Citrus: Leveraging Expert Cognitive Pathways in a Medical Language Model for Advanced Medical Decision Support

Guoxin Wang[1]*, **Minyu Gao**[1], **Shuai Yang**[1], **Ya Zhang**[1], **Lizhi He**[1], **Liang Huang**[1], **Hanlin Xiao**[1]†, **Yexuan Zhang**[1], **Wanyue Li**[1], **Lu Chen**[1], **Jintao Fei**[1], **Xin Li**[1]
[1] Citrus Team, JD Health International Inc.
https://github.com/jdh-algo/Citrus

Abstract

Large language models (LLMs), particularly those with reasoning capabilities, have rapidly advanced in recent years, demonstrating significant potential across a wide range of applications. However, their deployment in healthcare, especially in disease reasoning tasks, is hindered by the challenge of acquiring expert-level cognitive data. In this paper, we introduce Citrus, a medical language model that bridges the gap between clinical expertise and AI reasoning by emulating the cognitive processes of medical experts. The model is trained on a large corpus of simulated expert disease reasoning data, synthesized using a novel approach that accurately captures the decision-making pathways of clinicians. This approach enables Citrus to better simulate the complex reasoning processes involved in diagnosing and treating medical conditions.To further address the lack of publicly available datasets for medical reasoning tasks, we release the last-stage training data, including a custom-built medical diagnostic dialogue dataset. This open-source contribution aims to support further research and development in the field. Evaluations using authoritative benchmarks such as MedQA, covering tasks in medical reasoning and language understanding, show that Citrus achieves superior performance compared to other models of similar size. These results highlight Citrus's potential to significantly enhance medical decision support systems, providing a more accurate and efficient tool for clinical decision-making.

1 Introduction

Recent advancements in the reasoning capabilities of LLMs have become a focal point in research and are increasingly seen as a benchmark for assessing the intelligence level of these models[1, 2]. While the progress in reasoning capabilities has been rapid in domains like mathmetics and programming, the development in healthcare remains relatively limited[3–6]. The open-ended nature of medical practice presents a more complex challenge for Language Models. Medical expertise is cultivated through real-world clinical practice, making it essential for medical reasoning models to learn from the diagnostic and treatment processes of human experts. As a result, emulating the reasoning pathways of medical professionals becomes a crucial step for developing effective medical reasoning models.

Clinical practice, requiring highly sophisticated medical reasoning skills, encompasses patient consultation, diagnosis, differential diagnosis, and treatments[7, 8]. Medical experts have systematically

*Project Lead
†Work done during the internship at Citrus Team.

Preprint. Under review.

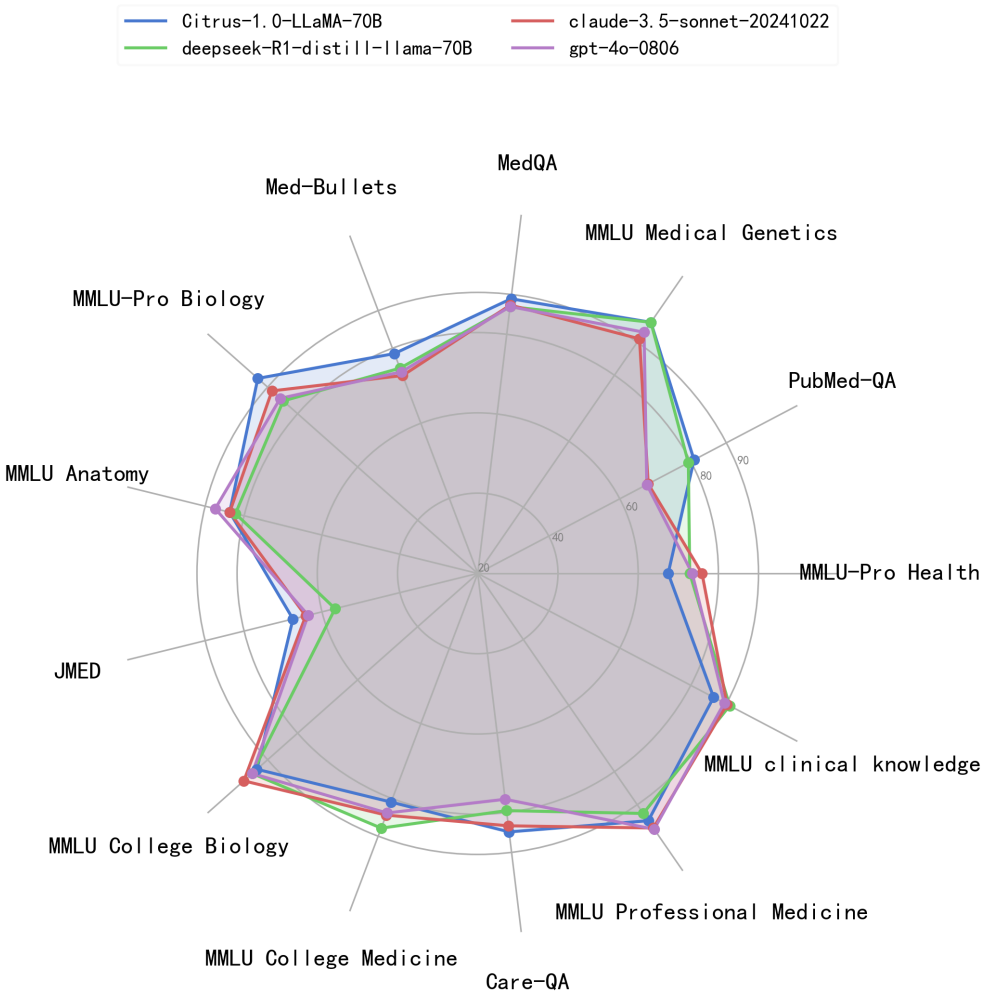


Figure 1: Citrus rank high in several authoritative medical benchmarks, comparing with two widely used LLMs, GPT-4o and Claude and a powerful 70B scale LLM, which is distilled from DeepSeek-R1.

summarized the thought processes involved in clinical practice[9–12]. For medical language models to successfully assist in clinical decision-making, they must not only process vast amounts of medical data but also emulate the complex cognitive processes of expert medical professionals[13]. This requires LLMs to understand not only the explicit medical knowledge but also the implicit reasoning steps that experts use when diagnosing and treating patients. Furthermore, as medical decisions often involve ambiguity, incomplete data, and uncertainty, models must be able to handle these complexities in a way that mirrors expert judgment.

To achieve this, it is necessary for medical language models to be trained on data that closely mirrors the decision-making processes of clinicians. However, obtaining real-world expert-level clinical reasoning data is challenging, as it requires capturing the nuances of expert thought, which are often difficult to quantify. Moreover, existing datasets, although valuable, often fail to replicate the dynamic and often ambiguous nature of clinical practice. In response, new approaches to data synthesis and model training are required to bridge these gaps, enabling models to mimic expert reasoning while also adapting to the complexities and variability inherent in medical practice.

In collaboration with human medical experts, we have designed a methodology for enhancing complex reasoning capabilities in large medical language models, specifically tailored for clinical scenarios. This training-free approach emulates the cognitive reasoning processes of medical professionals, resulting in significant improvements in complex reasoning tasks across multiple models, including

Llama3.1[14] and GPT-4[15]. Inspired by the results, we took further steps and carefully designed a multi-stage training method incorporating multiple phases of continuous pre-training (CPT), supervised fine-tuning (SFT) and reinforcement learning (RL). In this paper, we present Citrus, a medical language model that leverages expert cognitive pathways to simulate the reasoning processes of clinicians. By training Citrus on a large corpus of simulated medical reasoning data, we replicate the dynamic and iterative nature of clinical decision-making. This enables the model to engage in more accurate and effective reasoning, forming the foundation for future AI-driven medical decision support systems. Additionally, we successfully translated this methodology into a trainable approach, resulting in substantial performance improvements in several open-source base models across a variety of medical benchmarks. The advanced medical reasoning and decision-making support provided by Citrus have enabled it to outperform 70B parameter models in the medical domain.

At the same time, we identified a key challenge in current medical language model evaluations: the structured nature of assessment questions often fails to capture the inherent ambiguity of patient symptoms in real-world clinical practice. By leveraging real-world consultations at JD Health's internet hospital, we have created a clinical practice evaluation dataset JDH MEDical Practice Dataset (JMED) that reflects real-world disease distribution, and can be regularly updated.

The contributions of this paper are as follows:

1. We propose a training-free reasoning approach that emulates the cognitive processes of medical experts, enabling large language models to enhance their medical capabilities in clinical diagnosis and treatment.
2. In conjunction with the data construction method, we introduce a multi-stage post-training approach to further improve the model's medical performance.
3. We have made the Citrus model and its training data publicly available as open-source resources to advance research in AI-driven medical decision-making.
4. We have developed and open-sourced a large-scale, updatable clinical practice evaluation dataset based on real-world data, accurately reflecting the distribution of patients in real-world settings.

2 Related Works

Medical reasoning in clinical practice In clinical practice, determining the most rational expert thinking process has always been a key research focus[8–12]. The hypothetico-deductive method[16–18] is a reasoning process from general to specific, which determines diseases based on symptom combinations according to known medical theories. According to this method, some diagnostic hypotheses or conclusions has been raised firstly after collecting information from patients and will be waiting for testing. And these hypotheses, to some extent, guided the subsequent diagnosis and treatment. Pattern-recognition method[19, 20] is a reasoning process from specific to general, which discovers patterns based on clinical observations and empirical summaries. Physicians quickly establish preliminary diagnoses through certain typical descriptions and specific combinations of symptoms that have been repeatedly validated in long-term clinical practice. The dual-processing theory (DPT)[21, 22], which integrates hypothesis testing methodologies with pattern recognition approaches, has gained widespread recognition and acceptance among medical experts. DPT includes system 1 and system 2[23]. System 1 is a fast, intuitive, non-analytical process which is similar to pattern-recognition method, while System 2 is a slow, deliberate, analytical process related to hypothetico-deductive method[9, 24]. DPT posits that the reasoning pathway in clinical practice necessitates the concurrent integration of both intuitive and analytical processes[23, 25, 26].

Application of Large Language Models in Medical Reasoning Researchers have realized the great potential of LLMs reasoning in medical problems solving[27, 28, 3]. Recent advancements in LLMs for healthcare have been propelled by improved training methodologies, including CPT, SFT, and RL, which significantly enhance medical dialogue comprehension and question-answering capabilities[27, 29–48]. Training-free techniques, such as advanced prompt engineering, have enabled general-purpose LLMs to perform specific medical tasks without retraining, as evidenced by studies like MedPrompt[49, 39, 50–55]. The implementation of multi-agent systems simulating experts from various medical departments has improved decision-making and overall medical performance by

supporting complex tasks such as multi-step reasoning and treatment planning[56–59]. Research has highlighted the potential of generating intermediate steps to enhance reasoning abilities, exemplified by OpenAI's O1[1]. Additionally, R1 has demonstrated that training with large-scale synthetic data can yield exceptional reasoning models[2]. Inspired by these innovations, models such as Huatuo-O1[60], O1-Journey[61, 62], and Baichuan-M1[63] have been developed, utilizing inference-time scaling to produce extended reasoning outputs, thereby excelling in diagnostic tasks involving complex medical cases[64]. Huatuo-O1 focuses on advancing the complex reasoning capabilities of LLMs in healthcare by constructing verifiable medical problems and employing medical validators to ensure output accuracy. In contrast, O1-Journey emphasizes enhancing LLMs' ability to handle intricate medical tasks through reasoning augmentation. Baichuan-M1, developed from scratch and specifically optimized for medical applications, is designed to excel in both general domains such as mathematics and programming, as well as specialized medical fields including diagnostic support, medical research, and treatment recommendations. Building on these advancements, our objective is to effectively emulate doctors' cognitive processes in clinical practice to enhance the medical capabilities of large language models.

Evaluation of medical capabilities in large language models Large language models have shown considerable promise in the medical field, and several benchmarks exist to evaluate their capabilities in this domain. Some studies compile medical license examination questions into medical competency assessment datasets, evaluating large language models' medical capabilities in the same way medical students are tested[62]. Some works collect key questions from medical papers, requiring large language models to read medical paper abstracts to answer medical research questions, examining the models' ability to comprehend medical literature[65]. Furthermore, to more accurately assess and differentiate the reasoning abilities of large models, the MMLU-Pro[66] dataset selects more challenging and reasoning-focused questions from MMLU[67] and expands the number of answer choices from four to ten. We aim to combine the advantages of these works to construct a clinical practice evaluation dataset that aligns with the distribution characteristics of real patients and can be regularly updated.

3 Training Data

3.1 Understanding Clinical Reasoning

Recent studies have used the Chain-of-Thought (COT)[68] generation technique to enhance the reasoning capabilities of medical models. We argue that structuring the reasoning process to mirror the cognitive pathways of expert doctors in a structured COT approach is more effective in activating the model's reasoning potential compared to unstructured, base-model-driven processes. Additionally, structured reasoning is easier for human experts to verify. Upon observing the reasoning processes of medical professionals, we identify two primary reasoning methods: the hypothetico-deductive method and the pattern-recognition method. The hypothetico-deductive method involves generating hypotheses based on available information, testing these hypotheses against further data, and revising them to form conclusions. This method emphasizes critical thinking and careful hypothesis testing, making it a robust approach in clinical practice, especially when facing complex, uncertain cases.In contrast, the pattern-recognition method relies on the recognition of patterns or symptoms that closely match previous cases or well-established medical knowledge. This method is often more intuitive and is useful when dealing with familiar or straightforward cases. It involves rapid decision-making based on experience rather than hypothesis testing. Experienced experts typically combine both methods in clinical decision-making to ensure efficient and accurate outcomes.

Leveraging the cognitive pathways of medical experts, we propose a multi-stage data construction methodology that allows LLMs to integrate both reasoning methods, emulating expert reasoning patterns in medical decision-making[37], shown in Figure.2. The main approaches are as follows:

1. Pattern-recognition capabilities are typically developed through CPT. Through exposure to large-scale, high-quality medical datasets, LLMs can most intuitively learn the logical relationships and probability distributions among various medical entities.
2. Hypothetical-deductive reasoning capability requires LLMs to manipulate medical knowledge sophisticatedly. To emulate this complex cognitive pattern, we synthesized extended

COT data by simulating expert reasoning processes. Additionally, a two-stage curriculum learning strategy is implemented as a prerequisite to smooth the model's learning trajectory.

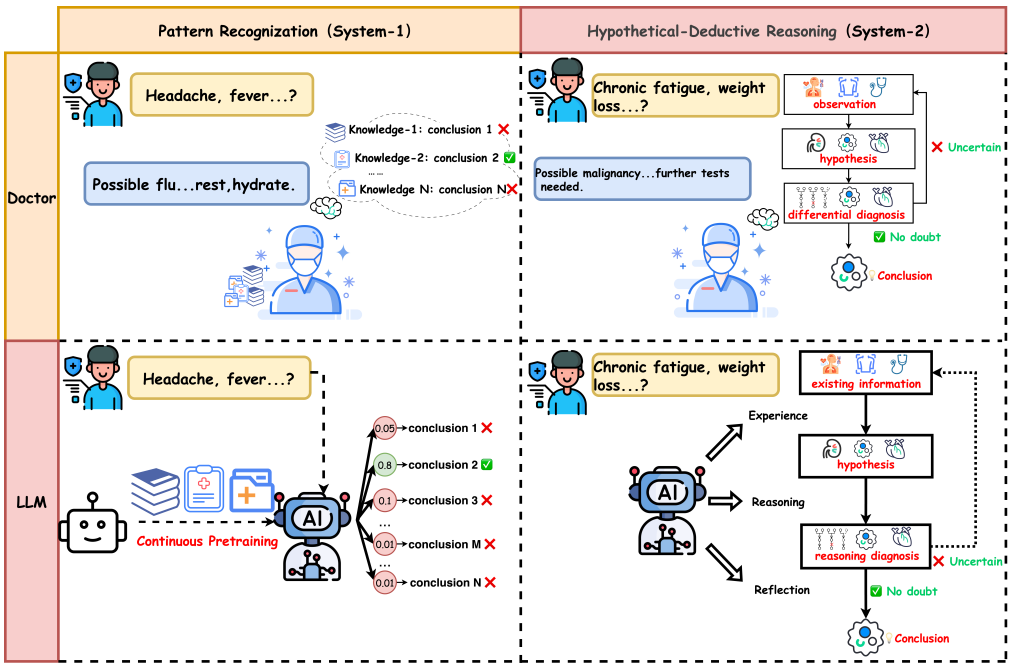


Figure 2: LLMs preforms similar cognitive pathways as medical experts. CPT enabled LLMs to learn medical knowledge and perform pattern recognition as doctors do, meanwhile LLMs are capable to handle hypothetical-deductive reasoning by executing several specific reasoning steps, which can be trained through SFT and RL procedure.

3.2 CPT Data for Pattern Recognition

The pattern-recognition method is typically embedded in the pre-training of large language models and is refined further during domain-specific training. Through the CPT process, a comprehensive medical domain dataset is used to enhance the pattern recognition capabilities of LLMs in addressing medical challenges. By collecting medical field data and preprocessing it, we have obtained a CPT dataset that enables LLMs to learn medical knowledge and perform pattern recognition.

Data Collection The sources of CPT data include the following aspects:

- Web Data
- Medical Textbooks
- Medical guidelines and literature.

Data Process Data sourced from the web requires careful attention to data cleaning and labeling processes. Following the RedPajama[69] approach, we applied natural language processing techniques, such as entity recognition, relationship extraction, and text classification, for data cleaning and labeling. Additionally, we performed deduplication to ensure the quality of the dataset.

For handling PDFs with complex structures, we leveraged certain computer vision solutions. In contrast to web text, data from research papers and medical guidelines presents a different challenge. While this type of data often boasts high-quality content, extracting and transforming it from highly complex structured formats into a form suitable for the CPT training process is particularly challenging. We processed over one million PDF documents using methods such as MAPneo[70].

Regarding medical textbooks, we applied data augmentation techniques to synthesize data. Data derived from medical textbooks inherently represents an optimal training corpus. However, due to

limitations on the number of medical books available for collection, this dataset is smaller compared to web and academic materials. In order to ensure that these high-quality data points have a significant impact during training without being overshadowed by the other two data sources, we utilized techniques such as WizardLM's self-evolution method to diversify and expand the knowledge from individual books into a variety of medical queries[71].

Scale and Distribution To mitigate the performance degradation caused by catastrophic forgetting, we cannot train the model solely on medical corpora. Therefore, we cleaned and selected approximately 200 billion tokens of CPT corpus from the following public datasets (CCI[72], PubMed[73], SlimPajama[74], WuDao[75], ChineseWebText[76], Math Pile[77], Stack Code[78], etc.) and purchased medical book data. After categorizing and labeling the data, we found that medical data accounted for about 30% of the total.

3.3 Data Synthesis for Hypothetico-Deductive Reasoning

The hypothetico-deductive method is typically characterized by the following steps in an expert's thinking process: collecting information, analyzing symptoms, generating diagnostic hypotheses, conducting differential diagnosis, and reaching conclusions. In this process, hypothesis generation and hypothesis testing are the core components of the reasoning process. To model this, we propose a comprehensive data series including general ability sft data with data course and medical ability sft data inspired from training-free dual expert data synthesis system.

3.3.1 General Medical Instruction Data

To enhance the model's fundamental instruction-following capabilities and improve its ease of training, we design a two-stage data course that is not limited to medical problems. We refer to these as Stage-1 and Stage-2 SFT Data[79]. We recognize that it is an impractical task to directly train a base model, which has undergone medical knowledge CPT, to acquire medical reasoning abilities. LLMs struggle to effectively address complex medical reasoning problems when starting with no prior task-handling capabilities. To address this, we adopted a two-stage, general-purpose SFT data approach as part of our data curriculum.

In the first stage, we train the model with approximately 7 million basic instruction examples to improve its ability to follow simple instructions. In the second stage, we use 1.4 million higher-quality and more complex instructions, aiming to enhance the model's multi-turn dialogue handling and complex instruction-following capabilities while preserving the abilities gained from the first stage. This process results in the stage-2 SFT model, which provides a solid foundation for more specialized task training in subsequent phases.

3.3.2 Dual Expert Reasoning Method

In this section, we present the Dual-Expert Reasoning Method. Through this approach, LLMs can emulate medical experts by employing hypothetico-deductive reasoning processes to address medical problems.

To emulate the Hypothetico-Deductive Process , we established a Reasoning Expert. When confronted with a problem, this role analyzes the available information, formulates new hypotheses, and conducts thorough reasoning. During the Training-free experiments, we observed that this process allows considerable flexibility. When the model does not engage in reflection, a significant amount of invalid reasoning processes are generated. This is unacceptable in terms of both reasoning accuracy and training efficiency. To address this, a multi-expert ensemble approach proves to be an effective solution. Thus, we designed a second expert, called the Reflection Expert. The Reflection Expert is tasked with evaluating the reasonableness of the reasoning process and discarding unreasonable or irrelevant steps. We then designed a cognitive flow loop to ensure the model generates a sufficient number of reasonable and accurate reasoning steps:

1. The model lists the existing information as the starting point for reasoning.
2. Based on the existing information, the model proposes possible diagnoses as the endpoints of the reasoning.

3. Perform forward reasoning, attempting to establish the logical path from the starting point to the endpoint.
4. Use the Reflection model to evaluate the validity of the reasoning.
5. Repeat steps 3-4: the model's prior logical path and reasoning feedback should be visible to the model, which will attempt to establish more distinct logical paths. If all reasoning endpoints (diagnoses) have been fully discussed, and then rank the reasoning endpoints (diagnoses). Determine if a diagnosis can be made.
6. If a diagnosis is made, output the result and conclude the reasoning.
7. If a diagnosis cannot be made, return to step 1 and attempt to request external knowledge to gather more information.

This method allows the model to emulate the structured, logical reasoning used by physicians in medical decision-making. By using dual-expert reasinong method, the model can generate multiple possible conclusions, evaluate the validity of each, and progressively refine its reasoning to arrive at the most plausible conclusion. Furthermore, when faced with incomplete or ambiguous information, the model can request external knowledge to assist in making a diagnosis, mimicking the diagnostic approach of medical professionals.

3.3.3 Medical Reasoning Instruction Data

In this section, we describe the construction of Stage-3 SFT Data using the Dual-Expert Reasoning Method. This dataset, called Citrus_S3, designed to improve medical reasoning abilities in LLMs, will be open-sourced to promote further research and development in the field. To ensure accuracy and diversity, we propose several advanced data processing techniques, outlined below.

Reasoning Model with Ground Truth The key to generating reliable medical COT training data using this dual-expert method without additional supervised training is ensuring that the model generates a reasonable and accurate reasoning process. To achieve this, we modified the training-free method by providing the reflection model with the ground truth for the medical questions faced by the reasoning model. In this setup, the reflection model evaluates the reasonableness of the steps generated by the reasoning model and subtly guides it toward the correct answer, without directly providing the solution.This design results in a redefined step 4 in the dual-expert method.

Question Seeds Another indispensable part to successfully execute this data generation procedure is to have extensive medical questions, which should be complicated enough to ignite reasoning process as well as equipped with ground truth that has been properly verified.

Question Rewriting The training question seeds in datasets like MedQA[80] are closed-form questions. We believe that providing answer options limits the model's reasoning capacity, restricting its ability to explore different reasoning paths. To improve generalization, we made the following adjustments:

- We removed the options from the original closed-form questions and converted them into open-ended questions. This allows the model to focus on reasoning and conclusions without predefined answers.
- We created a prompt for rewriting the open-ended questions, removing dependencies on options (e.g., "Which of the following statements is incorrect?").

Question Quality Control Simple medical questions can be answered based on the model's existing knowledge, but they do not require complex medical reasoning. To filter data useful for learning reasoning abilities, we used models such as GPT-4o-2024-0513, Qwen2.5-7B[81], and Llamba-3.1-8B[82] to answer closed-form questions. If these models answered correctly, the data was categorized as easy data, which does not require reasoning. Otherwise, it was categorized as hard data. During the SFT data sampling stage, we used all the hard data and a small portion of easy data to ensure the quality of the training dataset.

Data rewriting Data rewriting is essential to transform multi-expert problem analysis into a first-person thought process. We use LLMs to accomplish this task with several strict constraints:

- Keep thought scale.
- Use narrative words for transition words such as furthermore,therefore, then ,wait...
- Discard duplicated steps
- Keep the ground truth

3.3.4 Data Analysis

DataSet Name	Training Stage	Scale	Field	Construction Method
Web Data	CPT	287B	General	Open-Source
Medical Textbooks	CPT	4.6B tokens	modern medicine	Collect
Medical guidelines and literature	CPT	73B tokens	modern medicine	Collect
Infinity-Instruct-7M	SFT stage 1	7M lines	General& Instruction Fellowing	Open-Source and in-house data
Infinity-Instruct-gen	SFT stage 2	1.4M lines	General& Instruction Fellowing	Open-Source and in-house data
Citrus_S3	SFT stage 3	60K lines	Long COT on Medical Reasoning	Data Synthesis
Citrus_xpo	RL	50K pairs	Long COT on Medical Reasoning	Rejection Sampling

Table 1: Training Data statistics. Our training data is incorporating with CPT data, SFT data and RL data. The table shows the scale, field and construction method of each dataset.

This data synthesis approach, which is shown in Table.1, enables LLMs to generate medical COT data that aligns with medical logic, thereby enhancing their medical capabilities without the need for additional model training. Furthermore, by utilizing data synthesized using the hypothetico-deductive method for model training, the model can acquire medical reasoning capabilities similar to those of doctors.

4 Model Training

In this section, we present a comprehensive training procedure that integrates multiple stages, including CPT, SFT, and RL, referred in Figure.3. Through this multi-phase approach, we aim to transform a base model, initially lacking domain-specific medical knowledge and reasoning abilities, into a robust medical reasoning model capable of performing complex cognitive processes to effectively address and solve clinical problems. The training procedure leverages both general-purpose and medical-specific data, progressively refining the model's ability to handle medical tasks and engage in sophisticated reasoning when confronted with real-world clinical scenarios.

4.1 CPT Stage

This phase focuses on the continuous pre-training of existing foundation models to enhance their comprehension of medical domain knowledge. A primary challenge lies in adapting the same dataset for different foundation models, which possess distinct training data ratios and quality control mechanisms.

In the continuous pre-training of large language models, the ratio of data from different sources is a critical topic. Here, we employ an AutoML approach to dynamically determine the proportion of each data source during training. Specifically, we frame the data ratio problem as a multi-armed bandit problem[83]. We hypothesize that the benefit of encountering previously seen content in continuous pre-training is relatively small, so the model should be encouraged to learn new knowledge. Therefore, we treat the training loss from each data source as a reward. Through this methodology, base models are exposed to training corpora with dynamically adjusted sampling ratios across different training phases, resulting in substantially improved convergence efficiency.

4.2 SFT Stage

4.2.1 Medical Reasoning Ability SFT Training

We propose a three-stage SFT training framework to enhance the model's medical reasoning capabilities. As discussed in Section 3.3, the SFT datasets across these three stages are arranged in ascending

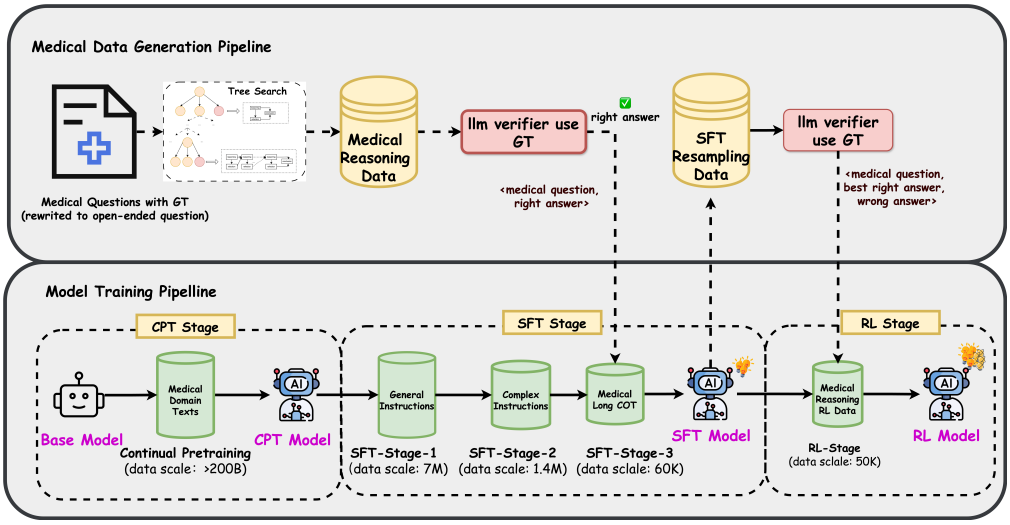


Figure 3: Overview of training stages and training data pipeline . The training process consists of three stages: CPT, SFT, and RL. We shows training purposes and dataset scale on each stage, also, we points out the data pipeline in corresponding stage.

order of difficulty. The underlying rationale is that the model should first master general knowledge application skills before proceeding to complex medical reasoning logic. In the following section, we will focus on elaborating the third stage of SFT training.

The third phase of SFT training focuses on improving the model's performance in the target task domain: medical reasoning. We used data obtained from the Training-Free approach and fine-tuned the Stage-2 SFT model in this phase. The main objective of this phase is to enhance the model's ability to perform long COT in medical reasoning tasks. We used approximately 100,000 medical benchmark problems with GroundTruth gold-standard answers to generate reasoning data, which were used to train the model's medical reasoning capabilities. To maintain the model's general-purpose capabilities during this process, we included reasoning data from other domains, such as logical and mathematical reasoning, in quantities comparable to the medical data. A toekn distribution statistics of stage 3 SFT training data is shown in Figure.4.

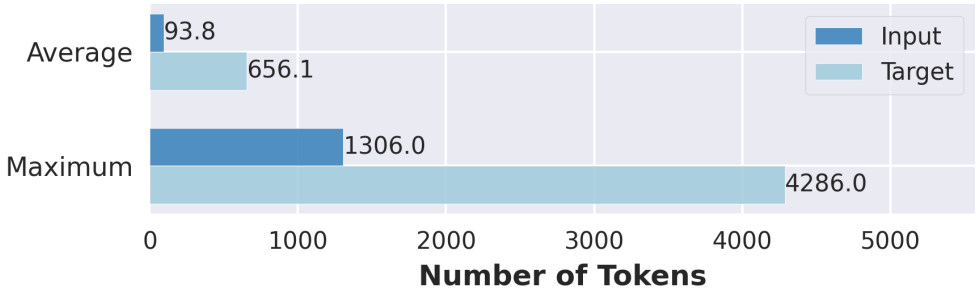


Figure 4: Token distribution statistics of stage 3 SFT training data. The data are designed and manufactured to simulate the long COT reasoning process of medical experts. We confirm a average length of 656 tokens and upper-bound of the length is around 4k.

4.2.2 SFT Data Format

For all SFT data, we follow an unified template format: <sft-input, sft-target>. The sft-input consists of open-ended questions, and we expect the reasoning model to follow a structured thought process before providing answers. The resulting sft-target outputs follow the format: <think>thinking

tokens</think><answer>answer tokens</answer>, where the reasoning process is contained within <think> and the final conclusion is given in <answer>.

4.3 RL Stage

After obtaining the Stage-3 SFT model, an effective reinforcement learning (RL) training phase is necessary. Compared to online methods, the RL techniques used in this phase, such as SIMPO[84] and CPO[85], have distinct advantages. SimPO completely eliminates the dependency on reference models, by directly using the average log probability generated by the policy model as an implicit reward. This not only reduces computational and memory consumption but also simplifies the training process, avoiding the complexity brought by multi-stage optimization. By introducing length normalization, SimPO effectively prevents the model from being biased toward generating lengthy but low-quality responses due to the reward mechanism. However, SimPO is quite sensitive to the learning rate, so we introduced NLL loss, similar to the CPO method, to enhance training stability. These methods offer more stable and efficient learning compared to traditional online reinforcement learning methods. For the RL training, we used data that shares the same origin as that used in Pre-RL SFT, sampling and training on a dataset of approximately 50,000 instances.

4.3.1 RL Data Sampling

We use the best-performing checkpoint after the Stage-3 SFT to perform rejection sampling. The process is as follows:

Repeat Sampling: Open-ended questions (without answer options) are given to the model, which generates 20 responses at a high temperature (temperature = 1.2).

Construct Preference Data: To teach the model reasoning methods instead of just generating reasoning-like statements, we use rule-based rewards based on answer correctness, other than neural reward models. This ensures that rewards are accurately aligned with the correct reasoning steps. Specifically:

- Answer Mapping: For each response, we assess whether the open-ended answer corresponds to the correct option in the original closed-form question. Responses that align with the correct answer are classified as good responses, while others are considered bad responses. We only retain data that contains both good and bad responses.
- Response Scoring: Each response is also scored using GPT-4o. The reasoning process and conclusion are assessed, and the highest-scoring good response is selected as the chosen response. For bad responses, the one with the lowest score is retained if multiple bad responses exist with the same incorrect option.

4.3.2 RL Data Format

For the RL stage, the data format is <RL-input, chosen, rejected>. The RL-input format matches the sft-input format, and chosen and rejected follow the sft-target format.

5 JDH Medical Practice Dataset: Construction and Validation of a Real-World Clinical Dialogue Benchmark

Evaluating medical models is inherently challenging, especially when aligning them with real-world clinical settings. Effective evaluations should ensure that these models can be applied successfully in clinical practice.We systematically analyzed several widely-used medical QA datasets (e.g., MedQA[80], PubMedQA[86], MedMCQA[87], MedBullets[88], MMLU[67], MMLU-Pro[66], and CARE-QA[89]), as shown in Table 2. This analysis revealed three distinctive characteristics: (1) Most datasets are exclusively sourced from medical journal literature or professional medical examinations, with none incorporating real-world hospital data; (2) Question formats primarily consist of multiple-choice questions (MCQs) with 4-5 options, except for MMLU-Pro, which uses a 10-option format. These questions feature clear conditions and fixed options, failing to capture the ambiguity and

limited diagnostic information encountered in real clinical settings; (3) With the exception of CareQA, the remaining datasets lack continuous updates after their initial release.

To address this, we developed the JMED, a novel dataset based on real-world medical data distributions. Unlike existing datasets, JMED closely mimics authentic clinical data while facilitating effective model training. Although based on real consultation data, it is not directly sourced from actual medical data, allowing us to incorporate key elements necessary for model training. We ensured compliance with ethical and legal standards throughout the data collection process, safeguarding privacy and meeting ethical guidelines. Due to the open-ended nature of medical consultations, where definitive answers are often elusive, the evaluation process is more challenging. To address this, each question includes 21 response options, with a "None of the above" choice. This design significantly increases the complexity and difficulty of distinguishing the correct answers, thereby providing a more rigorous assessment framework.

Compared to existing medical QA datasets, JMED has three principal advantages: First, it more accurately reflects the ambiguity in patient symptom descriptions and the dynamic nature of clinical diagnosis in real-world scenarios. Second, the expanded response options require enhanced reasoning capabilities to identify the correct answers among numerous distractors. Additionally, leveraging the vast amount of consultation data from JDH Internet Hospital, we can continuously generate data that aligns with the distribution characteristics of real patients.

	Data Source	**Answer Format**	**Test Dataset Size**	**Released Time**	**Latest Update Time**
MedQA	Examination	4-option MCQs	1,273	2022	2022
PubMedQA	Literature	3-option MCQs	1,000	2019	2019
MedMCQA	Examination	4-option MCQs	4,183	2022	2022
MedBullets	Examination	5-option MCQs	308	2024	2024
MMLU	Examination	4-option MCQs	1,871	2021	2021
MMLU-Pro	Examination	10-option MCQs	818	2024	2024
CareQA	Examination	4-option MCQs	5,410	2020	2024
JMED	Hospital	21-option MCQs	1,000	2025	updatable

Table 2: Comparison of our dataset JMED with existing medical QA datasets. JMED outperforms the most amount of options in MCQs and is the only one based on real-world hospital data. These two factors make JMED more challenging and realistic.

5.1 Data Collection and Construction Pipeline

5.1.1 Raw Data Processing

The dataset originates from anonymized doctor-patient dialogues at JD Health Internet Hospital, filtered to retain consultations adhering to standardized diagnostic workflows. The initial release contains 1,000 high-quality clinical records spanning all age groups (0-90 years) and multiple specialties.

- Privacy Protection: Automated de-identification of sensitive information (names, institutions, locations) via regular expression matching.
- Data Balancing: Ensured statistical representativeness across age, gender, and medical specialties based on platform-wide consultation patterns.
- Deduplication: Applied semantic similarity algorithms to eliminate redundant chief complaints.

5.1.2 Structured Transformation

We constructed a set of multiple-choice questions (MCQs) based on the preprocessed data, as illustrated in Figure 5.

- Electronic Medical Record (EMR) Generation: Extracted key clinical elements using prompt engineering to create structured EMRs.
- Question Formulation: Employed the DeepSeek-r1 model to parse EMRs and generate clinically coherent questions aligned with diagnostic reasoning pathways.

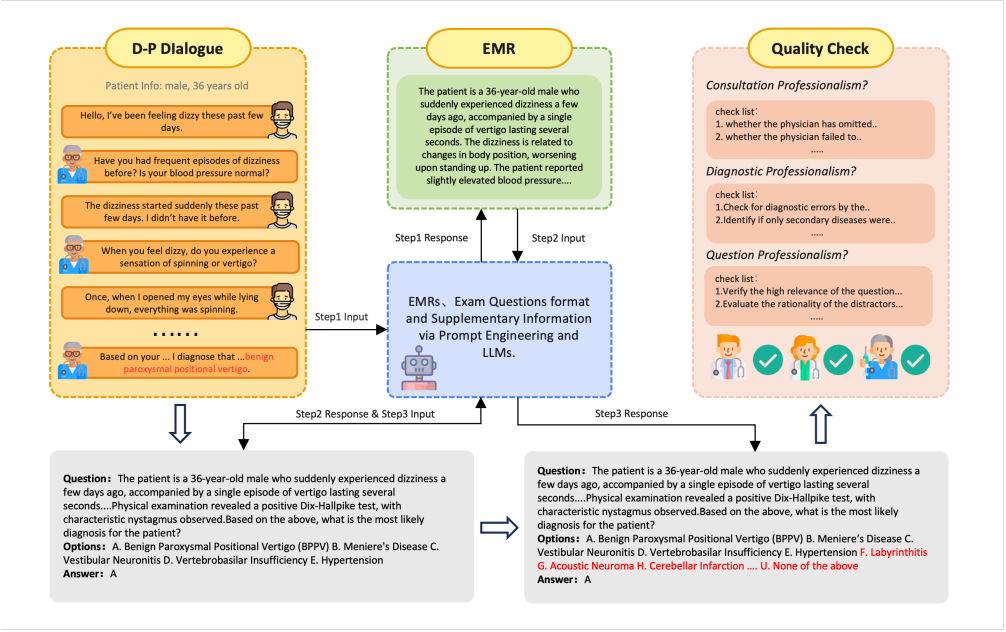


Figure 5: The construction framework of our dataset, JMED, is illustrated with arrows indicating the type of data used as input for the LLMs and the corresponding response obtained at each step. We begin by consolidating the dialogue data into EMRs, then transform it into the format of medical examination questions, and finally, through option expansion and quality checks, we obtain our dataset.

- Option Expansion: Generated 21 mutually exclusive diagnostic options (including standardized ICD-10 terms and plausible differential diagnoses) using LLMs, ensuring compliance with the International Classification of Diseases, 10th Revision (ICD-10).

5.2 Quality Assurance Framework

Considering the seriousness and precision required in the medical field, a three-tier quality control system was established. This primary review process involves collaboration with physicians from 15 departments, with each department having two attending or associate attending-level doctors review the questions. Secondary validation is distributed to associate experienced physicians to conduct a re-evaluation, leveraging their expertise to ensure data quality and accuracy, and final audit is processed by chief physicians. All manual review processes must adhere to the criteria as describe in appendix D.

Based on the aforementioned criteria, we have constructed a set of 1000 multiple-choice questions derived, encompassing multiple departments and age groups. Each data entry includes a unique ID, department, question, options, and the correct answer. The options adhere to the ICD-10 standard for disease nomenclature and have been reviewed and validated by professional physicians to ensure the appropriateness of the questions, options, and correct answers.

5.3 Dataset Characteristics

As shown in Figure 6, the validated subset comprises 1000 clinically rigorous multiple-choice questions with the following demographic distributions:

- Age Coverage: Full spectrum (0-90 years), with 83.37% of cases from the 21-40 age group after outlier adjustment.
- Gender Ratio: 58% male vs. 42% female.
- Specialty Distribution: Covers 15 clinical disciplines.

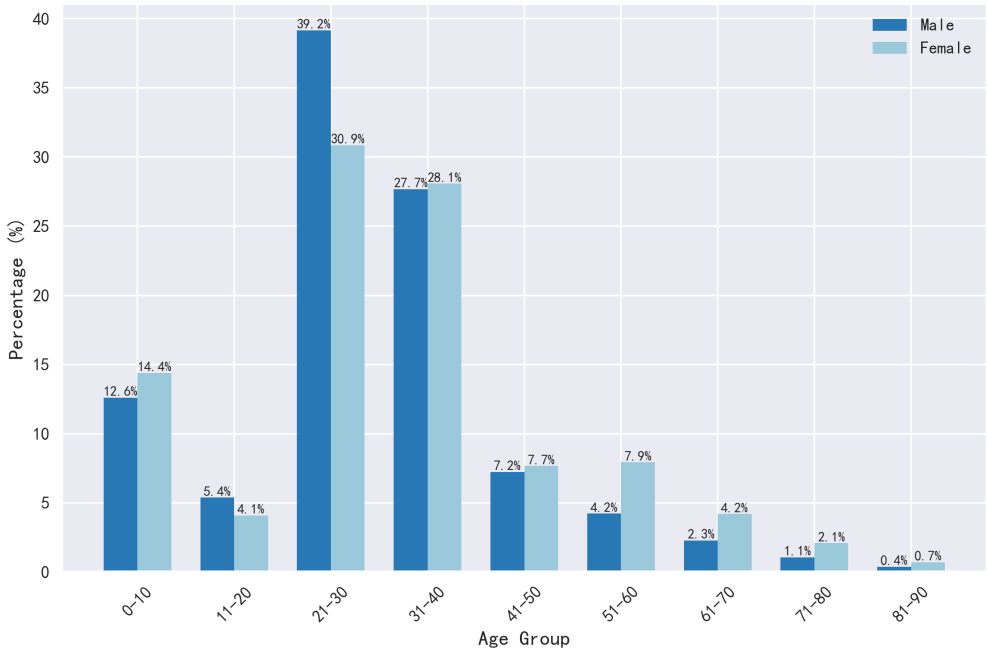


Figure 6: Age distribution statistics of males and females in our dataset. We got the main age distribution of the dataset is 21-40 years old with 83.38% in portion. Meanwhile, we observe an unbalanced distribution in gender, that male patients are 58%.

This distribution aligns with real-world usage patterns of online consultation platforms: adults constitute the primary user base, while consultations for pediatric and elderly patients are typically initiated by caregivers. Statistical analysis confirms that internet-based healthcare platforms have emerged as the preferred channel for adult populations seeking medical services.

6 Experiments

6.1 Experimental Setup

Implementation Details We tested our methodology on Qwen2.5-72B[81], LLama3.1-70B[14] as our base models due to their foundational capabilities. The knowledge capacity of such large-scale models is a prerequisite for stimulating medical reasoning abilities. We use a two-stage SFT to enhance the general capabilities of the model, and in the third stage, we use a small amount of reasoning data to improve the model's medical reasoning ability. In the final stage, we performed reject sampling and alignment on the model to further enhance its reasoning capabilities. We use DeepSpeed ZeRO-3 and Accelerate to train the LLM, with AdamW as the optimizer. The $\beta 1$ and $\beta 2$ of AdamW are 0.9 and 0.95, respectively. We apply a weight decay of 1e-4 and clip the gradient norm to 1.0.

Training Hyperparameters The hyperparameter settings for our model training are shown in the table below. The training parameters for Qwen and LLaMA differ only in the learning rate during the alignment phase. Details are shown in Table 3.

6.2 Benchmarks

We utilized the **MedQA**[80], **PubMedQA**[86], **MedMCQA**[87], **MedBullets**[88], **MMLU**[67], **MMLU-Pro**[66], and **CARE-QA**[89] as benchmarks, with **JMED** serving as a medical reasoning evaluation dataset specifically developed by us.

model name	train phase	learning rate	Batch size	Epochs	Other hyperparameters
Citrus-Qwen-72B	stage1	5e-6	512	2	-
Citrus-Qwen-72B stage1	stage2	5e-6	512	1	-
Citrus-Qwen-72B stage2	stage3	5e-6	512	2	-
Citrus-Qwen-72B stage3	cpo-simpo	1e-6	16	3	$\alpha = 0.05, \beta = 10, \gamma = 5.4$
Citrus-Llama-70B	stage1	5e-6	512	2	-
Citrus-Llama-70B stage1	stage2	5e-6	512	1	-
Citrus-Llama-70B stage2	stage3	5e-6	512	2	-
Citrus-Llama-70B stage3	cpo-simpo	3e-7	16	3	$\alpha = 0.05, \beta = 10, \gamma = 5.4$

Table 3: Model Training Hyperparameter Settings. We show the training hyperparameters for both Qwen2.5-72B and Llama3.1-70B in different stages. We find only a slight difference in the learning rate during the alignment phase since Llama is harder to constrain with such a high learning rate which Qwen is working at.

MedQA dataset is derived from multiple-choice questions of the United States Medical Licensing Examination (USMLE), covering English, Simplified Chinese, and Traditional Chinese. It is designed to evaluate a model's understanding and reasoning ability in medical knowledge.

PubMedQA is a biomedical question-answering dataset collected from PubMed abstracts, containing 1,000 expert-annotated, 61,200 unlabeled, and 211,300 artificially generated QA instances to evaluate a model's understanding and reasoning ability in biomedical research texts.

MedMCQA is sourced from multiple-choice questions in the AIIMS and NEET PG entrance exams. The dataset comprises over 194,000 multiple-choice questions, covering 2,400 healthcare topics and 21 medical subjects. Its purpose is to evaluate and improve models for generating answers to multiple-choice questions in the medical field.

MedBullets is a free learning and collaboration community that offers a large collection of USMLE-style questions and study resources. The dataset includes over 1,000 free USMLE Step 1-style questions, along with extensive study materials. The question type is primarily USMLE Step 1-style multiple-choice questions. Its purpose is to provide a learning and collaboration platform for medical students.

MMLU is a large-scale multitask language understanding benchmark dataset designed to evaluate the knowledge and reasoning abilities of large language models across multiple subjects.

MMLU-Pro is an improved and upgraded version of MMLU, designed to provide more challenging and difficult test questions.

CareQA is a healthcare QA dataset. The dataset originates from official sources of the Spanish Specialized Healthcare Training (FSE) examinations, including the biology, chemistry, medicine, nursing, pharmacology, and psychology tests from 2020 to 2024.

JMED dataset comes from JD Health's online internet hospital and is designed to simulate real clinical data.

6.3 Main Results

We evaluated multiple open-source and closed source LLMs on medical tasks, as shown in Table below.

According to the Main Result Table.4, Citrus1.0-Llama-70B reach a top class performance on 70B scale models, especially on MedQA,PubMedQA,MedBullets,CareQA benchmark. Citrus also surpasses many close-source top LLMs such as Claude-sonnet and GPT-4o. Our model consistently demonstrates strong performance across a wide range of medical benchmarks, highlighting the effectiveness of our proposed approach. Observing the loss curve in Figure.7, it can be seen that the model gradually converges at each SFT stage. In the alignment phase, the reward curve of CPO-SimPO gradually rises and converges. The evaluation results indicate that the performance of the aligned model is the best among all stages.

6.4 Future Discussion

In the ablation experiments, we explore the impacts on different stage of training, including SFT stage 1,2,3 and RL stage. As the most distinguishable and influential benchmarks for medical scenario

	MedQA	PubMed-QA	Care QA	JMED	Med-Bullets	MMLU-Pro Health	MMLU-Pro Biology
LLMs around 70B							
deepseek-R1-distill-llama-70B	0.8696	0.793	0.7952	0.571	0.7468	**0.7286**	**0.848**
Llama3.1-70B-instruct	0.7722	0.793	0.5333	0.559	0.6429	0.6467	0.7978
huatuoGPT-o1-70B	0.835	**0.812**	0.7095	-	0.763	0.7164	0.8382
qwen2.5-72B-instruct	0.7455	0.756	-	0.667	-	0.665	0.834
O1-Journey Learning-llama-70B	0.8648	-	-	-	0.7727	-	-
Citrus1.0-llama-70B	**0.8892**	0.809	**0.8486**	**0.684**	**0.7857**	0.6748	0.8326
LLMs beyond 70B							
claude-3.5-sonnet-20241022	0.8735	0.68	0.8333	0.669	0.7273	0.7592	0.8856
gpt-4o-0513	0.8743	0.697	0.8095	0.668	0.7435	0.7323	0.8577
gpt-4o-0806	0.8696	0.676	0.7667	0.644	0.737	0.7347	0.8577
deepseek-v3	0.7824	0.732	0.7667	0.646	0.6558	0.6993	0.8173
deepseek-R1	0.9097	**0.767**	**0.9123**	**0.751**	0.8149	0.7518	0.8577
gpt-o1-mini	0.8955	0.706	0.6571	0.629	0.8084	0.7213	0.855
gpt-o1-preview	**0.9513**	0.725	0.8714	0.716	**0.8896**	**0.7714**	**0.894**

	MMLU Anatomy	MMLU clinical knowledge	MMLU College Biology	MMLU College Medicine	MMLU Medical Genetics	MMLU Professional Medicine
LLMs around 70B						
deepseek-R1-distill-llama-70B	0.8222	**0.9094**	0.9514	**0.8786**	**0.96**	0.9265
Llama3.1-70B-instruct	0.8148	0.8566	0.9306	0.7861	0.95	0.9118
huatuoGPT-o1-70B	0.837	0.883	**0.9583**	0.8092	**0.96**	**0.9632**
qwen2.5-72B-instruct	**0.8519**	0.8528	0.9306	0.8208	0.89	0.9044
O1-Journey Learning-llama-70B	-	-	-	-	-	-
Citrus1.0-llama-70B	0.837	0.8642	0.9357	0.8092	**0.96**	0.9485
LLMs beyond 70B						
claude-3.5-sonnet-20241022	0.837	0.9019	**0.9792**	0.8439	0.91	0.9706
gpt-4o-0513	0.8889	0.883	0.9653	0.8555	0.96	0.9559
gpt-4o-0806	0.8741	0.8943	0.9514	0.8382	0.93	**0.9743**
deepseek-v3	0.837	0.8868	0.9514	0.8092	0.9	0.9301
deepseek-R1	**0.9259**	**0.9283**	**0.9792**	**0.8844**	**0.98**	0.9596
gpt-o1-mini	0.8074	0.8604	0.9444	0.8439	0.96	0.9596
gpt-o1-preview	0.9185	0.9094	0.9514	0.8728	0.97	0.9706

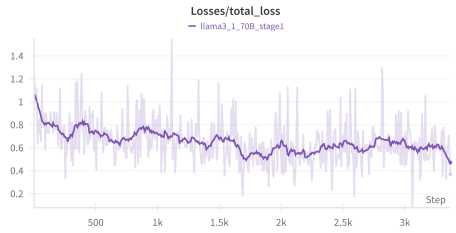
Table 4: Main Results on Medical Benchmarks. LLMs are seperated into 70B scale group and beyond 70B group. Citrus leads most benchmarks among 70B LLMs, moreover, Citrus also surpasses several LLMs beyond 70B on medical benchmarks. **bold** highlights the best scores, and underlines indicate the second-best.

challenges, MedQA is carefully selected as "North star" during our training procedure from the base model all the way to the final one. The results shown in Table.5 provide insights into the importance of each procedure in the training pipeline, discussed as below.

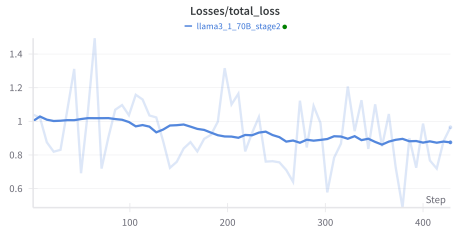
SFT Training Stages The first two stages of SFT primarily focus on grounding the model with general knowledge and reasoning tasks. Training on these stages reach an acceptable level for the model to handling reasoning tasks. The third stage is where the model's medical reasoning capabilities are fine-tuned.This stage's effectiveness is demonstrated in the performance improvements on the MedQA benchmark as the model progresses from 77.06 to 84.13.

SFT data size impact The influence of data size on the model's performance is evident when considering the results from different configurations of Stage 3. Fine-tuning with 20k SFT data yields a performance of 83.12, whereas utilizing 60k SFT data boosts performance to 84.13. However, further increasing the data size to 130k results in a slight performance dip to 83.74, suggesting diminishing returns as the model approaches an optimal configuration for this stage.

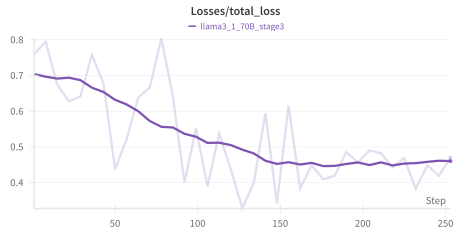
RL Data Proportion We experimented with varying the composition of the rejection sampling data. The most successful configuration involved using 45k medical questions and an additional 5k non-medical questions, resulting in a performance of 88.92. This configuration demonstrates that introducing non-medical question data in other scientific fields into the training process can help



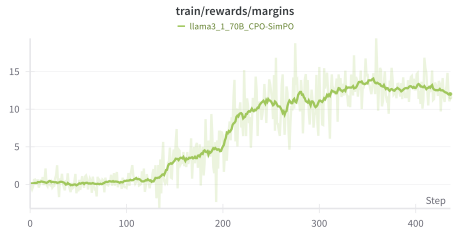
(a) SFT Stage1 Loss



(b) SFT Stage2 Loss



(c) SFT Stage3 Loss



(d) CPO-SimPO Rewards

Figure 7: Training Loss. The figure shows the loss curve of each stage in the training process. The model gradually converges at each SFT stage in (a)(b)(c). In figure7(d), the alignment phase reward curve of CPO-SimPO gradually rises and converges.

	MedQA
Baseline LLMs	
Llama-3.1-70B-base	48.95
Llama-3.1-70B-instruct	78.40
Ours LLMs	
Citrus1.0-Qwen-72B	87.12
Citrus1.0-Llama-70B	88.92
Supervised Finetune (SFT)	
Citrus-Llama-70B stage1/2	77.06
Citrus-Llama-70B stage1/2 + stage3 w/ 20k sft data	83.12
Citrus-Llama-70B stage1/2 + stage3 w/ 60k sft data	<u>84.13</u>
Citrus-Llama-70B stage1/2 + stage3 w/ 130k sft data	83.74
Reinforcement Learning (RL)	
Citrus-Llama-70B-stage3 CPO+SIMPO w/ 17k Rejection Sample data	87.82
Citrus-Llama-70B-stage3 CPO+SIMPO w/ 45k Rejection Sample data	87.20
Citrus-Llama-70B-stage3 CPO+SIMPO w/ 45k + 5k non-med Rejection Sample data	**88.92**

Table 5: The ablation experiments on *Citrus1.0-Llama-70B*. The SFT stage results and RL stage results are shown in sequence to show different contributions to final perform of model from different stages. We also implemented several experiments to reveal the impact from data size and data portion. "w/o" and "w/" denote "without" and "with". **Bold** highlights the best scores in each segment.Use **MedQA** benchmark to evaluate the influence on different training stages and data sizes.

balance the model's understanding of both domain-specific and general reasoning tasks, enhancing overall model performance.

RL Data Size Impact We also explore different RL data size on 17k and 45k. For 45k scale, we use the core, most-difficult 17k data combined with 28k other data, which is not challenging enough for complex medical cognitive task from same sources. The results show that the use of 17k rejection sample data yields a better performance 87.82, which is slightly higher than the model training by

larger dataset improvements. This illustrated that a smaller size is enough for the model to capture the core ability of medical reasoning.

7 Conclusion

We present Citrus, a medical language model designed to enhance medical reasoning by emulating the cognitive processes of medical experts. Through a novel data synthesis approach and a multi-stage training methodology, we have developed a model capable of efficiently handling complex medical decision-making tasks. By releasing the model and its training data, we aim to promote further research in AI-driven medical reasoning and decision-making, thereby contributing to the advancement of healthcare technologies.

Thinking like an Expert We have constructed a medical reasoning dataset modeled from the cognitive processes of doctors, and have effectively demonstrated that such data significantly enhances the problem-solving capabilities of LLMs in medical scenario. Through an exploration of doctorś thought processes, design of experimental data and attempts at model training, we have ultimately developed an LLM capable of effectively leveraging Long COT generated data to address medical issues. From a high-order perspective, we envision that our approach could be widely applicable across domains. By deconstructing the cognitive strategies of experts and utilizing representative core data to generate training data through our approach, models could potentially learn to abstract thinking specific to a given domain. We believe this approach can serve as a comprehensive alternative to human feedback. As a criterion, it effectively replaces the necessity of human feedback in training, allowing the model to understand the underlying characteristics of thinking. Through this understanding, the model can approach generalizable problems from an elevated level of cognitive abstraction, thereby becoming a domain expert.

Complex Training Pipeline We developed a comprehensive multi-phase training pipeline for Citrus, incorporating CPT, SFT, and RL, to enable the model to efficiently learn and adapt to complex medical reasoning tasks. By understanding the problem-solving thought processes of medical experts, we identified the dual-process theory and applied distinct cognitive strategies to various training phases using CPT and SFT. While we believe that extensive pre-training data and clinical examples will help the model perform pattern recognition, there is currently no effective method to equip the model with the complex reasoning abilities that medical experts use to solve problems. We employed a warm-up training phase using data courses and a carefully designed COT data generation method. By training the base model in a specific order, it is gradually enhanced into a medical reasoning model. In the final stage, we claim that offline RL training could further enhance the model's reasoning ability, ultimately ranking it among the top models of similar parameter scales on several authoritative benchmarks.

References

[1] Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024. 1, 4

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 4

[3] Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, et al. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*, 2024. 1, 3

[4] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*, 2022.

[5] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

[6] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022. 1

[7] Andrew B. Symons and Robert H. Seller. *Differential Diagnosis of Common Complaints*. Elsevier - Health Sciences Division, 7th edition, 2017. 1

[8] Geoffrey Norman. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427, 2005. 1, 3

[9] Joy Higgs, Mark A Jones, Stephen Loftus, and Nicole Christensen. *Clinical Reasoning in the Health Professions E-Book: clinical Reasoning in the Health Professions E-Book*. Elsevier Health Sciences, 2008. 2, 3

[10] Linda Adams. Clinical reasoning and causal attribution in medical diagnosis. 2013.

[11] Alan Schwartz and Arthur S Elstein. Clinical reasoning in medicine. *Clinical reasoning in the health professions*, 3:223–234, 2008.

[12] Henk G Schmidt, Geoffrey R Norman, and Henny P Boshuizen. A cognitive perspective on medical expertise: theory and implication [published erratum appears in acad med 1992 apr; 67 (4): 287]. *Academic medicine*, 65(10):611–21, 1990. 2, 3

[13] Sandra M Monteiro and Geoffrey Norman. Diagnostic reasoning: Where we've been, where we're going. *Teaching and learning in medicine*, 25(sup1):S26–S32, 2013. 2

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 13

[15] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[16] Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. Medical problem solving: A ten-year retrospective. *Evaluation & the Health Professions*, 13(1):5–36, 1990. 3

[17] Vimla L Patel and Guy J Groen. Knowledge based solution strategies in medical reasoning. *Cognitive science*, 10(1):91–116, 1986.

[18] Joy Higgs. Developing clinical reasoning competencies. *Physiotherapy*, 78(8):575–581, 1992. 3

[19] Howard S Barrows and Paul J Feltovich. The clinical reasoning process. *Medical education*, 21(2):86–91, 1987. 3

[20] SM Case, DB Swanson, and PL Stillman. Evaluating diagnostic pattern recognition: the psychometric characteristics of a new item format. In *Research in medical education: proceedings of the... annual Conference. Conference on Research in Medical Education*, volume 27, pages 3–8, 1988. 3

[21] Seymour Epstein. Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8):709, 1994. 3

[22] Kenneth R Hammond. *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press, 2000. 3

[23] Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59(1):255–278, 2008. 3

[24] Thierry Pelaccia, Jacques Tardif, Emmanuel Triby, and Bernard Charlin. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical education online*, 16(1):5890, 2011. 3

[25] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011. 3

[26] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013. 3

[27] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023. 3

[28] Pinja Karttunen. Large language models in healthcare decision support. *Tampere University*, 2023. 3

[29] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 3

[30] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[31] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, 2023.

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

[33] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

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

[35] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, et al. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*, 2023.

[36] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376, 2024.

[37] Jiageng Wu, Xiaocong Liu, Minghui Li, Wanxin Li, Zichang Su, Shixu Lin, Lucas Garay, Zhiyun Zhang, Yujie Zhang, Qingcheng Zeng, et al. Clinical text datasets for medical artificial intelligence and large language models—a systematic review. *NEJM AI*, 1(6):AIra2400012, 2024. 4

[38] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025.

[39] 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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. 3

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

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

[42] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

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

[44] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pages arXiv–2305, 2023.

[45] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

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

[47] Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*, 2024.

[48] Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*, 2024. 3

[49] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023. 3

[50] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023. 3

[51] Jordi Bayarri Planas. Prompt engineering for medical foundational models. Master's thesis, Universitat Politècnica de Catalunya, 2024.

[52] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024.

[53] Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason Cheung, Teng Zhang, and Honghan Wu. Knowlab_aimed at mediqa-corr 2024: Chain-of-though (cot) prompting strategies for medical error detection and correction. In *proceedings of the 6th clinical natural language processing workshop*, pages 353–359, 2024.

[54] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

[55] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. 3

[56] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*, 2024. 4

[57] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[58] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

[59] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024. 4

[60] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. 4

[61] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024. 4

[62] Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*, 2025. 4

[63] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025. 4

[64] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023. 4

[65] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019*, 2019. 4

[66] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. 4, 10, 13

[67] 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. 4, 10, 13

[68] 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. 4

[69] Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *Advances in Neural Information Processing Systems*, 37:116462–116492, 2025. 5

[70] Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, et al. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv:2405.19327*, 2024. 5

[71] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023. 6

[72] Liangdong Wang, Bo-Wen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, TengFei Pan, and Guang Liu. Cci3.0-hq: a large-scale chinese dataset of high quality designed for pre-training large language models, 2024. 6

[73] Courtesy of the U.S. National Library of Medicine. pubmed. Hugging Face Datasets, 2024. 6

[74] Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, 2023. 6

[75] Zhao Xue, Hanyu Zhao, Sha Yuan, and Yequan Wang. WuDaoCorpora Text, December 2022. 6

[76] Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model, 2023. 6

[77] Zengzhi Wang, Rui Xia, and Pengfei Liu. Generative ai for math: Part i–mathpile: A billion-token-scale pretraining corpus for math. *arXiv preprint arXiv:2312.17120*, 2023. 6

[78] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024. 6

[79] Beijing Academy of Artificial Intelligence (BAAI). Infinity instruct. *arXiv preprint arXiv:2406.XXXX*, 2024. 6

[80] 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. 7, 10, 13

[81] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 7, 13

[82] Aviv Bick, Tobias Katsch, Nimit Sohoni, Arjun Desai, and Albert Gu. Llamba: Scaling distilled recurrent models for efficient language processing. *arXiv preprint arXiv:2502.14458*, 2025. 7

[83] Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023. 8

[84] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2025. 10

[85] Mohamed Abdel-Basset, Reda Mohamed, and Mohamed Abouhawwash. Crested porcupine optimizer: A new nature-inspired metaheuristic. *Knowledge-Based Systems*, 284:111257, 2024. 10

[86] 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)*, pages 2567–2577, 2019. 10, 13

[87] 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*, pages 248–260. PMLR, 2022. 10, 13

[88] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024. 10, 13

[89] Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. Automatic evaluation of healthcare llms beyond question-answering, 2025. 10, 13

A Ethical Statement

Although the proposed model is a medical LLM with complex reasoning capabilities, it may still produce content that includes hallucinations or inaccuracies. Therefore, the current model is not suitable for real-world applications. Consequently, we will impose strict limitations on the use of our model. The models are not permitted for use in clinical or other industry applications where such inaccuracies could lead to unintended consequences. We emphasize the ethical responsibility of users to adhere to these restrictions in order to safeguard the safety and integrity of their applications.

B Prompt

Here are the prompt examples.

Reasoning Expert Prompt

You are tasked with addressing a medical examination question. Please carefully read the question, provide a detailed thought process, and then present your final answer.

Here is the question:
<Question>
{Q}
</Question>

Facing on the previous question, you are an assistant that engages in extremely thorough, self-questioning reasoning.
Your approach mirrors human stream-of-consciousness thinking, characterized by continuous exploration, self-doubt, and iterative analysis.
With the expectation that when facing this medical issue, you will be able to apply professional medical reasoning methods, such as differential diagnosis, to further reason and think about the problem."

Below is the definition of the differential diagnosis method in medical reasoning:
Differential diagnosis refers to the process of systematically considering different possible diseases, ruling out diagnoses that do not match the condition, and ultimately determining the most likely disease. It involves the following steps:
• Collecting information: Inquire about the medical history, conduct physical exams, and perform necessary laboratory tests.
• Listing possible diagnoses: Based on the medical history, signs, symptoms, and laboratory results, list all possible diseases.
• Gradually eliminating: Through further tests, symptom evaluations, and diagnostic tests, gradually eliminate impossible diagnoses, ultimately confirming the most likely disease.
Differential diagnosis is a process of comparison and contrast, where doctors judge each potential diagnosis based on its characteristics, finding the disease that most closely matches the patient's symptoms and signs.

Please establish the following process in your logical reasoning:
1. List all the known information in the problem, including the complete medical history and all test results.
2. List possible diagnoses.
3. Attempt to build a logical reasoning process.

Below are the reasoning requirements; please ensure each step of the reasoning process meets the following criteria:
(more details are listed in https://github.com/jdh-algo/Citrus)

Your reasoning steps should follow these requirements:
<Reasoning>
[Your extensive internal monologue goes here]
- Begin with small, foundational observations
- Question each step thoroughly
- Show natural thought progression
- Express doubts and uncertainties
- Revise and backtrack if you need to
- Continue until natural resolution
</Reasoning>

Please review the question again:
<Question>
{Q}
</Question>
Please review the output format requirements again; your reasoning steps should be formatted as:
<Reasoning>
[Insert your reasoning step here.]
</Reasoning>
Please follow this output format for the next valid and useful reasoning step:

Reflection Expert Prompt

Please, as a very professional doctor, you will review the thought process of an ordinary doctor regarding a specific medical issue.
You will see the question <Question>, the previously established thought process<Previous Thought> and current reasoning step <Current Reasoning Step>
Most importantly, you know the final answer <Ground Truth>.
Please carefully evaluate, from an objective and professional perspective, whether the doctor's reasoning step is logically sound.
You may think carefully, step by step, and provide rigorous reasoning to argue whether this reasoning step is logically valid.
Finally, you should rate its effectiveness with either 0 or 1, where 0 represents invalid, and 1 represents valid.
Regardless of whether the reasoning is valid or not, please provide your feedback.
If it is valid, please explain the logical reasoning form that is correct and suggest more possible directions for thinking.
If it is invalid, please point out the flaws in the reasoning and provide a revised thought direction.
The feedback should be heuristic, and you may guide them towards the correct answer.
Below are the question and knowledge:
<Question>
{Q}
</Question>
{GT}
Below is the previously established thought process:
<Previous Thought>
{previous_thought}
</Previous Thought>
Below is the current reasoning step:
<Reasoning Step>
{reasoning_step}
</Reasoning Step>
The output format should strictly follow the format below:
<Feedback>
step feedback
</Feedback>
<Rating>
1
</Rating>
Please consider whether the current reasoning step is positively helpful in answering the question, and remember to follow the output format at the end by providing feedback in a concise and precise text form, followed by the rating (0 or 1).
You can include the Ground Truth in the feedback to provide necessary heuristic guidance, but do not mention terms like Ground Truth, Answer, etc., in the feedback.
Please use English for output:

C Standard Inquiry Process

User-submitted consultation requests Users submit information regarding their medical conditions, symptoms, and medical history on the JDH platform. This information is typically submitted in the form of text, images, or videos.

Conversation records between doctors and users After receiving a user's consultation request, doctors engage in real-time text, voice, or video communication with the user. These conversation records contain key information such as the doctor's inquiries about the user's condition, the diagnostic process, and treatment recommendations.

Diagnostic plans and prescriptions Based on the user's condition description and conversation content, doctors provide diagnostic results, treatment plans, and medication prescriptions.

D Quality-Check

Consultation Professionalism

- Assess whether the physician has omitted any critical or non-critical questions during the consultation regarding the patient's chief complaint, current medical history, or past medical history, which could lead to insufficient grounds for the final conclusion.
- Evaluate whether the physician failed to inquire about allergy history or special past medical conditions when recommending antibiotics or other treatments, potentially causing significant harm to the patient's physical and mental well-being.
- Determine whether the physician neglected to routinely collect information on the patient's allergy history, liver and kidney function, or special disease history when advising on medications or products that may cause allergic reactions or other harm. Consultation quality must be comprehensive, detailed, and without any deficiencies.

Diagnostic Professionalism

- Check for diagnostic errors by the physician.
- Identify if only secondary diseases were diagnosed while primary diseases were overlooked.
- Ascertain if the physician provided only a broad diagnosis when the patient's description allowed for a more specific subtype diagnosis.
- Ensure that the diagnostic basis is sufficient. Diagnostic quality must be comprehensive, accurate, and well-supported by evidence.

Question Professionalism

- Verify the high relevance of the question content to the dialogue content.
- Evaluate the rationality of the distractor options and whether they have a hierarchical relationship with the correct option, leading to ambiguity in the answer.
- Ensure the information in the question is sufficient for diagnosis. The quality of the question must meet the standards of being informative, well-evidenced, and having reasonably set options.