekhman et al., 2024). These issues limit the practical use of LLMs in healthcare.

To create a more practical medical LLM, in this paper we propose a novel training pipeline consisting of two progressive fine-tuning stages: miscellaneous knowledge aggregation (MKA) and downstream alignment (DA). In the first stage, we introduce two modules, KNOWLEDGE AGGREGATOR (KA) and NOISE AGGREGATOR (NA), to encode advantageous knowledge and noisy contents, respectively. The NOISE AGGREGATOR is asymmetric with respect to KNOWLEDGE AGGREGATOR, which explicitly induces KNOWLEDGE AGGREGATOR to perform multi-task knowledge extraction. To avoid catastrophically parameterized knowledge disruption, we innovatively remove the updated NOISE AGGREGATOR but retain only the KNOWLEDGE AGGREGATOR after training. Following this stage, we introduce an alignment module to cater to the downstream alignment requirements from the specific task. We add an orthogonal regularization term to ensure non-overlapping between the optimization space of alignment and knowledge space in the first stage. Built upon this pipeline, we propose MEDCARE with three sizes (1.8B, 7B, 14B), a specialized LLM tailored for both knowledge-intensive tasks and alignment-required tasks, derived from the Qwen1.5 series. Figure 2 provides a visual representation of MEDCARE.

In summary, our contributions are as follows:

- We introduce a taxonomy for medical tasks, which divides all tasks into knowledge-intensive tasks and alignment-required tasks. This taxonomy reinforces the practical usability requirements for medical LLMs.
- We introduce a two-stage fine-tuning pipeline that balances knowledge maintenance and downstream alignment. This approach not only encodes knowledge without disruption but also adapts to specific tasks with minimal knowledge forgetting.
- Based on the progressive pipeline, we introduce MEDCARE, a medical LLM to simultaneously encode massive knowledge and align with practical requirements under the medical multi-task taxonomy. To our knowledge, MEDCARE is the first LLM to effectively handle such a wide spectrum of tasks in the medical domain with few alignment taxes.
- We conduct extensive experiments on over 20 knowledge-intensive and alignment-required tasks, benchmarking MEDCARE against a diverse array of established language models. Our results demonstrate the superior performance of MEDCARE, affirming its effectiveness for both genres of tasks.

2 Preliminaries

Low-rank Adaptation (LoRA) Given an input sequence $\mathbf{X} \in \mathbb{R}^{m \times d}$, LoRA (Hu et al., 2021) proves that the update of original linear layers ΔW in large language models is of low-rank, and can be decomposed into the multiplication of two compact matrices AB. The pretrained weight W is frozen in the training phase and does not undergo gradient updates, while A and B are trainable parameters and contribute together to the forward pass:

$$\mathbf{H} = \mathbf{X}W + \mathbf{X}\Delta W = \mathbf{X}W + \frac{\alpha}{r}\mathbf{X}AB \quad (1)$$

where $W \in \mathbb{R}^{d \times d'}, A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times d'}$ and $r \ll d, d'$. $\mathbf{H}$ is the processed output. α is the scaling factor. Without loss of generality, we omit the layer index for the following formula. At the

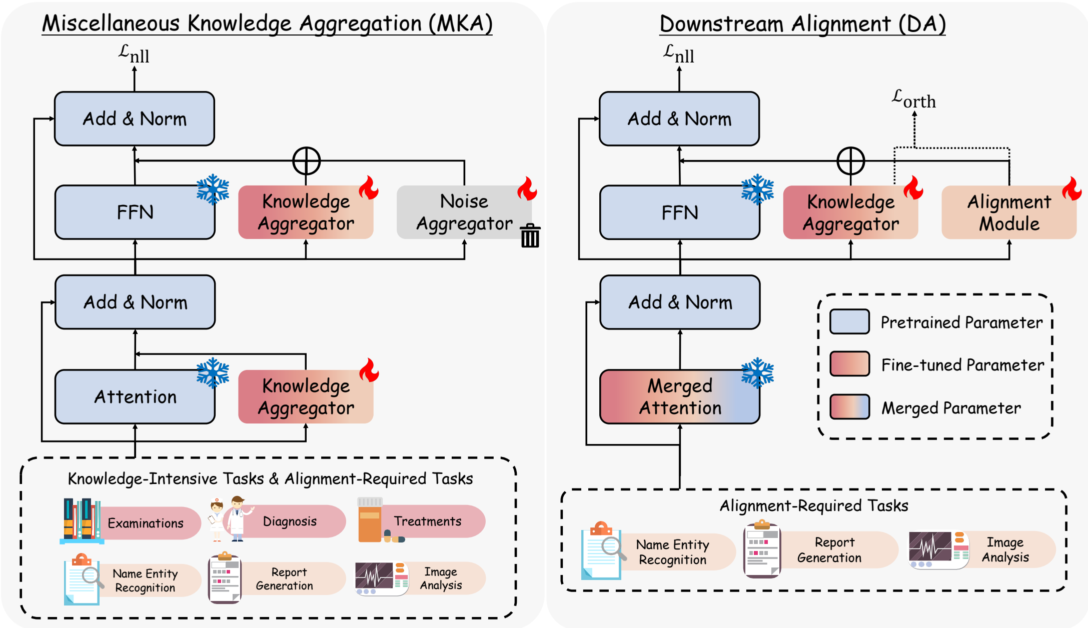


Figure 2: Overview of the proposed MEDCARE. In the MKA stage, MEDCARE encodes advantageous knowledge and noisy contents with KNOWLEDGE AGGREGATOR and NOISE AGGREGATOR from both types of tasks, respectively. The updated NOISE AGGREGATOR is removed to avoid the knowledge disruption. In the DA stage, an additional alignment module and orthogonal regularization are introduced to cater to the requirements of the alignment tasks.

beginning of training, B is initialized to an all-zero matrix and A uses Gaussian initialization to make sure that the product AB is zero at initialization.

Mixture of LoRA The mixture of the LoRA (MoLoRA) module is based on the Feed-forward Network (FFN) module in the LLMs since experts of FFN can store diverse task knowledge (Geva et al., 2020) from multi-task learning. In the popular LLaMA-style FFN, the input $\mathbf{X}$ is computed using an SiLU (Elfwing et al., 2018) gate as follow:

$$\mathbf{H} = (\mathbf{X}W_u \cdot \text{silu}(\mathbf{X}W_g))W_d \tag{2}$$

where $W_g \in \mathbb{R}^{d \times d'}, W_u \in \mathbb{R}^{d \times d'}, W_d \in \mathbb{R}^{d' \times d}$, and $d' = 8/3d$. In the LoRA forward passes, each linear layer $W \in \{W_d, W_u, W_g\}$ is updated using the LoRA module described in Eq.1. Consider an MoLoRA module with E experts, and each expert is denoted as $\{E_i\}_{i=1}^{E}$, the forward pass of the linear layer is formulated as:

$$\mathbf{H} = \mathbf{X}W + \mathbf{X}\Delta W \tag{3}$$

$$\mathbf{X}\Delta W = \sum_{i=1}^{K} G(\mathbf{X})_i E_i(\mathbf{X}) \tag{4}$$

where $G(\cdot) = \text{softmax}(\mathbf{X}W_r), W_r \in \mathbb{R}^{d \times E}$ is the router in MoLoRA with top-K selection which holds the top-K affinity with respect to current input $\mathbf{X}$. Assuming each LoRA expert shares the same rank r and α and we obtain the weight w_i from the router $G(\cdot)$ as $w_i = G(\mathbf{X})_i$, the overall output of the MoLoRA linear layer can be formulated as:

$$\mathbf{h} = W\mathbf{x} + \frac{\alpha}{r} \sum_{i=1}^{K} w_i \cdot A_i B_i \mathbf{x} \tag{5}$$

where $\mathbf{x}, \mathbf{h} \in \mathbb{R}^d$ is any token representation of $\mathbf{X}$ and the frozen W inherits from the pre-trained linear weight.

3 MEDCARE

In this paper, we categorize the medical task into two primary genres: knowledge-intensive tasks and alignment-required tasks. Knowledge-intensive tasks require LLMs to seamlessly transfer both medical knowledge and their general domain reasoning capabilities into the medical domain. Models leverage the internal broad pre-trained knowledge without significantly disrupting the pretrained distribution, to fulfill domain-specific medical reasoning tasks. In contrast, alignment-required tasks pose a greater challenge. These tasks typically deviate more substantially from the model's original training paradigms, which can lead to hallucinations (Gekhman et al., 2024) and the aliasing of pre-trained knowledge (Dou et al., 2023). To address this challenge, we propose MEDCARE, a method designed to simultaneously acquire the necessary knowledge for the general adaptation of LLMs and to achieve specialized alignment for specific tasks. The fine-tuning process of MEDCARE comprises two stages: miscellaneous knowledge aggregation (MKA; §3.2) and downstream alignment (DA; §3.3). The training procedure for MEDCARE is illustrated in Figure 2.

3.1 Problem Formulation

In the MKA stage, the dataset $\mathcal{D}_{\text{MKA}}$ comprises N samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, derived from tasks that are both knowledge-intensive and require alignment. Here, $\mathbf{x}_i$ and $\mathbf{y}_i$ represent the input and target for the models, respectively. LLMs assimilate diverse knowledge from these tasks to augment their knowledge and reasoning capabilities. Subsequently, these LLMs are exposed to a specific subset of the alignment-required data $\mathcal{D}_{\text{DA}}$, which is necessary for learning the intricacies of various alignment-required tasks, including named entity recognition with specificity and report analysis that demands explanatory detail.

3.2 Miscellaneous Knowledge Aggregation

In this stage, the models acquire common medical knowledge from both types of tasks. To absorb miscellaneous knowledge into single LLMs without interference, we adopt the MoLoRA structure (Li et al., 2024; Su et al., 2024; Feng et al., 2024; Luo et al., 2024) as the NOISE AGGREGATOR and introduce shared experts as the KNOWLEDGE AGGREGATOR to further circumvent low generalization due to the high specialization of each expert (Gou et al., 2023) and parameter redundancy (Dai et al., 2024) on the FFN module of the models. Through backpropagation, the KA acquires common knowledge from multiple datasets, while NA learns the distinct alignment-required tasks. Consequently, the KA absorbs common knowledge encoded in each task, while the NA learns the interfering alignment requirements.

$$\mathbf{H} = \mathbf{X}W + \text{KA}(\mathbf{X}) + \text{NA}(\mathbf{X}) \tag{6}$$
$$= \mathbf{X}W + \text{LoRA}_s(\mathbf{X}) + \text{MoLoRA}(\mathbf{X}) \tag{7}$$

where W denotes any weight matrix in FFN modules. The rank of NA is r, and the rank of KA is $r' = sr$, where s is the number of the share experts. For the attention module, we only add a vanilla LoRA module to each projection layer W as Eq.1 to guarantee sufficient optimization space. As most task knowledge is encoded in the common knowledge aggregator, we discard the separate knowledge aggregator after this training stage. In other words, we introduce redundant parameters to avoid erroneous leverage of task-agnostic knowledge through decoupling the KA and NA modules, employing only the KA module:

$$\mathbf{H} = \mathbf{X}W + \text{KA}(\mathbf{X}) \tag{8}$$

While discarding the separate KA prevents the model's knowledge from being disturbed, it also compromises the model's alignment capabilities, thus leading to suboptimal performance on alignment-required tasks. Therefore, a second stage of training is necessary to enable the model to learn the requirements of alignment tasks while preserving its knowledge reasoning capabilities.

3.3 Downstream Alignment

In this stage, MEDCARE merges back the LoRA weights of self-attention into the pretrained model and freezes the self-attention module to ensure the instruction-following ability of LLMs (Wu et al., 2023). We introduce an additional alignment LoRA module $\text{Align}(\cdot)$ in the FFN module to acquire specific alignment knowledge from the alignment dataset, the forward pass of linear layers of FFN can be formulated as:

$$\mathbf{H} = \mathbf{X}W + \text{KA}(\mathbf{X}) + \text{Align}(\mathbf{X}) \tag{9}$$

Following Wang et al. (2023c), we introduce an orthogonal loss to ensure that the new alignment task is learned in a direction orthogonal to the original knowledge task. For the LoRA weights of KA $\{A_k^p, B_k^p\}_{p=1}^{P}$ and Align $\{A_d^p, B_d^p\}_{p=1}^{P}$, where P is the total number of modules, we achieve the orthogonal subspaces of the alignment stage as:

$$\arg\min_{A_d^p} O_{k,d}^p = A_k^{p\top} A_d^p \tag{10}$$

Based on the subspace learning, the learning objective of this stage is expressed as

$$\mathcal{L} = \mathcal{L}_{\text{nll}} + \lambda \mathcal{L}_{\text{orth}} \tag{11}$$
$$= -\sum_{\mathbf{x},\mathbf{y}\in\mathcal{D}_{\text{DA}}} \log p_\theta(\mathbf{y}|\mathbf{x}) + \lambda \sum_{p=1}^{P} \left| O_{k,d}^p \right| \tag{12}$$

Model	MedQA	MMB.	CMB	CMExam	CMMLU	CEval	PLE Pha	PLE TCM	Avg.
Llama3-8B	59.40	63.78	41.63	44.99	51.40	53.66	38.33	33.54	48.34
ChatGLM3-6B	44.51	51.34	39.81	43.21	46.97	48.80	34.60	32.90	42.77
Baichuan2-7B	45.97	52.39	46.33	50.48	50.74	51.47	44.60	42.10	48.01
Baichuan2-13B	49.42	56.71	50.87	54.90	52.95	58.67	44.20	41.70	51.18
Qwen1.5-7B	74.46	78.58	61.33	58.24	66.77	68.29	32.29	29.17	58.64
Qwen1.5-14B	**81.93**	84.33	64.33	57.79	68.76	78.05	48.33	38.33	65.23
ChatGPT	37.51	40.08	43.26	46.51	50.37	48.80	41.20	31.20	42.81
HuatuoGPT-II-7B	59.22	62.03	60.39	65.81	59.08	62.40	47.70	47.50	58.02
HuatuoGPT-II-13B	75.77	78.69	63.34	68.98	61.45	64.00	52.90	51.60	64.59
MEDCARE-1.8B	56.80	65.62	43.04	47.83	48.82	51.22	38.13	35.21	48.33
↦ *w/o* DA	61.70	64.62	49.28	52.21	51.18	60.98	36.04	39.58	51.95
MEDCARE-7B	76.77	80.79	60.13	65.33	64.33	65.85	47.29	52.08	64.07
↦ *w/o* DA	75.16	79.36	61.60	66.85	66.25	70.73	50.83	53.33	65.51
MEDCARE-14B	81.44	**84.76**	64.17	68.74	**71.57**	78.05	54.17	**54.58**	69.69
↦ *w/o* DA	77.41	83.36	**66.79**	**71.77**	69.42	**82.93**	**57.29**	54.17	**70.39**

Table 1: The results on medical knowledge exams. The results of the MedQA are evaluated on the Chinese Mainland subset. 'MMB.' indicates the Chinese subset of the MMedBench. 'PLE Pha' and 'PLE TCM' indicate the Pharmacy and Traditional Chinese Medicine tracks of the 2023 Chinese National Pharmacist Licensure Examination. Note that for the general benchmarks, CMMLU and CEval, we only chose questions related to the medical domain.

where $|O_{k,d}|$ denotes the sum of the absolute value of each entry of $O_{k,d}$, and λ is a hyperparameter to control the weights of orthogonal loss. During the training process, we do not follow Wang et al. (2023c) to fix the knowledge aggregator module since they can further acquire the specific knowledge that is lacking in the previous learning stage. After training, we can merge the updates of the knowledge aggregator and alignment module into the pretrained weights W to avoid the increased inference latency and GPU overhead:

$$W' = W + \frac{\alpha}{r'} A_k B_k + \frac{\alpha}{r'} A_d B_d \tag{13}$$

4 Experiments

MEDCARE is built upon the Qwen1.5-Chat [1] series. For knowledge-intensive tasks, we adopt the Chinese Mainland test set of **MedQA** (Jin et al., 2021), the Chinese subset of **MMedBench** (Qiu et al., 2024), and two comprehensive Chinese medical exam datasets, **CMB** (Wang et al., 2023d) and **CMExam** (Liu et al., 2024a). We also collect the medical parts of the general benchmarks, including **CMMLU** (Li et al., 2023a) and **CEval** (Huang et al., 2023). Besides, we test the performance of the models on the flash exam questions from the 2023 Chinese National Pharmacist Licensure Examination (**PLE**) collected by Chen et al. (2023a). For the alignment-required tasks, we use **CBLUE** (Zhang et al., 2022a) and our newly developed Chinese Clinical Task Evaluation (**CCTE**) dataset. More details of the experiment settings can be found in Appendix B.

[1] https://github.com/QwenLM/Qwen1.5

4.1 Results on Knowledge-Intensive Tasks

We utilized eight distinct datasets to rigorously assess the medical knowledge capabilities of the proposed model MEDCARE for comprehensive evaluation. As detailed in Table 1, MEDCARE has demonstrated exceptional proficiency in medical knowledge, particularly notable in the MEDCARE-14B, which has outperformed both ChatGPT and the leading open-source medical model, HuatuoGPT-II (13B). Remarkably, the MEDCARE-1.8B model matches the performance of ChatGLM3-6b, while MEDCARE-7B achieves comparable results to the HuatuoGPT-II (13B). Besides, without the DOWNSTREAM ALIGNMENT, MEDCARE achieves better performance.

To mitigate potential biases caused by data leakage, we further evaluated our models on the latest 2023 Chinese National Pharmacist Licensure Examinations (PLE) (Chen et al., 2023a), with the detailed results shown in Table 7. This assessment confirms that MEDCARE continues to outpace the competition, with MEDCARE-14B *w/o* DA surpassing the performance of GPT-4. These findings underscore MEDCARE's superior utilization of medical knowledge, achieving higher expertise levels with fewer parameters.

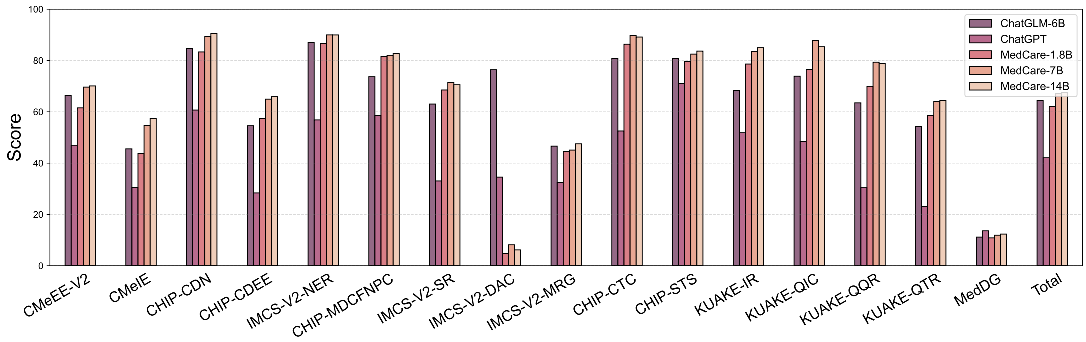


Figure 3: Results on 16 tasks in CBLUE. ChatGLM-6B is fine-tuned on the CBLUE dataset with LoRA and ChatGPT is augmented by in-context learning. The results are obtained from the official implementation of CBLUE.

Model	Report Generation					Image Analysis					Discharge Instruction					Examination Education					Avg.
	Flu.	Rel.	Com.	Pro.	Avg.	Flu.	Rel.	Com.	Pro.	Avg.	Flu.	Rel.	Com.	Pro.	Avg.	Flu.	Rel.	Com.	Pro.	Avg.	
ChatGLM2-6B	4.16	2.66	2.64	2.52	3.00	4.88	4.06	4.04	3.70	4.17	4.98	3.46	3.12	3.24	3.70	3.68	2.94	2.44	2.76	2.96	3.46
ChatGLM3-6B	4.52	2.78	2.66	2.52	3.12	4.98	4.16	4.20	3.94	4.32	5.00	3.34	3.24	3.04	3.66	4.52	2.88	2.14	2.56	3.03	3.53
Baichuan2-7B	4.88	3.36	3.30	3.40	3.74	5.00	4.40	4.40	4.02	4.46	5.00	3.96	3.84	3.82	4.16	3.98	2.74	1.68	2.20	2.65	3.75
Baichuan2-13B	4.94	3.42	3.40	3.46	3.81	5.00	4.42	4.42	4.22	4.52	5.00	3.92	3.92	3.88	4.18	4.56	2.88	1.40	2.18	2.76	3.81
Qwen1.5-7B	4.72	3.22	3.26	3.12	3.58	5.00	4.64	4.64	4.42	4.68	5.00	4.18	4.40	4.22	4.45	4.94	3.34	1.80	2.94	3.26	3.99
Qwen1.5-14B	4.86	3.54	3.50	3.50	3.85	5.00	4.66	4.66	4.44	4.69	5.00	4.10	4.16	4.10	4.34	4.80	3.28	1.68	2.56	3.08	3.99
ChatGPT	4.88	3.32	2.96	3.12	3.57	4.98	4.30	4.04	4.00	4.33	5.00	3.90	3.68	3.64	4.06	4.60	3.36	2.44	2.96	3.34	3.82
HuatuoGPT-II-7B	4.80	3.16	3.12	3.30	3.60	5.00	4.22	4.26	4.04	4.38	5.00	3.96	3.90	3.88	4.19	5.00	4.08	3.88	3.90	4.22	4.09
HuatuoGPT-II-13B	4.78	3.22	3.16	3.40	3.64	5.00	4.34	4.30	4.14	4.45	5.00	3.90	3.88	3.84	4.16	4.90	4.30	4.02	3.96	4.30	4.13
MedCare-1.8B	5.00	3.66	3.54	3.78	4.00	4.98	3.92	4.00	3.86	4.19	5.00	4.18	4.24	4.12	4.39	4.72	4.20	3.90	3.88	4.18	4.19
MedCare-7B	5.00	3.78	3.68	3.80	4.07	5.00	4.12	4.24	4.12	4.37	5.00	4.36	4.42	4.38	4.54	4.96	4.42	4.40	4.34	4.53	4.38
MedCare-14B	4.98	3.74	3.62	3.70	4.01	5.00	4.14	4.24	4.10	4.37	5.00	4.46	4.46	4.40	4.58	4.94	4.54	4.48	4.44	4.60	**4.39**

Table 2: Results on medical alignment task CCTE. 'Flu.' indicates 'Fluency', 'Rel.' indicates 'Relevance', 'Com.' indicates 'Completeness', and 'Pro.' indicates 'Proficiency'. The maximum value of all scores is 5.

4.2 Results on Alignment-Required Tasks

To evaluate the medical alignment ability of the models, we tested CBLUE and CCTE with 20 distinct alignment-required tasks. CBLUE requires models to generate outputs in a prescribed format. A higher score indicates the model aligns more closely with the task requirements. As depicted in Figure 3, the results indicate that even when augmented by In-Context Learning (ICL), ChatGPT still fails to fulfill such specific format requirements. Besides, MedCare-7B surpasses the established CBLUE baseline ChatGLM on nearly all the tasks with an average of more than 3 points. Since CBLUE strictly requires the output format, we mainly evaluate model alignment capabilities through ablation studies discussed in §5 instead of comparing it with other LLMs.

CCTE encompasses four clinical tasks that, while not specifying the format of model outputs, require a higher level of professional expertise. For this assessment, we employed GPT-4 to score the outputs across four dimensions comprehensively. As shown in Table 2, it is evident that even the MedCare-1.8B variant outperforms both general and medical-specific large language models, further confirming MedCare 's robust potential in clinical settings. More details about CCTE are discussed in Appendix C.

4.3 Ablation Experiments

The results of the ablation experiments are shown in Table 3. For the methods only fine-tuned with the MKA stage, inference with NA demonstrated minimal improvement in knowledge examination performance. Without the disturbance of the NA, MedCare achieved the best knowledge performance but failed to complete the alignment-required tasks. For the two-stage fine-tuning methods (MKA+DA), further fine-tuned with the NA module or without orthogonal regularization led to a noticeable reduction in performance on knowledge-intensive tasks. The orthogonal regularization can avoid parameter redundancy, facilitating more effective alignment learning while mitigating the loss of reasoning capability.

5 Discussion

In this section, we discuss the following research questions (RQ) of the proposed MedCare:

- **RQ1:** Why does Noise Aggregator have

MKA		DA		Knowledge-Intensive					Alignment-Required	
KA	NA	*w/o* $\mathcal{L}_{\text{orth}}$	*w/* $\mathcal{L}_{\text{orth}}$	CMMLU	CEval	PLE Pha	PLE TCM	Avg.	CBLUE	CCTE
Base Model				48.45	53.66	31.46	25.63	39.80	3.76	3.36
✓	✗	✗	✗	46.23	41.46	33.54	33.96	38.80	51.41	4.04
✓	✓	✗	✗	50.07	45.00	36.04	33.54	41.16	57.09	4.15
✓	☑	✗	✗	51.18	60.98	36.04	39.58	**46.95**	8.15	3.80
✓	✓	✗	✓	45.79	41.46	37.50	34.38	39.78	56.26	4.02
✓	☑	✓	✗	48.52	46.34	37.50	35.42	41.95	55.89	4.00
✓	☑	✗	✓	48.74	53.66	33.33	37.08	43.20	**62.05**	**4.22**

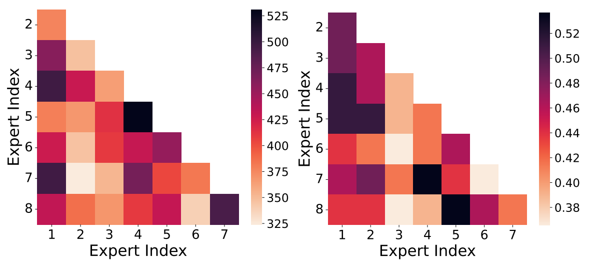
Table 3: Ablation Experiments of the MEDCARE methods. Note that ☑ indicates MEDCARE drops NOISE AGGREGATOR after MKA fine-tuning stage.

Model	Knowledge-Intensive					Alignment-Required	
	CMMLU	CEval	PLE Pha.	PLE TCM.	Avg.	CBLUE	CCTE
Base Models	48.45	53.66	31.46	25.63	39.80	3.76	3.36
Parallel LoRA	48.67	43.90	35.00	34.17	40.43	55.05	4.09
↦ *w/o* LoRA 1	51.26	51.22	34.79	35.00	43.07	11.43	3.89
↦ *w/o* LoRA 2	50.52	48.78	35.63	33.75	42.17	11.87	3.88
MEDCARE	50.07	45.00	36.04	33.54	41.16	**57.09**	**4.15**
↦ *w/o* NA	51.18	60.98	36.04	39.58	**46.95**	8.15	3.80
↦ *w/o* KA	49.19	39.02	34.17	32.50	38.72	50.22	4.06

Table 4: Performance of partial experts with MKA fine-tuning stage. 'Parallel LoRA' has two identical LoRAs on FFN modules, which are numbered with LoRA1 and LoRA2 for distinguishment.

a negative impact on the task performance?

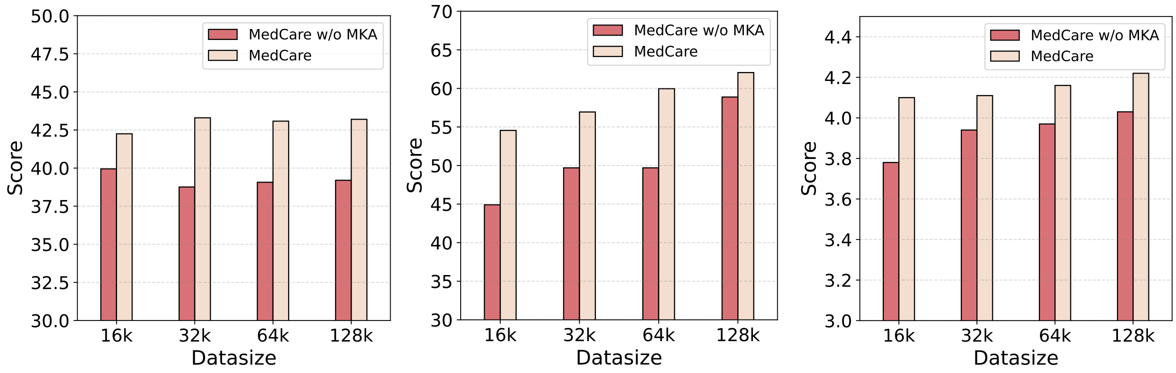
- **RQ2:** Are the roles of KNOWLEDGE AGGREGATOR and NOISE AGGREGATOR determined by the model architecture?
- **RQ3:** How do both two fine-tuning stages improve the model's capabilities?
- **RQ4:** Can MEDCARE still learn the knowledge effectively when the scale of fine-tuning corpus becomes smaller?
- **RQ5:** How does the effect of MEDCARE compare to other PEFT methods?
- **RQ6:** How does MEDCARE perform in other languages?



(a) (b)

Figure 4: Mismatch between expert activation times and performance on CEval. (a) Activation times of each expert combination. (b) Performance of only used each expert combination. The i-th raw and j-th column indicates the combinations of the i-th and j-th experts. The accuracy of the vanilla MoLoRA inference is 45.00.

Response to RQ1: NOISE AGGREGATOR interferes with models because of the routing mismatch. We investigate the effect of different combinations of MoLoRA experts in NOISE AGGREGATOR on the final performance to demonstrate the routing mismatch. We first compute the activation frequencies for each combination of the experts in vanilla MoLoRA inference. Meanwhile, different expert combinations are designated to test the performance of the models. The results in Figure 4 show the mismatch between the experts' activation times and their corresponding performance, which indicates that the router in the MoLoRA module failed to select the experts with optimal performance for each task and even caused the per-



(a) (b) (c)

Figure 5: Performance with different sizes of alignment-required datasets for DA fine-tuning stage. (a) The average performance on knowledge-intensive tasks. (b) The average performance on CBLUE. (c) The average performance on CCTE. 'MEDCARE *w/o* MKA' indicates the model is fine-tuned with only the second stage. Note that the score of the knowledge examinations is the average of the CMMLU, CEval, PLE Pharamy, and PLE TCM.

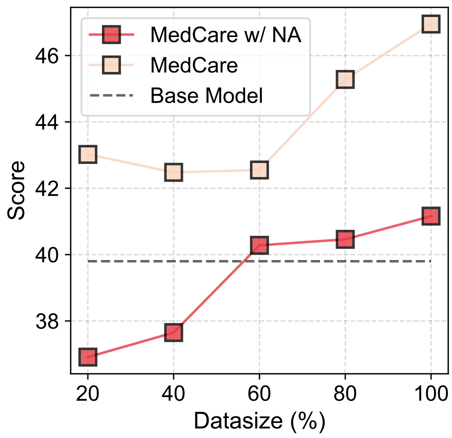


Figure 6: Average performance on the medical knowledge examination with different knowledge aggregation learning data size. 'Base Model' indicates the performance of the Qwen1.5-1.8B without fine-tuning.

formance drop during inference. By discarding the NOISE AGGREGATOR, LLMs can learn knowledge more effectively and surpass the vanilla MoLoRA models with an average of more than 5 points, as shown in Table 4.

Response to RQ2: Yes. Symmetric structure failed to decouple the learning process into the common knowledge and task-specific requirements. To demonstrate this, we tested the models with two parallel LoRA modules to investigate the role of the NOISE AGGREGATOR. As shown in Table 4, the parallel LoRA module failed to decouple the learning process due to its symmetric structure. However, it is surprising that removing one of the FFN LoRA modules can improve the model performance on knowledge-intensive tasks. Similar phenomena are also found in previous work (Jiang et al., 2024a). They found that removing the fine-tuned FFN parameters can fully utilize the general capacity of the base model by producing a more similar hidden state. In the structure of MEDCARE, the KNOWLEDGE AGGREGATOR predominantly acquires common knowledge from the fine-tuning corpus, while the NOISE AGGREGATOR focuses more on learning alignment formats.

Response to RQ3: The KA stage improves the knowledge capacity of the models and the DA stage adapts the models to learn the target formats. We explore the respective role of each fine-tuning stage and empirically validate each merit. Figure 5 shows the results on knowledge examination, CBLUE, and CCTE, respectively. For the first stage, it is obvious that the models with the MKA fine-tuning stage not only achieve better performance on the knowledge examination tasks but also align with the format of the downstream dataset with greater ease. Although discarding NOISE AGGREGATOR after the first stage of fine-tuning reduces the alignment performance of the model, the target format can be learned faster by retraining the model with the second-stage medical adaptation. As more second-stage aligned data is added, the model's alignment capability significantly improves without compromising its knowledge capacity. This suggests that the second stage of the fine-tuning process primarily helps the model align with the format, rather than learning new knowledge.

Model	Knowledge-Intensive					Cross-Lingual Generalization			
	CMMLU	CEval	PLE Pha	PLE TCM	Avg.	MedQA	MedMCQA	MMLU	Avg.
Qwen1.5-7B	**66.77**	68.29	32.29	29.17	49.13	40.46	45.33	61.43	49.07
Huatuo-II-7B	59.08	62.40	47.70	47.50	54.17	40.69	44.75	44.33	43.26
LoRA (Hu et al., 2021)	65.21	60.98	46.67	50.83	55.92	40.53	44.94	61.24	48.90
LoRA+ (Hayou et al., 2024)	63.29	68.29	45.63	45.63	55.71	38.65	45.21	61.69	48.52
MoRA (Jiang et al., 2024b)	65.07	63.41	49.17	47.92	56.39	38.65	45.66	61.11	48.47
DoRA (Liu et al., 2024b)	65.14	**70.73**	50.00	50.83	59.18	39.75	46.02	60.60	48.79
MEDCARE	66.25	**70.73**	**50.83**	**53.33**	**60.29**	**40.96**	**46.98**	**62.27**	**50.07**

Table 5: Comparison with other parameter efficient fine-tuning (PEFT) methods on knowledge-intensive tasks and cross-lingual generalization. The results of PEFT methods are fine-tuned on Qwen1.5-7B. The datasets in cross-lingual generalization are all in English. Note that the 'MedQA' indicates the US subset.

Response to RQ4: Yes. Even with limited data, MEDCARE can enhance its medical knowledge proficiency. To show the effective knowledge learning ability of MEDCARE, we compare the performance of models with varying scales of the first-stage fine-tuning corpus, and present the results in Figure 6. The vanilla MoLoRA suffers from performance degradation when the size of the training-corpus is smaller than 50%, while MEDCARE utilizes the knowledge in the corpus among all sizes of the data and surpasses the performance of the base model consistently, demonstrating that MEDCARE has better knowledge utilization ability.

Response to RQ5: MEDCARE outperforms the commonly used PEFT methods on knowledge-intensive tasks. We choose four PEFT methods to show the effectiveness of MEDCARE, including LoRA (Hu et al., 2021), LoRA+ (Hayou et al., 2024), MoRA (Jiang et al., 2024b), and DoRA (Liu et al., 2024b). It is observed in Table 5 that most of the PEFT methods suffer from performance drops on the medical parts of the general benchmarks CMMLU and CEval. This shows that domain adaptation will destroy the generalization ability of LLM and reduce its ability in scenarios outside the distribution of training data. The more effective methods, like DoRA and MEDCARE, can not only improve the domain capacity but also keep the generalization ability of the LLMs simultaneously, where MEDCARE demonstrates more superiority than the DoRA method.

Response to RQ6: MEDCARE shows strong cross-lingual generalization ability with only fine-tuned with monolingual data. We choose three English medical datasets, MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and the medical part of MMLU (Hendrycks et al., 2021), to validate the cross-lingual capacity of the LLMs. The results in Table 5 show that all the other PEFT methods with only Chinese fine-tuning data achieve worse performance on English medical benchmarks when compared to the base model Qwen1.5-7B. However, with only Chinese fine-tuning data, MEDCARE can still improve the model's performance on the English dataset with 1 point, showing strong cross-lingual ability and applicability.

6 Conclusions

In this paper, we first categorize medical tasks into knowledge-intensive tasks and alignment-required tasks to reinforce the practical usability requirements for medical LLMs. Then, we propose a two-stage fine-tuning framework to adapt LLMs to medical domains and mitigate knowledge performance degradation and propose MEDCARE. The experiment results show that MEDCARE can achieve promising performance on over 20 knowledge-intensive and alignment-required tasks.

Limitations

In this paper, we propose a two-stage fine-tuning framework that mitigates the damage and loss of pre-trained knowledge in large language models during downstream fine-tuning for medical tasks. Despite its effectiveness, our approach still exhibits limitations. Through knowledge-aggregative learning, our method enables large language models to more effectively assimilate the knowledge of the fine-tuning corpus, thereby enhancing the knowledge capabilities of the final model beyond the baseline. However, the alignment fine-tuning in the second stage still adversely affects the model's knowledge capabilities. Furthermore, our method does not yet allow for the decoupling of knowledge and format learning directly, but instead requires two-stage training. Addressing these two issues will be the focus of our future work.

Ethic Considerations

In this article, we introduce a medical large language model, designated as MEDCARE, which incorporates several ethical considerations crucial for its deployment and usage in sensitive settings:

Performance vs. Potential Risks While MEDCARE demonstrates enhancements over previous general and specialized medical models in knowledge reasoning and performance on downstream tasks, it's important to acknowledge the inherent limitations of large language models. Notably, these models can exhibit "hallucinations" or generate misleading information. Additionally, the training datasets might harbor undiscovered biases, which could inadvertently influence the model's outputs. Given these concerns, we emphasize that MEDCARE is not suitable for providing medical advice or for use in direct clinical applications.

Data Ethics and Privacy Compliance The datasets employed in this study, including the knowledge testing data and the CBLUE dataset, are all publicly available and open-source and thus do not pose ethical dilemmas concerning their usage. However, the clinical data from CCTE used in training and testing, which involves report generation, image analysis, and discharge instructions, originates from hospital inpatient records. We have taken stringent measures to ensure the privacy and confidentiality of this information. All personal identifiers have been removed to maintain anonymity, ensuring no individual can be recognized from the data used. During the data collection, patients signed informed consent forms and were fully aware of the data usage methods described in this paper. Additionally, the usage of this data has been reviewed and approved by the corresponding hospital ethics committees. The specific approval numbers will be provided after the end of the review. This ensures that the data usage in this paper fully complies with ethical standards and privacy protection regulations.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2022ZD0162101), the National Natural Science Foundation of China (No. 62106140), and STCSM (No. 21511101100, No. 22DZ2229005)

References

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. https://ai.meta.com/blog/meta-llama-3/. [Accessed 05-06-2024].

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jie Cai, Shenglin Chen, Siyun Guo, Suidong Wang, Lintong Li, Xiaotong Liu, Keming Zheng, Yudong Liu, and Shiling Chen. 2023. Regemr: a natural language processing system to automatically identify premature ovarian decline from chinese electronic medical records. *BMC Medical Informatics and Decision Making*, 23(1):126.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023b. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-LoRAs: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11371–11380, Torino, Italia. ELRA and ICCL.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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)*.

Zan Hongying, Li Wenxin, Zhang Kunli, Ye Yajuan, Chang Baobao, and Sui Zhifang. 2021. Building a pediatric medical corpus: Word segmentation and named entity annotation. In *Chinese Lexical Semantics: 21st Workshop, CLSW 2020, Hong Kong, China, May 28–30, 2020, Revised Selected Papers 21*, pages 652–664. Springer.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Shuyang Jiang, Yusheng Liao, Ya Zhang, Yu Wang, and Yanfeng Wang. 2024a. Taia: Large language models are out-of-distribution data learners. *arXiv preprint arXiv:2405.20192*.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, et al. 2024b. Mora: High-rank updating for parameter-efficient fine-tuning. *arXiv preprint arXiv:2405.12130*.

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.

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. *arXiv preprint arXiv:2402.10373*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *arXiv preprint arXiv:2404.15159*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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. *Cureus*, 15(6).

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 441–459, Online. Association for Computational Linguistics.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023a. Lawgpt: Chinese law large language models. https://github.com/LiuHC0428/LAW_GPT.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024a. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023b. Moelora: An moe-based parameter efficient finetuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024c. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022a. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023c. Gpt understands, too. *AI Open*.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Website. https://openai.com/blog/chatgpt.

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.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2402.13963*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Zhan Su, Fengran Mo, Prayag Tiwari, Benyou Wang, Jian-Yun Nie, and Jakob Grue Simonsen. 2024. Mixture of experts using tensor products. *arXiv preprint arXiv:2405.16671*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9.

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

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med: Chatglm model fine-tuning based on chinese medical knowledge. https://github.com/SCIR-HI/Med-ChatGLM.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023c. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, Singapore. Association for Computational Linguistics.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2023. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv preprint arXiv:2310.00492*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, et al. 2022b. Cblue: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yu Zhao, Yunxin Li, Yuxiang Wu, Baotian Hu, Qingcai Chen, Xiaolong Wang, Yuxin Ding, and Min Zhang. 2022. Medical dialogue response generation with pivotal information recalling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4763–4771.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. Promptcblue: A chinese prompt tuning benchmark for the medical domain. *arXiv preprint arXiv:2310.14151*.

A Related Works

Bilingual medical large language models Large language models such as GPT-4 (OpenAI, 2022), PaLM (Chowdhery et al., 2023) and LLaMA (Touvron et al., 2023a,b) have achieved superior zero-shot performance across tasks and serve as interactive chatbots to interact with humans. However, trained on little medical-oriented data and Chinese data, these LLMs are not useful for medical conversations and consultations, especially for Chinese medical scenarios. Therefore, a lot of work has been done to fine-tune base models on medical data to obtain large medical models. Med-PaLM (Singhal et al., 2022) is grounded on Flan-PaLM (Chung et al., 2022; Chowdhery et al., 2023) to encode clinical knowledge. Med-PaLM2 (Singhal et al., 2023b) as a successor, reduce the gap with doctors by fine-tuning in the medical data and using modern prompting strategies. Apart from grounding on super-large language models such as PaLM or GPT, many works attempt to build medical-LLM on deployable sizes of LLMs, including 7B and 13B. DoctorGLM (Xiong et al., 2023), ChatGLM-Med (Wang et al., 2023b), and Bianque-2 (Chen et al., 2023b) are all built on ChatGLM to support acceptable bilingual medical consultation. Other work like MedicalGPT (Xu, 2023), Huatuo (Wang et al., 2023a) and HuatuoGPT-II (Chen et al., 2023a) are built on LLaMA-series and enlarge the vocabulary to support Chinese conversations.

Parameter efficient fine-tuning Full fine-tuning effectively adapts base large language models to downstream tasks but also consumes significant computational resources with the increasing size of models and the number of tasks. To address this, Parameter-Efficient Fine-Tuning (PEFT) methods have been introduced. These methods freeze the base language models, modifying only a negligible number of parameters during the training phase, yet achieving similar or even superior performance with limited fine-tuning data. Among these methods, Adapter-Tuning (Rebuffi et al., 2017; Houlsby et al., 2019; Lin et al., 2020; Pfeiffer et al., 2021) was the pioneering architecture that connected two additional projection layers to the pretrained language model. In addition to incorporating additional modules, Prefix-tuning (Li and Liang, 2021) introduces learnable prefix tokens and prepends them before the input prompts. Differently, Prompt-Tuning (Lester et al., 2021) utilizes learnable prompt tokens for each task, as a multi-task PEFT approach in NLU scenarios. Following these, P-Tuning (Liu et al., 2023c) and P-Tuning-v2 (Liu et al., 2022b) move away from explicit prompts and employ a prompt-generator to convert pseudo prompts into task prompts, allowing decoder-only models to also perform NLU tasks. However, these methods introduce additional priors and significant inference latency. In contrast, Low-Rank Adaptation (LoRA)(Hu et al., 2022) and its variant, Weight-Decomposed Low-Rank Adaptation (DoRA)(Liu et al., 2024c), take a different approach. LoRA updates original parameters with two low-rank matrices without assuming any specific task or architecture, eliminating inference latency by merging back these two matrices to the original weight. DoRA extends this by incorporating weight decomposition, achieving performance comparable to full fine-tuning. Nonetheless, transferring these methods to multi-task learning scenarios without manual adjustments remains a challenge.

B Experiments Details

B.1 Implementation Detail

For all sizes of MEDCARE, we set the batch size to 128 and fine-tuned the models with 1 epoch for the miscellaneous knowledge aggregation step. The learning rate is 2e-4, with a `warmup_ratio`=0.03 and the cosine learning schedule. The maximum length of the training sample is configured to 3072. We fine-tune the model with MKA stage for 1 epoch and DA stagê for 3 epochs. The orthogonal weight factor $\lambda = 1$. For the configuration of the PEFT, the LoRA rank r and α are fixed at 16 and 32, respectively. The number of the shared experts is set to 2, and the total number of mixture experts is set to 8, with 2 experts activated for each token during the training. We only adopt MoLoRA for the linear layers in the feed-forward network (FFN) blocks and adopt normal LoRA for the linear layers in the attention blocks. All the experiments are conducted on 8×A100 80G GPUs.

B.2 Baseline Models

We selected various models as baselines for comparing their performance on medical tasks. For the general open-sourced models, we choose Baichuan2-7B/13B-Chat (Yang et al., 2023), Qwen1.5-7B/14B-Chat (Bai et al., 2023), ChatGLM2/3-6b (Zeng et al., 2022), LLaMA2-

Type	Task	Description	Size	Metrics
Knowledge-Intensive Tasks	MedQA	Chinese Mainland Medical License Exams (USMLE)	3,425	Accuracy
	MMedBnech	Chinese subset of the Multilingual Medical Benchmark	3,425	Accuracy
	CMB	Comprehensive Multi-level Assessment for Medical Knowledge	11,200	Accuracy
	CMExam	Chinese National Medical Licensing Examination	6,811	Accuracy
	CMMLU†	Chinese Massive Multitask Language Understanding	1,354	Accuracy
	CEval†	A Multi-Level Multi-Discipline Chinese Evaluation	41	Accuracy
	PLE Pharmacy	Pharmacist Licensure Examination Pharmacy track	480	Accuracy
	PLE TCM	Pharmacist Licensure Examination Traditional Chinese Medicine track	480	Accuracy
CCTE	Report Generation	Analysis of Abnormal Indicators in Physical Examination Reports	50	GPT-4 Eval
	Image Analysis	Analysis of the Medical Image Reports	50	GPT-4 Eval
	Discharge Instruction	Providing patients with clear and comprehensive guidance	50	GPT-4 Eval
	Examination Education	Guide students in understanding the thought process behind examination questions	50	GPT-4 Eval
CBLUE	CMeEE	Chinese Medical Named Entity Recognition	500	Micro-F1
	CMeIE	Chinese Medical Text Entity Relationship Extraction	600	Micro-F1
	CHIP-CDN	Clinical Terminology Normalization	600	Micro-F1
	CHIP-CDEE	Clinical Discovery Event Extraction	600	Micro-F1
	IMCS-V2-NER	Intelligent Medical Conversation System Named Entity Recognition	600	Micro-F1
	CHIP-MDCFNPC	Medical Dialog Clinical Findings Positive and Negative Classification	600	Micro-F1
	IMCS-V2-SR	Intelligent Medical Conversation System Symptom Recognition	600	Micro-F1
	IMCS-V2-DAC	Intelligent Medical Conversation System Dialogue Action Classification	800	Macro-F1
	IMCS-V2-MRG	Intelligent Medical Conversation System Medical Report Generation	600	RougeL
	CHIP-CTC	Clinical Trial Criterion	1100	Micro-F1
	CHIP-STS	Semantic Textual Similarity	600	Micro-F1
	KUAKE-IR	Information Retrieval	600	Micro-F1
	KUAKE-QIC	Query Intent Criterion	660	Micro-F1
	KUAKE-QQR	Query Query Relevance	600	Micro-F1
	KUAKE-QTR	Query Title Relevance	600	Micro-F1
	MedDG	Medical Dialog Generation	600	RougeL

Table 6: Statistics of the evaluated medical tasks. "†" indicates that we only choose the questions related to the medical domain for the task.

7B/13B-Chat (Touvron et al., 2023a), and LLaMA3-8B-Instruct (AI, 2024). We choose HuatuoGPT-II (7B/13B) (Chen et al., 2023a) for the medical open-sourced models. Additionally, we also choose the ChatGPT (OpenAI, 2022) as the type of closed-source model with strong performance.

B.3 Fine-tuning Corpus

The corpus used for the MKA fine-tuning stage contains nearly 400k samples in total. For the publicly available parts of data, it contains 80k multiple-choice of the question-answering sample with rational from the training set of the MMedBench (Qiu et al., 2024) and CMExam (Liu et al., 2024a), and 68k samples from the PromptCBLUE (Zhu et al., 2023), which transfer CBLUE (Zhang et al., 2022a) into a pure text format using specific prompts, 30k multi-turn medical conversations sampled from the `HuatuoGPT-sft-data-v1` (Wang et al., 2023a) and 50k single-turn question-answering data generated by GPT-4 from `HuatuoGPT2_sft_instruct_GPT4_50K` (Chen et al., 2023a). The private portion of the data contains about 200k training samples, with each of the four clinical tasks having 50k training samples each.

B.4 Testing Benchmarks

A detailed summary of the descriptions, quantities, and evaluation methods of all test data can be found in Table 6.

Medical Knowledge Exams Medical licensing exams measure large language models' medical knowledge and reasoning capabilities, serving as a common testing method. For this type of Benchmarks, we follow the experiments setting of Chen et al. (2023a). The examination benchmarks include the Chinese Mainland test set of MedQA (Jin et al., 2021), the Chinese test set of MMedBench (Qiu et al., 2024), and two comprehensive Chinese medical exam datasets, validation set of the CMB (Wang et al., 2023d) and the test set of the CMExam (Liu et al., 2024a). We also collect the medical parts of the general benchmarks, including the validation set of the CMMLU (Li et al., 2023a) and CEval (Huang et al., 2023). We also test the performance of the models on the flash exam questions from the 2023 Chinese National Pharmacist Licensure Examination, collected by Chen et al. (2023a).

Medical Alignment Tasks The characteristic of medical alignment Tasks lies in their unique input and output format requirements, primarily evaluating the models' ability to follow the instructions in the clinical scenario. We mainly choose two types of alignment tasks to evaluate the models. The first is CBLUE (Zhang et al., 2022a), a Chinese multi-task medical dataset encompassing 16 distinct medical NLP tasks. The second is the proposed Chinese Clinical Task Evaluation (CCTE), which comprises four clinical tasks: Report Generation, Image Analysis, Discharge Instruction, and Examination Education.

C Chinese Clinical Tasks Evaluation

C.1 Tasks Descriptions

Report Generation In the Report Generation task, there are over 600 different types of test items and more than 3,600 test reports containing various combinations of these test items. The input comprises patient-specific laboratory test report data. The model is tasked with analyzing and identifying which of the patient's test indicators deviate from normal ranges, providing insights into potential underlying causes and offering relevant recommendations for further action. Typically, the number of test items presented in the reports ranges from 2 to 16. To effectively accomplish this task, the model must demonstrate a robust capability for recognizing and interpreting extended contextual information, as well as possess an advanced understanding of medical knowledge. This is crucial for ensuring accurate analysis and the provision of actionable medical advice based on the test results. The data example is shown in Figure 7.

Image Analysis In the Image Analysis task, the primary focus is on generating, analyzing, and diagnosing based on descriptive reports derived from medical imaging. This task involves interpreting detailed reports associated with 14 types of medical image reports, including Rapid Pathology, Endoscopy, Magnetic Resonance Imaging, Electrocardiogram, Computed Tomography, Color Ultrasound, Digital Subtraction Angiography, Computed Radiography, Routine Pathology, Nuclear Medicine, Gastrointestinal Pathology, Endoscopy Examination, Radio Frequency, and Immunohistochemistry Pathology. The model must effectively parse and understand these textual descriptions to identify any noted abnormalities, correlate them with potential medical conditions, and provide diagnostic insights. The data example is shown in Figure 8.

Discharge Instruction In the Discharge Instruction task, the primary objective is to generate comprehensive medical advice and instructions for patients at the time of their discharge based on their admission details, treatment processes, and discharge conditions. This task involves synthesizing information from a patient's entire hospital stay, encompassing initial symptoms, diagnostic findings, treatments administered, and the patient's response to those treatments. The model must adeptly process and integrate this wide array of medical data to formulate clear and precise discharge instructions. These instructions typically include guidelines on medication management, wound care, lifestyle adjustments, follow-up appointments, and signs of potential complications that should prompt immediate medical attention. The data example is shown in Figure 9.

Examination Education The primary focus of the Examination Education task is on enhancing models' ability to explain their decision-making processes, particularly in medical contexts. This task involves evaluating the model's capability to provide detailed explanations and analyses of its responses. This not only aids in increasing the interpretability of large language models in healthcare applications but also serves as a valuable tool for medical students preparing for exams. For the data of the Examination Education, we sample the 50 samples from the test set of the MMedBench (Qiu et al., 2024) for each question in it contains the ground truth rational. The data example is shown in Figure 10.

C.2 Evaluation Method

For the evaluation method of the CCTE dataset, we adopt the GPT4 evaluation to solve the problem that clinical tasks are difficult to evaluate. Following the evaluation pipeline of Wang et al. (2023d), we score each response across four aspects — Fluency, Relevance, Completeness, and Medical Proficiency — using a grading scale from 1 to 5. We input the original data from clinical data for this task, along with the model's responses, into GPT-4 as evaluation criteria to prevent biases in scoring due to medical knowledge hallucinations in GPT-4. Simultaneously, GPT-4 leverages its capabilities to score language metrics such as fluency. The prompt of the GPT-4 is shown in Figure 11.

Report Generation

Question:

下面是一份体检报告结果，请对其中的结果异常的检查项目进行逐项分析。24小时尿蛋白检测+尿电解质\n钠：42.00000 mmol/L\n钾：13.40000 mmol/L\n氯：44.00000 mmol/L\n微量总蛋白：49.00000 mg/L\n微量白蛋白：4.60000 mg/L\n24小时尿钠：126.00000 mmol/24h\n24小时尿钾：40.20000 mmol/24h\n24小时尿氯：132.00000 mmol/24h\n尿量：3000.00000 ml\n24小时尿总蛋白：147.00000 mg/24h\n24小时尿微量白蛋白：13.80000 mg/24h

(**Translations**: Below is a medical examination report. Please analyze each of the abnormal test results.\n24-hour Urine\n Protein Test + Urine Electrolytes\n Sodium: 42.00000 mmol/L\n Potassium: 13.40000 mmol/L\n Chloride: 44.00000 mmol/L\n Trace Total Protein: 49.00000 mg/L\n Trace Albumin: 4.60000 mg/L\n 24-hour Urine Sodium: 126.00000 mmol/24h\n 24-hour Urine Potassium: 40.20000 mmol/24h\n 24-hour Urine Chloride: 132.00000 mmol/24h\n Urine Volume: 3000.00000 ml\n 24-hour Urine Total Protein: 147.00000 mg/24h\n 24-hour Urine Trace Albumin: 13.80000 mg/24h)

Targets:

根据您提供的体检报告结果，下面是对异常项目的分析：

1. **24小时尿钠**:\n - **范围**：130-260 mmol/24h\n - **分析**：您的检测结果为126 mmol/24h，低于正常范围，表明存在尿钠排泄量不足。\n - **潜在原因**：尿钠排泄不足可能由肾脏功能异常、饮食中盐分摄入不足等因素引起。\n - **建议**：建议适量增加饮食中的盐分摄入，同时关注肾脏功能，如有必要可以进行肾功能检查。
2. **24小时尿氯**:\n - **范围**：170-250 mmol/24h\n - **分析**：您的检测结果为132 mmol/24h，明显低于正常范围，表明存在尿氯排泄量不足。\n - **潜在原因**：尿氯排泄不足可能与肾脏功能异常、饮食中氯盐摄入不足等因素有关。\n - **建议**：建议适量增加饮食中氯盐摄入，同时关注肾脏功能，如有必要可以进行肾功能检查。
3. **尿量**:\n - **范围**：1000-2500 ml\n - **分析**：您的尿量为3000 ml，高于正常范围，表明存在多尿情况。\n - **潜在原因**：多尿可能由多种原因引起，如饮水过多、尿糖增加、肾脏功能异常等。\n - **建议**：如果多尿持续存在，建议咨询医生进行进一步检查，以确定具体原因并进行相应治疗。
4. **24小时尿总蛋白**:\n - **范围**：24-141 mg/24h\n - **分析**：您的检测结果为147 mg/24h，略高于正常范围，表明存在轻度蛋白尿。\n - **潜在原因**：蛋白尿可能由肾脏疾病、高血压、糖尿病等引起。\n - **建议**：建议进一步检查肾脏功能，如肾小球滤过率等，以确定蛋白尿的原因并进行相应治疗。

(**Translations**: Based on the results of your medical examination report, here is an analysis of the abnormal items:

1. **24-hour Urine Sodium**:\n - **Range**: 130-260 mmol/24h\n - **Analysis**: Your test result is 126 mmol/24h, which is below the normal range, indicating insufficient urinary sodium excretion.\n - **Potential Causes**: Insufficient urinary sodium may be caused by renal dysfunction, inadequate dietary salt intake, among other factors.\n - **Recommendations**: It is recommended to moderately increase dietary salt intake and monitor renal function. If necessary, a renal function test should be conducted.
2. **24-hour Urine Chloride**:\n - **Range**: 170-250 mmol/24h\n - **Analysis**: Your test result is 132 mmol/24h, significantly below the normal range, indicat Potential Causes:\n - **Recommendations**: It is recommended to moderately increase dietary chloride intake and monitor renal function. If necessary, a renal function test should be conducted.
3. **Urine Volume**:\n - **Range**: 1000-2500 ml\n - **Analysis**: Your urine volume is 3000 ml, above the normal range, indicating a condition of polyuria.\n - **Potential Causes**: Polyuria can be caused by various factors, such as excessive drinking, increased urinary glucose, renal dysfunction, among others.\n - **Recommendations**: If polyuria persists, it is advised to consult a doctor for further examination to determine the specific cause and appropriate treatment.
4. **24-hour Urine Total Protein**:\n - **Range**: 24-141 mg/24h\n - **Analysis**: Your test result is 147 mg/24h, slightly above the normal range, indicating mild proteinuria.\n - **Potential Causes**: Proteinuria may be caused by kidney disease, hypertension, diabetes, among other conditions.\n - **Recommendations**: It is recommended to further examine renal function, such as the glomerular filtration rate, to determine the cause of proteinuria and undertake appropriate treatment.)

Figure 7: Case Example of Report Generation tasks.

Image Analysis

Question:

请根据以下彩超检查报告的内容做出分析和诊断：【颈动脉】双侧颈动脉血管内膜毛糙，管腔通畅，血流充盈好，流速曲线正常范围内，右侧颈动脉分叉处前壁见强回声斑块，大小约4.7*1.4mm。\n【椎动脉】双侧椎动脉血管，管腔通畅，血流充盈好，流速曲线正常范围内，腔内未见明显斑块形成。

(**Translations**: Based on the content of the following color Doppler ultrasound report, an analysis and diagnosis are made: **[Carotid Arteries]** The intima of both carotid arteries is roughened, the lumens are patent, blood flow is well-filled, and the velocity curve is within the normal range. A hyperechoic plaque is observed on the anterior wall at the bifurcation of the right carotid artery, measuring approximately 4.7*1.4 mm. **[Vertebral Arteries]** Both vertebral arteries have patent lumens and good blood filling, with velocity curves within the normal range. No significant plaque formation is observed within the lumen.)

Targets:

据彩超检查报告显示，您的双侧颈动脉内膜存在毛糙，并且在右侧颈动脉分叉处前壁发现了一个大小约为4.7*1.4mm的强回声斑块。另外，双侧椎动脉未见明显异常。根据这些结果，我们可以得出以下诊断结果：您患有双侧颈动脉内膜面毛糙伴右侧斑块形成的情况。\n\n这种情况可能是由于动脉粥样硬化引起的。动脉粥样硬化是一种常见的血管疾病，它会导致血管内膜变得不光滑，并且在血管壁上形成斑块。这些斑块可能会阻碍血液流动，增加心脑血管疾病的风险。\n\n为了控制病情并预防进一步的并发症，我建议您采取以下措施：首先，您可以咨询心血管专科医生，以便进行进一步的评估和治疗。他们可以根据您的具体情况制定个性化的治疗计划。其次，您可以采取一些生活习惯的改变来改善血管健康，如戒烟、限制饮酒、保持健康的饮食习惯和进行适度的体育锻炼。此外，您的医生可能会建议您服用一些药物来控制血压、血脂和血糖水平，以减少动脉粥样硬化的进展。\n\n总之，您的彩超检查结果显示您患有双侧颈动脉内膜面毛糙伴右侧斑块形成的情况。请您尽快就医，并按照医生的建议进行治疗和生活习惯的改变，以控制病情并预防并发症的发生。祝您早日恢复健康！

(**Translations**: According to the color Doppler ultrasound report, roughening of the intimal surfaces in both carotid arteries is observed, and a hyperechoic plaque measuring approximately 4.7*1.4 mm is found on the anterior wall at the bifurcation of the right carotid artery. Additionally, no significant abnormalities are seen in both vertebral arteries. Based on these findings, the following diagnosis can be made: you have roughening of the intimal surfaces of both carotid arteries accompanied by plaque formation on the right side. This condition may be due to atherosclerosis, a common vascular disease that causes the inner lining of the arteries to become rough and plaques to form on the arterial walls. These plaques can obstruct blood flow and increase the risk of cardiovascular and cerebrovascular diseases. To manage the condition and prevent further complications, I recommend the following measures: Firstly, consult a cardiovascular specialist for further evaluation and treatment. They can develop a personalized treatment plan based on your specific condition. Secondly, adopting lifestyle changes to improve vascular health, such as quitting smoking, limiting alcohol intake, maintaining healthy dietary habits, and engaging in moderate physical exercise, can be beneficial. Additionally, your doctor may advise taking medications to control blood pressure, lipid levels, and glucose levels to slow the progression of atherosclerosis. In summary, your ultrasound results indicate you have bilateral carotid intimal roughening with right-sided plaque formation. Please seek medical attention promptly and follow the doctor's recommendations for treatment and lifestyle modifications to manage your condition and prevent complications. Wishing you a swift return to health!

Figure 8: Case Example of Image Analysis tasks.

Discharge Instruction

Question:
入院情况：患者因“反复腹痛2月余”入院；查体：T：35.7˚CP：73次/分R：20次/分BP：157/55mmHg；神志清晰正常面容对答切题构音清晰查体合作全身皮肤及巩膜无黄染未及浅表淋巴结肿大颈软无抵抗气管居中双肺呼吸音清未及啰音。心律齐无杂音。腹平软无压痛、反跳痛肝脾肋下未及双下肢不肿。\n**诊疗过程描述**：患者入院后完善相关检查因粪便中见孢子艰难梭菌及金黄色葡萄球菌予甲硝唑口服抗感染阿泰宁调节肠道菌群制霉菌素漱口。现患者一般情况可予以出院。\n**出院情况**：患者现无腹痛等不适查体：神志清晰正常面容\n**出院时症状与体征**：患者现无腹痛等不适查体：神志清晰正常面容\n\n请根据上述患者的诊断记录给出出院医嘱。
(**Translations**: **Admission Details**: The patient was admitted due to “recurring abdominal pain for over two months”. On examination: Temperature: 35.7°C, Pulse: 73 bpm, Respiration: 20 bpm, Blood Pressure: 157/55 mmHg; the patient was alert with a normal facial expression, responded appropriately, with clear articulation and cooperative during the examination. There was no jaundice of the skin or sclera, no superficial lymphadenopathy, neck was supple without resistance, trachea was central, lung sounds were clear without rales. Heart rhythm was regular without murmurs. The abdomen was soft and flat with no tenderness or rebound pain, and the liver and spleen were not palpable below the ribs. No swelling was noted in the lower limbs. **Treatment Process Description**: After admission, the patient underwent comprehensive tests. Due to the presence of Clostridium difficile spores and Staphylococcus aureus in the stool, the patient was treated orally with metronidazole for infection, atainin to regulate intestinal flora, and antifungal mouthwash. The patient‘s general condition is now stable for discharge. **Discharge Details**: The patient currently has no abdominal pain or other discomforts. On examination: the patient was alert with a normal facial expression. **Symptoms and Signs at Discharge**: The patient currently has no abdominal pain or other discomforts. On examination: the patient was alert with a normal facial expression. Based on the patient's diagnostic record, please provide the discharge instructions.)

Targets:
1. **饮食与休息：**请保证充足的休息，注意营养均衡。饮食上请避免刺激性食物，多食用易消化、富含纤维的食物，如燕麦、熟蔬菜等，以帮助肠道恢复。
2. **症状观察：**如有腹痛、腹泻、血便等不适，请立即前往医院就诊。
3. **随访安排：**请定期复查肛瘘MR报告，待报告至门诊进行复查。同时，发病3个月后，请进行肠镜检查，以评估病情。
4. **定期复查：**请定期复查血常规、肝肾功能、CRP等指标，以监测病情变化。
5. **专科门诊随访：**\n - 呼吸科：请定期随访，进行血液和肺部CT检查，以评估治疗效果。
6. **紧急情况处理：**如遇紧急情况，请立即前往医院急诊科就诊。

请按照上述医嘱进行自我管理，并按时回医院进行复查和随访。祝您早日康复！
(**Translations**:
1. **Diet and Rest:** Ensure adequate rest and maintain a balanced diet. Avoid irritant foods and consume easily digestible, fiber-rich foods such as oats and cooked vegetables to aid in intestinal recovery.
2. **Symptom Monitoring:** If you experience abdominal pain, diarrhea, or bloody stools, please immediately go to the hospital for medical attention.
3. **Follow-up Arrangements:** Please schedule regular reviews of your anal fistula MR report, and follow up at the outpatient clinic when the report is available. Additionally, three months post-onset, please undergo a colonoscopy to assess the condition.
4. **Regular Check-ups:** Regularly review blood routine, liver and kidney functions, and CRP levels to monitor changes in your condition.
5. **Specialist Outpatient Follow-up:** Pulmonology: Please follow up regularly with blood tests and lung CT scans to evaluate the effectiveness of the treatment.
6. **Emergency Handling:** In case of an emergency, please immediately go to the hospital emergency department.

Please manage your condition as per the above instructions and return to the hospital for re-evaluation and follow-up as scheduled. Wishing you a speedy recovery!

Figure 9: Case Example of Discharge Instruction tasks.

Examination Education

Question:
女，25岁。停经7周，1h前突感下腹部疼痛，伴肛门坠胀感。查体：皮肤苍白，下腹压痛反跳痛腹肌紧张，阴道后穹窿饱满，有压痛。该患者最有可能诊断是\nA. 输卵管妊娠\nB. 肠结核\nC. 急性输卵管炎\nD. 急性肠炎\nE. 胃溃疡穿孔
(**Translations:** Female, 25 years old. Amenorrhea for 7 weeks, suddenly experienced lower abdominal pain 1 hour ago, accompanied by a sensation of rectal heaviness. Examination: pallor of the skin, tenderness and rebound tenderness in the lower abdomen, abdominal muscle tension, and a full, tender posterior fornix of the vagina. The most likely diagnosis for this patient is: A. Ectopic pregnancy B. Intestinal tuberculosis C. Acute salpingitis D. Acute enteritis E. Gastric ulcer perforation)

Targets:
育龄期女性，停经7周（输卵管妊娠常有6～8周的停经史），突发下腹痛，伴肛门坠胀感（典型症状为停经后腹痛与阴道流血，输卵管妊娠破裂时突感一侧下腹部撕裂样疼痛），皮肤苍白（输卵管妊娠破裂时疼痛和出血引起），下腹部压痛、反跳痛、腹肌紧张（腹部检查：下腹有明显压痛及反跳痛，尤以患侧为著，但腹肌紧张轻微），阴道后穹隆饱满，有压痛（输卵管妊娠流产或破裂者，阴道后穹隆饱满，有触痛），最可能的诊断是输卵管妊娠（A对）。肠结核（八版内科学P379）（B错）的常见临床表现为腹痛、腹泻、便秘及低热、乏力、盗汗等全身症状，可触及腹部包块。急性输卵管炎（P261）（C错）的常见临床表现为下腹部两侧剧烈疼痛，阴道分泌物增多，有时伴有尿频、尿痛等症状。急性肠炎（D错）的主要症状为恶心、呕吐、腹痛、腹泻、发热。胃溃疡穿孔（八版外科学P355）（E错）突发上腹部剧痛，呈刀割样，腹痛迅速波及全腹，常伴恶心呕吐，面色苍白，血压下降。\n答案为A. 输卵管妊娠.
(**Translations**: A woman of childbearing age with a history of amenorrhea for 7 weeks (ectopic pregnancies often have a 6-8 week history of amenorrhea), experienced sudden lower abdominal pain accompanied by a sensation of rectal heaviness (typical symptoms include post-amenorrhea abdominal pain and vaginal bleeding; a tearing-like pain in one side of the lower abdomen occurs when an ectopic pregnancy ruptures), pallor of the skin (caused by pain and bleeding from the rupture of an ectopic pregnancy), tenderness, rebound tenderness, and abdominal muscle tension in the lower abdomen (abdominal examination shows significant tenderness and rebound pain, especially on the affected side, but slight abdominal muscle tension), and a full, tender posterior fornix of the vagina (in cases of miscarriage or rupture of an ectopic pregnancy, the posterior vaginal fornix is full and tender). The most likely diagnosis is ectopic pregnancy (A correct). Intestinal tuberculosis (B incorrect, as per Internal Medicine, 8th Edition, p.379) typically presents with clinical symptoms such as abdominal pain, diarrhea, constipation, and systemic symptoms like low-grade fever, fatigue, and night sweats, and abdominal masses may be palpable. Acute salpingitis (C incorrect, as per p.261) commonly presents with severe pain on both sides of the lower abdomen, increased vaginal discharge, and sometimes symptoms of urinary frequency and pain. Acute enteritis (D incorrect) mainly presents with nausea, vomiting, abdominal pain, diarrhea, and fever. Gastric ulcer perforation (E incorrect, as per Surgery, 8th Edition, p.355) suddenly causes severe upper abdominal pain, knife-like in quality, with the pain quickly spreading throughout the abdomen, often accompanied by nausea, vomiting, pallor, and a drop in blood pressure.)

Figure 10: Case Example of Examination Education tasks.

Prompt for GPT-4 Evaluation in CCTE
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 description - a question based on the description and conversation - the solution to the question - a model' s answer to the question [description] {description} [end of description] [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}}

Figure 11: The prompt of the GPT-4 evaluation for CCTE tasks.

MODEL	Pharmacist Licensure Examination (Pharmacy)					Pharmacist Licensure Examination (TCM)					AVERAGE
	Optimal Choice	**Matched Selection**	**Integrated Analysis**	**Multiple Choice**	**Total Score**	**Optimal Choice**	**Matched Selection**	**Integrated Analysis**	**Multiple Choice**	**Total Score**	
Llama2-7B	16.88	24.09	15.00	0.00	18.54	15.63	17.27	21.67	2.50	16.04	17.29
Llama2-13B	15.63	26.82	15.00	0.00	19.38	18.13	23.18	21.67	0.00	19.38	19.38
Llama3-8B	41.25	42.72	30.00	15.00	38.33	38.13	30.90	35.00	27.50	33.54	35.94
ChatGLM2-6B	37.00	36.80	25.00	31.70	35.30	33.10	37.30	35.00	37.30	33.70	34.50
ChatGLM3-6B	39.50	39.10	10.50	0.20	34.60	31.80	38.20	25.00	20.00	32.90	33.75
Biachuan2-7B	51.20	50.90	30.00	2.60	44.60	48.10	46.00	35.00	7.50	42.10	43.35
Biachuan2-13B	43.80	52.70	36.70	7.90	44.20	41.30	46.40	43.30	15.00	41.70	42.95
Qwen1.5-7B	29.38	37.73	30.00	17.50	32.29	30.63	34.09	18.33	12.50	29.17	30.73
Qwen1.5-14B	53.75	51.36	46.67	12.50	48.33	40.00	37.73	43.33	27.50	38.33	43.33
ChatGPT	45.60	44.10	36.70	13.20	41.20	34.40	32.30	30.00	15.00	31.20	36.20
HuatuoGPT-II-7B	41.90	61.00	35.00	15.70	47.70	52.50	51.40	41.70	15.00	47.50	47.60
HuatuoGPT-II-13B	47.50	64.10	45.00	23.70	52.90	48.80	61.80	45.00	17.50	51.60	52.25
GPT-4	66.88	65.91	46.67	40.00	61.67	39.38	50.45	45.00	32.50	44.58	53.13
MEDCARE-1.8B	33.75	42.73	18.33	0.25	33.33	40.00	40.91	31.67	12.50	37.08	35.21
↦ *w/o* DA	34.38	44.55	30.00	5.00	36.04	40.00	42.73	45.00	12.50	39.58	37.81
MEDCARE-7B	51.88	53.18	40.00	17.50	48.13	48.75	59.09	51.67	25.00	51.88	50.00
↦ *w/o* DA	50.63	60.91	38.33	15.00	50.83	52.50	62.27	38.33	30.00	53.33	52.08
MEDCARE-14B	53.75	60.91	40.00	20.00	52.50	51.88	58.64	48.33	32.50	52.92	52.71
↦ *w/o* DA	59.38	67.27	36.67	25.00	57.29	**56.88**	57.73	51.67	27.50	**54.17**	**55.73**

Table 7: Results of the 2023 Chinese National Pharmacist Licensure Examination. It consists of two separate Examinations including Pharmacy track and Traditional Chinese Medicine (TCM) Pharmacy track. The results of the baseline models are obtained from Chen et al. (2023a)