CMQCIC-Bench: A Chinese Benchmark for Evaluating Large Language Models in Medical Quality Control Indicator Calculation

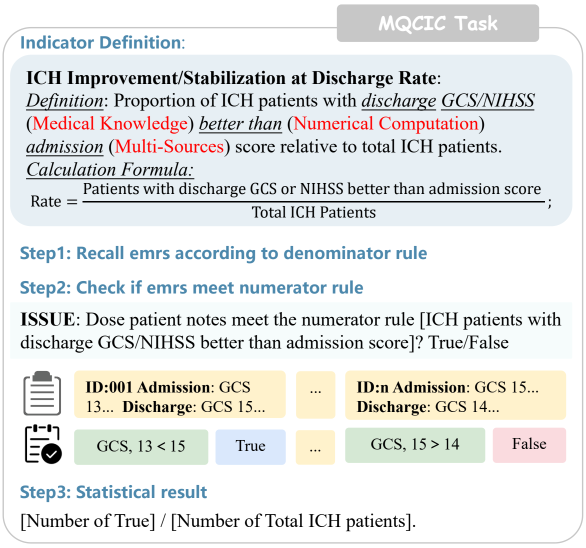
Guangya Yu, Yanhao Li, Zongying Jiang, Yuxiong Jin, Li Dai, Yupian Lin, Ruihui Hou, Weiyan Zhang, Yongqi Fan, Qi Ye*, Jingping Liu, Tong Ruan*
School of Information Science and Engineering,
East China University of Science and Technology, Shanghai, China
guangyayu@mail.ecust.edu.cn, {yeh_qi1125,ruantong}@ecust.edu.cn

Abstract

Medical quality control indicators are essential to assess the qualifications of healthcare institutions for medical services. With the impressive performance of large language models (LLMs) like GPT-4 in the medical field, leveraging these technologies for the **M**edical **Q**uality **C**ontrol **I**ndicator **C**alculation (MQCIC) presents a promising approach. In this work, (1) we introduce a real-world task MQCIC and propose an open-source Chinese electronic medical records (EMRs)-based dataset (CMQCIC-Bench) comprising 785 instances and 76 indicators. (2) We propose a semi-automatic method to enhance the rule representation. Then we propose the Clinical Facts-based Inferential Rule (CF-IR) method that disentangles the clinical fact verification and inferential rule reasoning actions. (3) We conduct comprehensive experiments on 20 representative LLMs, covering general and medical models. Our findings reveal that CF-IR outperforms Chain-of-Thought methods in MQCIC tasks. (4) We conduct an error analysis and investigate the capabilities of clinical fact verification and inferential rule reasoning, providing insights to improve performance in the MQCIC further. The dataset and code is available in this repository [1].

1 Introduction

Medical quality control indicators play an essential role in assessing the performance of healthcare institutions (Øvretveit, 2001; Wang et al., 2018; Anderson et al., 2017). Recently, Large Language Models (LLMs) like GPT-4 (Achiam et al., 2023) have shown promising capabilities in the medical domain. These include applications such as diagnostic reasoning (Dou et al., 2024), clinical note generation (Yang et al., 2023a), and automated clinical assessment (GU et al., 2024). Such capabilities also prove effective in **M**edical **Q**uality **C**ontrol **I**ndicator **C**alculation (MQCIC) (Ye et al., 2025).


MQCIC Task
Indicator Definition:
Step1: Recall emrs according to denominator rule
Step2: Check if emrs meet numerator rule
ID:001 Admission: GCS 13... Discharge: GCS 15...
...
ID:n Admission: GCS 15... Discharge: GCS 14...
GCS, 13 < 15
True
GCS, 15 > 14
False
Step3: Statistical result

ICH Improvement/Stabilization at Discharge Rate:
Definition: Proportion of ICH patients with *discharge* *GCS/NIHSS* (Medical Knowledge) *better than* (Numerical Computation) *admission* (Multi-Sources) score relative to total ICH patients.
Calculation Formula:

$$\text{Rate} = \frac{\text{Patients with discharge GCS or NIHSS better than admission score}}{\text{Total ICH Patients}};$$

ISSUE: Dose patient notes meet the numerator rule [ICH patients with discharge GCS/NIHSS better than admission score]? True/False

[Number of True] / [Number of Total ICH patients].

Figure 1: An example of calculation progress for ICH improvement/stabilization at discharge rate. Firstly collect patient records with intracerebral hemorrhage (ICH). Then identify those with discharge scores better than or equal to admission scores. Finally, the proportion of these patients among all ICH cases is calculated.

Traditionally, calculating quality indicators relied on manually constructed rules (regular expressions) (Tamang et al., 2015; Hsu et al., 2016), which is time-consuming (Ross et al., 2015). As illustrated in Figure 1, the indicator "ICH Improvement/Stabilization at Discharge Rate" contains <*Definition*, *Calculation Formula*>, which requires **(i) medical knowledge** regarding the Glasgow Coma Score(GCS) and NIH Stroke Score(NIHSS), reflecting different patient conditions; **(ii) multiple sources** of information, including both admission and discharge records; and **(iii) numerical computation or logical reasoning**. With such fine-grained rules, experts develop scripts to identify the relevant data from the unstructed text. However, these quality control indicators related to vari-

* Corresponding authors.

[1] https://github.com/YuY-2001/C-MQCIC

Findings of the Association for Computational Linguistics: ACL 2025, pages 609–626
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ous diseases are continually refined and expanded over time. Relying solely on fixed scripts or NLP extraction methods is inadequate and lacks generalizability (Lee et al., 2019; Raju et al., 2015).

Fortunately, LLMs have demonstrated excellent performance in the transformation as well as decomposition of rules (Wang et al., 2023a; Wu et al., 2024; Xu et al., 2024). However, several obstacles remain in developing LLM-based clinical applications (Huang et al., 2024), especially MQCIC: (i) LLMs struggle to provide accurate, reliable answers for complex clinical reasoning tasks, especially when using Chain-of-Thought (CoT) reasoning (Wei et al., 2022). (ii) Concerns over LLMs' reliance on opaque, "black-box" methods for clinical decisions, which may erode user trust. Increasing focus is being placed on improving LLM reasoning to follow explicit logical rules (Servantez et al., 2024a; Yang et al., 2024c; Sun et al., 2024), shifting from context-dependent to transparent, rule-based prompting.

Therefore, this work aimed to explore leveraging explicit rules to achieve automated indicator calculation using LLMs based on electronic medical records (EMRs). Firstly, we introduce a real-world task MQCIC, and propose an open-source dataset, CMQCIC-Bench, derived from Chinese EMRs on an online Chinese website. The dataset comprises 785 instances spanning 76 indicators. Each instance consists of a Patient Note, a Question, and an Answer. We also provide detailed annotations of clinical facts and explanations. Due to the ambiguity of existing rules that impairs the effectiveness of LLMs, we propose a semi-automatic method to enhance the rule representation. With these refined rules, we introduce the Clinical Fact-based Inferential Rule reasoning (CF-IR) method that disentangles the two abilities during the inference stage. We conducted extensive experiments on 20 representative LLMs across general and medical domain. The evaluation results demonstrate that CF-IR outperforms the CoT method. Furthermore, we investigated the capabilities of clinical fact verification and inferential rule reasoning.

In summary, the major contributions are as follows:

- We introduce a clinical scenario task **M**edical **Q**uality **C**ontrol **I**ndicator **C**alculation and propose CMQCIC-Bench, a Chinese open-source dataset with 785 instances, covering 76 different medical quality control indicators.
- We propose a semi-automatic method to enhance the rule representation. Then we propose the **C**linical **F**act-based **I**nferential **R**ule reasoning (CF-IR) method that disentangles the clinical fact verification and inferential rule reasoning actions.
- We conducted comprehensive experiments on 20 representative LLMs, where CF-IR improved performance by 0.43% in the zero-shot setting and 1.45% in the one-shot setting.
- We analyze errors and explore clinical fact verification and rule reasoning, offering insights to improve MQCIC performance.

2 The Medical Quality Control Indicator Calculation Task

Typically, MQCIC involves three steps: (1) Recall relevant EMRs from all cases based on the denominator rules of the indicator. (2) Identify the EMRs that meet the numerator rules from these relevant EMRs. (3) Finally, compute the proportion to determine the indicator's value. The first step can be addressed by matching the ICD-10 codes with diagnostic results. However, the second step is the most challenging, which is the focus of this work. Considering the type of answer is not unique, we define the task as a binary classification problem rather than a cloze task. Thus, the problem is defined as follows: given a Patient Note P and a Question Q related to the indicator's rule, the task of MQCIC is to generate the answer $A = \{True, False\}$.

3 Dataset Construction

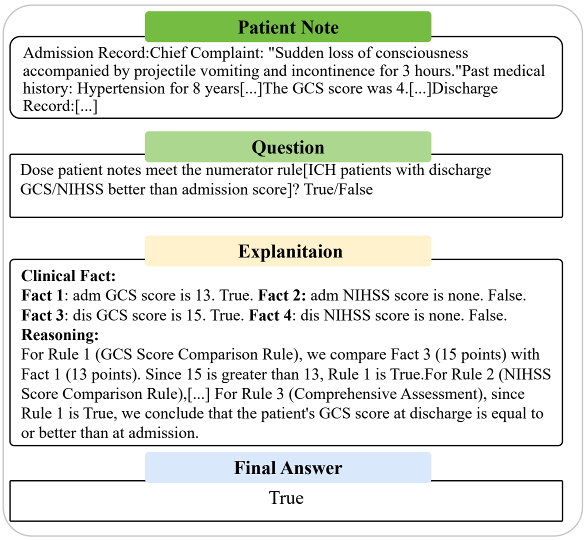
In this section, we construct a dataset, CMQCIC-Bench, for the MQCIC task. The main content includes the **data collection** of indicators and patient notes, **data annotation**, and **data characteristics**.

3.1 Data Collection

We collected indicators and patient notes from two sources. **Indicators Sources.** We manually curated 76 challenging indicators from authoritative documents[2], all developed by experts. For each indicator, a rule-related question was constructed for inclusion in the CMQCIC-Bench. **Patient Notes Sources.** We gathered raw data from a Chinese open-source medical website[3]. Patient notes meeting the denominator rules were filtered based on

[2] http://www.ncis.cn/home

[3] https://www.iiyi.com/



Patient Note

Admission Record:Chief Complaint: "Sudden loss of consciousness accompanied by projectile vomiting and incontinence for 3 hours."Past medical history: Hypertension for 8 years[...]The GCS score was 4.[...]Discharge Record:[...]

Question

Dose patient notes meet the numerator rule[ICH patients with discharge GCS/NIHSS better than admission score]? True/False

Explanitaion

Clinical Fact:
Fact 1: adm GCS score is 13. True. **Fact 2:** adm NIHSS score is none. False.
Fact 3: dis GCS score is 15. True. **Fact 4:** dis NIHSS score is none. False.
Reasoning:
For Rule 1 (GCS Score Comparison Rule), we compare Fact 3 (15 points) with Fact 1 (13 points). Since 15 is greater than 13, Rule 1 is True.For Rule 2 (NIHSS Score Comparison Rule),[...] For Rule 3 (Comprehensive Assessment), since Rule 1 is True, we conclude that the patient's GCS score at discharge is equal to or better than at admission.

Final Answer

True

Figure 2: Example instance of the CMQCIC-Bench dataset.

	CMQCIC-Bench	CMQCIC-Private
Indicators	76	42
Instance	785	314
Avg. L of Note	380.41	520.71
Avg. L of Q.	99.72	113.04
Min Facts	1	1
Max Facts	13	13
Avg. Facts	3.59	4.02
Open-source	Yes	No

Table 1: Statistics of CMQCIC-Bench and CMQCIC-Private datasets. Avg.: average; Q.: question.

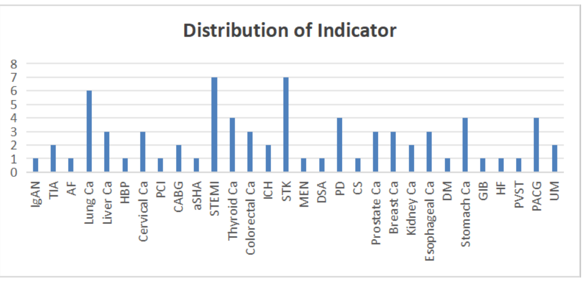


Figure 3: The distribution of indicator in CMQCIC-Bench dataset.

ICD-10 codes and diagnostic findings. Finally, we manually removed patient names, hospital information, and other sensitive data to ensure no privacy risks.

3.2 Data Annotation

Specifically, the annotation process uses the following three-step pipeline. **(1) Clinical fact extraction.** We leverage GPT-4o to extract the original information from EMRs without any modification, then reason based on the context to verify the clinical fact. The clinical facts contain GCS scores, lab exams, medications, etc. **(2) Answer and explanation generation.** Given the obtained facts, for each instance, we leverage GPT-4o to generate the step-by-step explanation through logical reasoning and a final answer $\{True, False\}$. **(3) Data quality control.** Finally, with the guidance of medical experts, annotators are required to check the answer in three facets: fact extraction, logical reasoning, and consistency. Fact extraction and logical reasoning verify the accuracy of the first two stages, while consistency ensures alignment within the patient notes to exclude low-quality cases. In the end, we curated **785** instances for CMQCIC-Bench, as shown in Figure 2, each instance consists of a Patient Note P, a Question Q, a step-by-step $Explanation$, and the final answer A. With the same process, we constructed a CMQCIC-Private dataset derived from patient notes of top-tier tertiary hospitals in China. Ethics committees and experts have rigorously de-identified these data to ensure no privacy leakage risk.

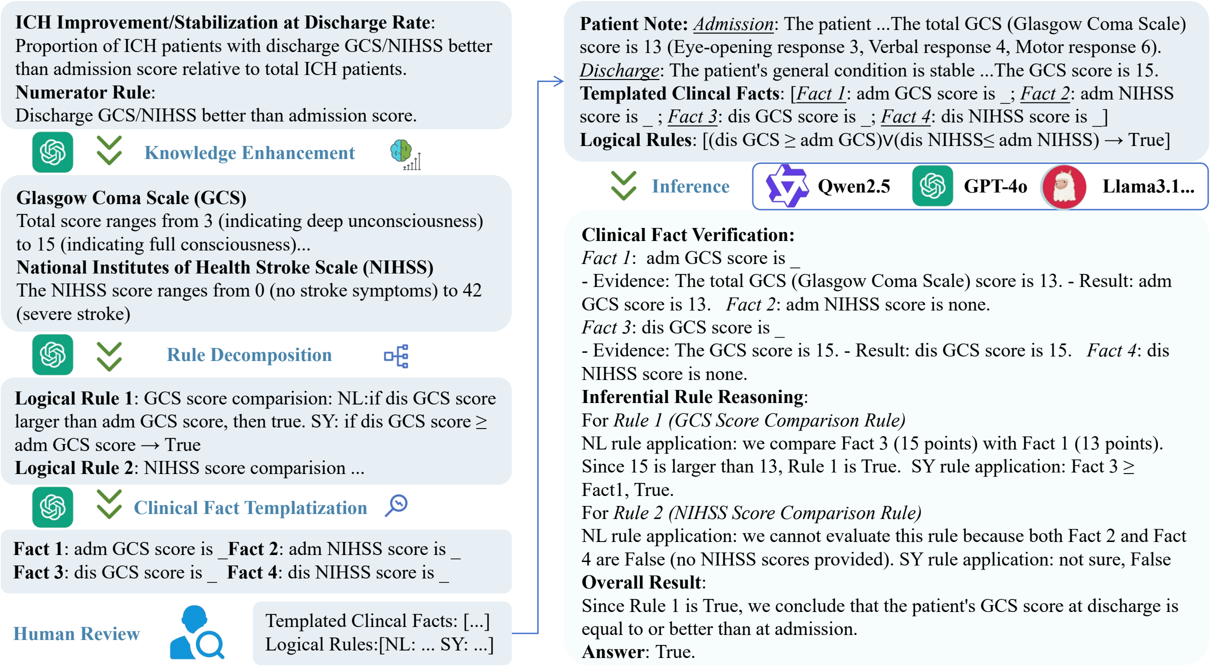
3.3 Data Characteristics

As shown in Table 1, we use Tiktoken[4] to measure sample lengths, yielding average lengths of 380.41 and 520.71, respectively. The shorter average length of each P in CMQCIC-Bench compared to the private dataset stems from the summarized nature of the source data (e.g., lab exams include only key findings). Despite this, the number of facts ranges from 1 to 13, with averages of 3.59 and 4.02, underscoring the task's demand for multi-step reasoning, consistent with real-world scenarios. Additionally, Figure 3 illustrates the indicator distribution, which spans 30 diseases.

4 Method

As shown in Figure 1, the indicators constructed by experts are quite vague, and the underlying medical knowledge can directly affect the implementation of the rules. Therefore, we propose a semi-automatic method that decomposes them into transparent, templated clinical facts and logical rules. We then introduce the Clinical Fact-based Inferential Rule (CF-IR) reasoning pattern for inference. As illustrated in Figure 4, the method comprises two key components: Rule Representation Enhancement and CF-IR. Additionally, we introduce Automatic CF-IR (ACF-IR) to explore

[4] https://github.com/openai/tiktoken



ICH Improvement/Stabilization at Discharge Rate: Proportion of ICH patients with discharge GCS/NIHSS better than admission score relative to total ICH patients.
Numerator Rule:
Discharge GCS/NIHSS better than admission score.

Glasgow Coma Scale (GCS)
Total score ranges from 3 (indicating deep unconsciousness) to 15 (indicating full consciousness)...
National Institutes of Health Stroke Scale (NIHSS)
The NIHSS score ranges from 0 (no stroke symptoms) to 42 (severe stroke)

Logical Rule 1: GCS score comparision: NL:if dis GCS score larger than adm GCS score, then true. SY: if dis GCS score ≥ adm GCS score → True
Logical Rule 2: NIHSS score comparision ...

Fact 1: adm GCS score is _**Fact 2**: adm NIHSS score is _
Fact 3: dis GCS score is _ **Fact 4**: dis NIHSS score is _

Templated Clincal Facts: [...]
Logical Rules:[NL: ... SY: ...]

Patient Note: *Admission*: The patient ...The total GCS (Glasgow Coma Scale) score is 13 (Eye-opening response 3, Verbal response 4, Motor response 6). *Discharge*: The patient's general condition is stable ...The GCS score is 15.
Templated Clincal Facts: [*Fact 1*: adm GCS score is _; *Fact 2*: adm NIHSS score is _ ; *Fact 3*: dis GCS score is _; *Fact 4*: dis NIHSS score is _]
Logical Rules: [(dis GCS ≥ adm GCS)∨(dis NIHSS≤ adm NIHSS) → True]

Clinical Fact Verification:
Fact 1: adm GCS score is _
- Evidence: The total GCS (Glasgow Coma Scale) score is 13. - Result: adm GCS score is 13. *Fact 2*: adm NIHSS score is none.
Fact 3: dis GCS score is _
- Evidence: The GCS score is 15. - Result: dis GCS score is 15. *Fact 4*: dis NIHSS score is none.
Inferential Rule Reasoning:
For *Rule 1 (GCS Score Comparison Rule)*
NL rule application: we compare Fact 3 (15 points) with Fact 1 (13 points). Since 15 is larger than 13, Rule 1 is True. SY rule application: Fact 3 ≥ Fact1, True.
For *Rule 2 (NIHSS Score Comparison Rule)*
NL rule application: we cannot evaluate this rule because both Fact 2 and Fact 4 are False (no NIHSS scores provided). SY rule application: not sure, False
Overall Result:
Since Rule 1 is True, we conclude that the patient's GCS score at discharge is equal to or better than at admission.
Answer: True.

Figure 4: An example overview of our method. On the left is the process of Rule Representation Enhancement, where human experts verify each result. On the right is an example illustrating the CF-IR method.

the automatic performance of rule representation enhancement.

4.1 Rule Representation Enhancement

- **Step 1**: Knowledge Enhancement. We leverage GPT-4o to recall relevant information instead of collecting additional medical books and guidelines. It aims to resolve rule ambiguities arising from the lack of knowledge. For example, the high and low scores of GCS and NIHSS have different meanings, and splitting the rules based solely on semantic information could result in incorrect logical rules.

- **Step 2**: Rule Decomposition. We use GPT-4o to break down complex rules into simpler logical rules in both natural and symbolic language. Symbolic language streamlines natural language by converting it into variables, combined with mathematical and logical symbols to build logical expressions.

- **Step 3**: Clinical Fact Templatization. GPT-4o further extracts the clinical facts involved in the logical rules, which are independent. Each clinical fact with a supposed answer set, such as True/False, numerical value unit, etc.

In the end, we enlisted human experts to review the enhanced rules for the 76 indicators.

4.2 Clinical Fact-based Inferential Rule Reasoning

Motivated by Jin et al. (2024), the model performs two steps during inference: **Clinical Fact Verification** and **Inferential Rule Reasoning**. LLM first extracts and verifies the related information using the templated clinical fact. Then, LLM reasoning on the verified clinical facts with explicit logical rules. We believe that disentangling these two distinct abilities improves performance during the inference process and enhances interpretability. Here are some discussions about the two abilities, and we explore deeply in Section 6.3.

(1) Clinical Fact Verification. Before engaging in clinical reasoning, it is crucial to obtain accurate clinical information(Wang et al., 2024a). However, extracting clinical facts and verifying them through reasoning in a long context with noises of over 380.41 tokens is quite challenging. This includes identifying synonyms, linking symptoms to facts (e.g., high GCS indicates better consciousness), understanding medications (e.g., dual antiplatelet therapy), and analyzing surgical indications, all requiring medical knowledge and clinical reasoning.

(2) Inferential Rule Reasoning. Reasoning can be categorized in multiple ways (Yu et al., 2024). In this study, we specifically define ’Inferential Rule Reasoning’ as the capability to deduce a final

		Zero-Shot			One-Shot		
		Standard	CoT	CF-IR	ACF-IR	CoT	CF-IR
General	MiniCPM3-4B (Hu et al., 2024)	63.31	72.10	68.91	78.98	83.56	82.67
	Internlm2.5-1.8B (Cai et al., 2024)	56.17	56.18	54.14	65.85	68.91	64.07
	Internlm2.5-7B (Cai et al., 2024)	63.31	73.12	79.49	77.07	84.07	84.45
	Internlm2.5-20B (Cai et al., 2024)	69.04	77.57	80.63	83.06	86.36	88.78
	Qwen2.5-0.5B (Yang et al., 2024a)	54.26	56.05	53.88	52.86	61.01	53.63
	Qwen2.5-1.5B (Yang et al., 2024a)	66.11	63.43	62.42	60.50	71.21	73.24
	Qwen2.5-3B (Yang et al., 2024a)	60.38	73.37	67.64	79.36†	77.83	82.03
	Qwen2.5-7B (Yang et al., 2024a)	66.49	82.80	82.92	85.73†	85.22	89.93
	Qwen2.5-14B (Yang et al., 2024a)	78.98	82.03	86.11	84.96	87.89	91.59
	Qwen2.5-32B (Yang et al., 2024a)	75.54	86.49	87.21	92.35†	89.80	94.77
	Qwen2.5-72B (Yang et al., 2024a)	87.77	87.51	92.73	91.33†	90.95	95.54
	llama3.1-8B (Dubey et al., 2024)	48.53	63.05	78.34	73.88	81.52	85.85
	llama3.1-70B (Dubey et al., 2024)	82.54	85.85	85.47	84.45	88.53	91.84
	GPT-4o (Achiam et al., 2023)	77.45	88.91	91.84	90.57	91.59	93.88
Medical	HuatuoGPT2-7B (Chen et al., 2024)	54.01	54.26	49.55	48.66	53.50	56.81
	HuatuoGPT2-14B (Chen et al., 2024)	53.63	55.28	46.36	37.19	52.10	43.31
	Apollo2-0.5B (Zheng et al., 2025)	39.55	41.96	41.14	57.19†	54.39	65.47
	Apollo2-1.5B (Zheng et al., 2025)	53.31	52.03	50.82	52.61	66.11	65.22
	Apollo2-7B (Zheng et al., 2025)	57.57	60.00	61.91	48.91	71.71	65.35
	Apollo-72B (Wang et al., 2024c)	68.91	76.24	72.61	80.63	86.11	86.36
Average		63.84	69.41	**69.71**	71.31	76.62	**77.73**
Human		**95.00**					

Table 2: Aggregated performance (micro-average accuracy) across all indicators on CMQCIC-Bench, using general and medical LLMs. **Bold** denotes the best performance. Underline denotes the second performance. Green denotes the best performance in certain LLM. † denotes ACF-IR outperforms the CoT.

conclusion by applying logical rules to multiple clinical facts. Specifically, for each templated clinical fact, the LLM first performs targeted information extraction and verification. Subsequently, it applies explicit logical rules to conduct structured reasoning based on the verified clinical facts.

5 Experiments

We now describe the experimental setup, empirically evaluate CF-IR, and compare it with existing methods using the CMQCIC-Bench dataset.

5.1 Settings

Here are the details of the models, methods, and evaluation metrics.

Models. We select a representative set of 14 general large language models including GPT-4[5] (Achiam et al., 2023), Qwen2.5 series (Yang et al., 2024a), Internlm2.5 series (Cai et al., 2024), Llama3.1 series (Dubey et al., 2024), MiniCPM3-4B (Hu et al., 2024), as well as 6 medical large language models, comprising the Apollo series (Zheng et al., 2025; Wang et al., 2024c), and HuatuoGPT2 series (Chen et al., 2024). We omit the "chat" and "instruct" in Table 2.

Methods. We assess six prompt-based methodologies: (1) Standard Prompt, utilizing solely the original rules and patient notes; (2) Zero-Shot CoT (Kojima et al., 2022), enhanced with the directive "Let's think step by step"; (3) For CF-IR prompt, we leverage human-reviewd fact templates and logical rules to derive the answer; (4) Specifically, we examine the **One-shot CoT** (Wei et al., 2022); (5) To explore the ability of LLM for rule representation enhancement, we introduce **ACF-IR**, an automated framework that enables LLMs to decompose rules and then conduct CF-IR; (6) Additonally, we set **One-Shot CF-IR**.

For each indicator, we selected an example outside of CMQCIC-Bench. The outputs for these examples were generated by GPT-4o in a zero-shot setting and were carefully reviewed and annotated by human evaluators.

Evaluation. Following prior works (Servantez et al., 2024b; Yang et al., 2024c), the final answer for each instance was labeled as '*True*/*Yes*' or '*False*/*No*', enabling us to use accuracy as the **outcome evaluation** metric. To evaluate **step-wise reasoning**, we compared responses against ground truth using DeepSeek (Liu et al., 2023), assessing

[5] gpt-4o-2024-08-06

Type	Question	Patient Note	Error of LLMs
A	Can the screening for anxiety and depression symptoms be considered complete based on the medical records of hospitalized Parkinson's disease patients?	On the initial visit and the 34th day after treatment, the patient's H-Y stage decreased from stage 3 to 2.5, the HAMD score dropped from 23 to 2, the Beck Depression Inventory (BDI) score decreased from 14 to 4, the PDSS-2 score dropped from 10 to 2,[...].	Based on the provided electronic medical record information, there are no clear mentions of specific results for anxiety and depression symptom screenings. Therefore, we are unable to determine the values for the following facts: Anxiety symptom screening = Not sure Depression symptom screening = Not sure. ##HAMD and BDI can evaluate the depression.
B	Please assess if the patient's coronary intervention meets post-operative success criteria: residual stenosis <20% after stenting or <50% after PTCA, with a TIMI grade of 3.	Treatment Process: [...] Post-stent implantation, there was no significant residual stenosis at the LAD lesion and TIMI grade II.	Based on the provided instructions and medical records, [...] After stent implantation, there was no significant residual stenosis in the patient's left coronary artery LAD. Therefore, the answer is "Yes" ##comment: TIMI grade must be III.

Table 3: Examples of two main error types. evidence, incorrect rationale, and ## comment is colored.

two dimensions: **Fact Faithfulness** (relevance to the original text) and **Fact Correctness** (accuracy of the fact result). Scores range from 0 (irrelevant/incorrect) to 1 (fully relevant/correct). We utilized DeepSeek [6] to extract facts from gold explanations and score model responses accordingly. Scores were averaged per instance to accommodate varying fact counts, resulting in an overall average score. Below is a formal definition of the metrics:

$$FC_i = \frac{\sum_j^m Judge(fact_j, r)}{m}, \quad (1)$$

$$FF_i = \frac{\sum_j^m Judge(fact_j, r)}{m}, \quad (2)$$

where Judge(·) represents the LLMs, outputting 0 or 1. The m denotes the number of facts in the i-th instance. The $fact_j$ denotes the j-th fact of the i-th instance. **Human evaluation**, we designed regex to extract key information, subsequently assessed by experts.

Implement Details. We conduct all experiments on H800 and use VLLM [7] to accelerate for general LLMs. Specifically, we load the medical LLMs directly. Additionally, we set the max_new_tokens = 1024; repetition_penalty = 1.2; temperature = 0.001. The experiments were run three times with random seeds, and the scores were averaged.

5.2 Main Results

Table 2 presents our evaluation results of various LLMs on the CMQCIC-Bench dataset.

(1) Current leading general LLMs perform better than medical LLMs. Qwen2.5-32B/72B-Instruct, and GPT-4o score similarly at 94.77, 95.54, and 93.88, respectively, while medical LLMs lag, with Apollo-72b scoring only 86.36. Only Qwen2.5-72B-Instruct nears human performance, highlighting the ongoing challenge of the MQCIC task for current methods and LLMs.

(2) CF-IR methods perform better than CoT across different parameters and models. In zero-shot and one-shot settings, the average score of CF-IR improves by 0.43% and 1.45%, respectively, compared to CoT. Unlike the CoT method, which performs reasoning along random paths, our approach integrates explicit logical rules with verifiable facts, enhancing the stability and interpretability of LLMs. While CF-IR demonstrated strong performance across various parameters in the one-shot setting, we observed that in the zero-shot scenario, CF-IR outperformed CoT only on general models with parameters $\geq$ 7B. We will analyze our improvement in Section 6.

(3) One-Shot setting can bring significant improvements. In general, after providing the examples, CoT and CF-IR achieved improvements of 10.38% and 11.50%, respectively, the performance of all models showed significant improvements in the one-shot setting for both the CoT and CF-IR methods except HuatuoGPT2. This may stem from HuatuoGPT's fine-tuning data being predominantly centered around QA tasks (Chen et al., 2024), without incorporating clinical scenarios, and weak in instruction-following.

(4) Automated rule representation enhancement remains challenging. While CF-IR achieves strong performance (**77.73**) with enhanced rule representation, ACF-IR's automated approach scores lower (71.31), underperforming CoT. Notably, only Apollo2-0.5B and specific Qwen2.5 variants (3B,

[6] https://www.deepseek.com/

[7] https://github.com/vllm-project/vllm

	Zero-Shot		One-Shot	
Error Type	CoT	CF-IR	CoT	CF-IR
clinical fact	0.19	0.23	0.17↓	0.17↓
reasoning	0.11	0.07	0.07↓	0.05↓
other	0.00	0.01	0.00-	0.00↓
Total	0.31	0.30	0.23↓	0.22↓

Table 4: Error type distribution of LLMs on CMQCIC-Bench. Arrows represent the changes from zero-shot to one-shot. We averaged all the models' performances.

7B, 32B, 72B) surpass CoT in one-shot settings, revealing the limitations of intrinsic model planning capabilities (Servantez et al., 2024b; Yang et al., 2024b). A promising direction involves leveraging advanced open-source models (e.g., GPT-4o) or specialized plan training (Wu et al., 2024) for rule decomposition, complemented by medical models for inference.

6 Empirical Analysis and Discussion

In this section, we analyze errors and evaluate step-wise reasoning, further exploring clinical fact verification and inferential rule reasoning capabilities.

6.1 Error Analysis

Firstly, we categorize errors into three types: Type A, B and C, representing errors in clinical fact verification, inferential rule reasoning, and other types, respectively. We display the example of two main error types in CMQCIC-Bench in Table 3.

Building on prior work (Khandekar et al., 2025), we employ DeepSeek to classify error types by comparing LLM outputs with ground truth in CMQCIC-Bench, facilitating a granular error analysis across LLMs. Since incorrect clinical facts can propagate and affect inferential rule reasoning, we focus on identifying the earliest error type. A manual review of 200 randomly sampled DeepSeek-annotated errors confirmed an 87% accuracy, validating our approach for analyzing error types in all CF-IR responses. As shown in Table 4, providing demonstrations reduces Type A and B errors, highlighting the value of exemplars. While CF-IR does not mitigate clinical fact verification errors, it significantly improves reasoning accuracy due to its structured logical framework. Further details are available in Appendix E.1 (Tables 11 and 12).

Models	Methods	FC	FF	ACC
Qwen2.5-72b	zero-shot CoT	68.09	68.07	87.51
	zero-shot CF-IR	76.34	76.83	92.73
	one-shot CoT	69.26	69.92	90.95
	one-shot CF-IR	90.45	86.20	95.54
Qwen2.5-32b	zero-shot CoT	66.63	66.42	86.49
	zero-shot CF-IR	72.86	71.53	87.21
	one-shot CoT	69.73	70.44	89.80
	one-shot CF-IR	84.61	76.78	94.77
Qwen2.5-14b	zero-shot CoT	68.25	66.44	82.03
	zero-shot CF-IR	70.87	65.14	86.11
	one-shot CoT	71.32	69.41	87.89
	one-shot CF-IR	83.48	78.22	91.59
Qwen2.5-7b	zero-shot CoT	67.31	63.36	82.80
	zero-shot CF-IR	67.03	65.76	82.92
	one-shot CoT	66.60	66.45	85.22
	one-shot CF-IR	77.97	71.88	89.93
llama3.1-70b	zero-shot CoT	65.24	62.49	85.85
	zero-shot CF-IR	71.70	65.23	85.47
	one-shot CoT	69.41	68.59	88.53
	one-shot CF-IR	83.15	77.50	91.84
llama3.1-8b	zero-shot CoT	57.02	57.52	63.05
	zero-shot CF-IR	64.89	61.25	78.34
	one-shot CoT	68.13	65.61	81.52
	one-shot CF-IR	78.40	70.38	85.85
Average		72.03	69.22	86.41

Table 5: Comparison of step-wise and outcome evaluation. **FC** denotes Fact Correctness. **FF** denotes Fact Faithfulness. The results of **ACC** sourced from Table 2.

6.2 Evaluation on Step-Wise Reasoning

As shown in Table 5, the step-aware evaluation metrics decreased by 14.38 and 17.19 points, respectively, compared to the outcome evaluation results. This suggests that the model often makes clinical fact verification errors during the reasoning process, even when the final result is correct.

6.3 Analysis on Clinical Fact Verification and Inferential Rule Abilities

While the CF-IR method enhances inference performance, we further investigate its two core capabilities: **Clinical Fact Verification** and **Inferential Rule Reasoning**. For **Clinical Fact Verification**, we define the input as <*Patient Note*, *Templated Clinical Fact*, *Question*> and the output as <*Reasoning*, *Final Answer*>, evaluated using **Fact Faithfulness** and **Fact Correctness**. Unlike step-wise reasoning, we test each fact independently to avoid contextual interference. For **Inferential Rule Reasoning**, providing verified facts as input to minimize errors, the input is <*Verified Clinical Facts*, *Logical Rules*>, and the output is <*Explanation*, *Final Answer*>, evaluated using labels like '*True*/*Yes*' or '*False*/*No*' for both natural (**NL**) and symbolic (**SY**) languages. All experiments are conducted in a zero-shot setting. Additional results are available

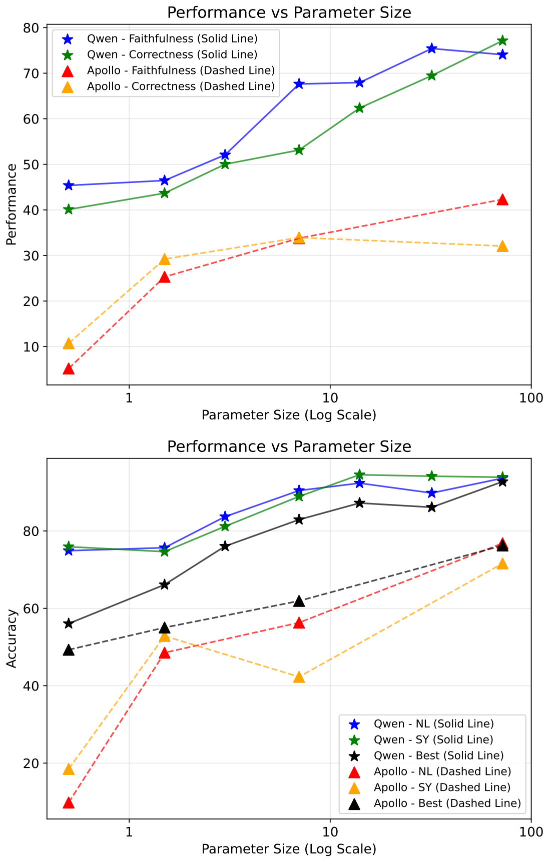


Figure 5: Clinical fact verification and inferential rule reasoning abilities of Qwen and Apollo series on CMQCIC-Bench. NL denotes natural language; SY denotes symbolic language.

in Table 13 in Appendix E.2.

As shown in Figures 5, we find that: **(1)** Both Qwen and Apollo exhibit performance-scale correlations across capabilities. **(2)** Fact verification performance significantly declines, consistent with Table 5. **(3)** For inferential reasoning, Qwen performs comparably in natural and symbolic settings, while Apollo shows stronger natural language robustness. **(4)** With correct facts, Qwen surpasses previous best results (standard, CoT, CF-IR) in zero-shot settings, whereas Apollo underperforms, likely due to Qwen's extensive logical reasoning training. See Appendix E.2 for additional results.

6.4 Benefit of fine-tuning

We fine-tuned the Qwen2.5-3B-Instruct model using LoRA (Hu et al., 2021) for 3 epochs on the CMQCIC-Bench, with evaluation on CMQCIC-Private. The data format follows ACF-IR: Input: <*Instruction*, *Patient Note*, *Rule*>; Output: <*Knowledge*, *Templated Clinical Facts*, *Logical Rules*, *Clinical Fact Verification*, *Inferential Rule Reasoning*, *Final Answer*>. As depicted in Figure 6, the fine-tuned 3B model achieves comparable or superior performance to the 7B model in real-world scenarios. In zero-shot settings, it demonstrates significant improvements across all methods, with gains of 6.7, 8.3, 14.3, and 13.7, confirming the feasibility of distillation rule enhancement for smaller models. More details are provided in Appendix B.

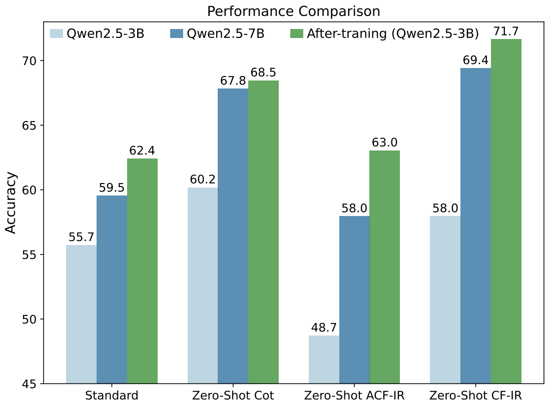


Figure 6: Analysis of fine-tuning benefit. Performance of three models on the CMQCIC-Private dataset.

7 Related Work

7.1 Rule-based LLM reasoning

While reasoning demonstrated a fundamental capability of LLM on applications (Li et al., 2024), there are many research such as CoT (Wei et al., 2022), CoT-sc (Wang et al., 2023b), ToT (Yao et al., 2024), etc. However, there are more attention on rule-enhanced methods (Yang et al., 2023b; Sun et al., 2023; Mu et al., 2023). Reasoning based on facts and deriving answers from logical rules is referred to as inferential rule following ability (Sun et al., 2024). Leveraging such ability that integrating explicit rules with LLMs has gained significant attention. For instance, Servantez et al. (2024b) utilized the IRAC framework to tackle legal tasks with LLMs, emphasizing the application of legal rules. Additionally, Wang et al. (2024b) proposed a neurosymbolic framework for multi-step rule application. Despite the current limitations of LLMs in rule-based reasoning (Yang et al., 2024c), our work demonstrates that such rule-based reasoning outperforms CoT reasoning in the MQCIC task.

7.2 LLM Evaluations in Clinical Scenarios

While LLMs have shown impressive capabilities in medical knowledge recall and reading comprehension on medical exams (Nori et al., 2023; Subramanian et al., 2024), their effectiveness in real-world clinical applications remains a critical area of evaluation. For example, Ouyang et al. (2024) assesses LLMs across 14 expert-curated clinical scenarios, including diagnosis, discharge summaries, and medical consultations. Similarly, Khandekar et al. (2025) introduces MedCal-Bench, a benchmark designed to evaluate inferential rule reasoning in medical contexts, while Hou et al. (2024) simulates a multi-step diagnostic process to test clinical reasoning capabilities. Furthermore, Munnangi et al. (2024); Chung et al. (2025) explore LLMs' abilities in clinical fact decomposition and verification. In this work, we focus on evaluating LLMs in the MQCIC task, with a specific emphasis on their performance in clinical fact verification and inferential rule reasoning, providing a detailed analysis of these two critical abilities.

8 Conclusion

In this work, we present MQCIC, a novel task, and CMQCIC-Bench, an open-source dataset derived from Chinese EMRs. We propose a semi-automatic approach to refine rule representation and introduce CF-IR, a disentangled inference method. Experimental results show that CF-IR surpasses CoT in performance. Error analysis reveals enhanced capabilities in clinical fact verification and inferential rule reasoning. Additionally, we evaluate step-wise reasoning and conduct a detailed investigation of the two abilities. Our work aims to advance the application of LLMs in MQCIC tasks and offers deeper insights into these essential capabilities.

Limitations and Future Work

While we construct a CMQCIC-Bench dataset and evaluate LLMs' clinical fact verification and inferential rule reasoning abilities, several limitations can be improved. (1) Due to the difficulty of manually verifying each sample, our dataset only contains 785 instances. (2) We have only located a comprehensive Chinese document on medical quality control indicators. As a result, our dataset consists solely of Chinese EMRs, and we are also leaning toward selecting Chinese LLMs for our analysis. (3) While we observed a significant improvement in model performance with the one-shot demonstration, benchmarking the model with few-shot instances could have further enhanced accuracy, a scenario we did not test. (4) Although we propose the CF-IR method, which performs well across various LLMs with an enhanced rule representation reviewed by humans, decomposing the rules with a smaller LLM that lacks strong planning capabilities remains a challenge.

Ethical Consideration

The medical cases are sourced from the iiyi website, where doctors voluntarily contribute and share their information. The data is explicitly authorized for use in research and educational activities. To safeguard patient privacy, our dataset excludes any personally identifiable details, such as patient names, hospital information, or other sensitive data. As a result, there is no risk of privacy violations related to our dataset. Furthermore, all data usage adheres to ethical guidelines and regulations governing medical information and research.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (No. 62306112), Shanghai Sailing Program (No. 23YF1409400), and Shanghai Pilot Program for Basic Research (No. 22TQ1400100-20).

References

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

Joseph C Anderson, Lynn F Butterly, Julia E Weiss, and Christina M Robinson. 2017. Providing data for serrated polyp detection rate benchmarks: an analysis of the new hampshire colonoscopy registry. *Gastrointestinal endoscopy*, 85(6):1188–1194.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Song Dingjie, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *First Conference on Language Modeling*.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2024. Integrating physician diagnostic logic into large language models: Preference learning from process feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2453–2473.

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

ZHANZHONG GU, Wenjing Jia, Massimo Piccardi, and Ping Yu. 2024. Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought. *Available at SSRN 4835586*.

Ruihui Hou, Shencheng Chen, Yongqi Fan, Guangya Yu, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: A benchmark for evaluating large language models in multi-step clinical diagnosis. *Preprint*, arXiv:2408.10039.

William Hsu, Simon X Han, Corey W Arnold, Alex AT Bui, and Dieter R Enzmann. 2016. A data-driven approach for quality assessment of radiologic interpretations. *Journal of the American Medical Informatics Association*, 23(e1):e152–e156.

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*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chaochao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Yining Huang, Keke Tang, and Meilian Chen. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2024. Disentangling memory and reasoning ability in large language models. *arXiv preprint arXiv:2411.13504*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew Zhang, et al. 2025. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jeffrey K Lee, Christopher D Jensen, Theodore R Levin, Ann G Zauber, Chyke A Doubeni, Wei K Zhao, and Douglas A Corley. 2019. Accurate identification of colonoscopy quality and polyp findings using natural language processing. *Journal of clinical gastroenterology*, 53(1):e25–e30.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, et al. 2024. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. 2023. Can llms follow simple rules? *arXiv preprint arXiv:2311.04235*.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi A Omiye, Mehr Kashyap, et al. 2024. Assessing the limitations of large language models in clinical fact decomposition. *arXiv preprint arXiv:2412.12422*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A large-scale Chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8428–8438, Miami, Florida, USA. Association for Computational Linguistics.

John Øvretveit. 2001. Quality evaluation and indicator comparison in health care. *The International journal of health planning and management*, 16(3):229–241.

Gottumukkala S Raju, Phillip J Lum, Rebecca S Slack, Selvi Thirumurthi, Patrick M Lynch, Ethan Miller, Brian R Weston, Marta L Davila, Manoop S Bhutani, Mehnaz A Shafi, et al. 2015. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. *Gastrointestinal endoscopy*, 82(3):512–519.

William A Ross, Selvi Thirumurthi, Patrick M Lynch, Asif Rashid, Mala Pande, Mehnaz A Shafi, Jeffrey H Lee, and Gottumukkala S Raju. 2015. Detection rates of premalignant polyps during screening colonoscopy: time to revise quality standards? *Gastrointestinal endoscopy*, 81(3):567–574.

Sergio Servantez, Joe Barrow, Kristian Hammond, and R. Jain. 2024a. Chain of logic: Rule-based reasoning with large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024b. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand. Association for Computational Linguistics.

Anand Subramanian, Viktor Schlegel, Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Vijay Prakash Dwivedi, and Stefan Winkler. 2024. M-QALM: A benchmark to assess clinical reading comprehension and knowledge recall in large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4002–4042, Bangkok, Thailand. Association for Computational Linguistics.

Wangtao Sun, Xuanqing Yu, Shizhu He, Jun Zhao, and Kang Liu. 2023. ExpNote: Black-box large language models are better task solvers with experience notebook. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15470–15481, Singapore. Association for Computational Linguistics.

Wangtao Sun, Chenxiang Zhang, Xueyou Zhang, Xuanqing Yu, Ziyang Huang, Pei Chen, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Beyond instruction following: Evaluating inferential rule following of large language models.

Suzanne Tamang, Manali I Patel, Douglas W Blayney, Julie Kuznetsov, Samuel G Finlayson, Yohan Vetteth, and Nigam Shah. 2015. Detecting unplanned care from clinician notes in electronic health records. *Journal of Oncology Practice*, 11(3):e313–e319.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024a. DiReCT: Diagnostic reasoning for clinical notes via large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024b. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654*.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024c. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *Preprint*, arXiv:2403.03640.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhuofeng Wu, Richard He Bai, Aonan Zhang, Jiatao Gu, V.G.Vinod Vydiswaran, Navdeep Jaitly, and Yizhe Zhang. 2024. Divide-or-conquer? which part should you distill your LLM? In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E. Gonzalez, and Bin CUI. 2024b. Buffer of thoughts: Thought-augmented reasoning with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023a. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024c. Can LLMs reason in the wild with programs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9806–9829, Miami, Florida, USA. Association for Computational Linguistics.

Zeyuan Yang, Peng Li, and Yang Liu. 2023b. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. *arXiv preprint arXiv:2310.15746*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Qi Ye, Guangya Yu, Jingping Liu, Erzhen Chen, Chenjie Dong, Xiaosheng Lin, Zelei Liu, Han Yu, and Tong Ruan. 2025. Imqc: A large language model platform for medical quality control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28810–28818.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2025. Efficiently democratizing medical LLMs for 50 languages via a mixture of language family experts. In *The Thirteenth International Conference on Learning Representations*.

A Models

- **Qwen2.5-Instruct series.** We choose the {0.5, 1.5, 3, 7, 14, 32, 72} sizes.
- **Internlm2.5-Chat series.** We choose the {1.8, 7, 20} sizes.
- **Llama3.1-Instruct series.** We choose the {8, 70} sizes.
- **MiniCPM3-4B** [8]. It's a lightweight Chinese LLM.
- **Apollo series** We choose the Apollo2 {0.5, 1.5, 7} and Apollo 72. Apollo models trained on Qwen and Qwen2 with a high quality medical dataset.
- **HuatuoGPT2 series** We choose the HuatuoGPT2 {7, 14}. We strictly followed the default load method of HuatuoGPT2-34B, but the inference time was too long, and the final results were not satisfactory. As a result, we did not conduct further experiments on HuatuoGPT2-34B.

B Training Details

As shown in Figure 5 and Table 6, Qwen2.5-7B-Instruct is an excellent foundation model, considering both performance and time cost. However, for practical purposes, we prefer using a more lightweight model. We trained the model with LLaMA-Factory. [9] We use the default ds_z3_config and Lora fine-tuning. Detail parameters: per_device_train_batch_size: 3; gradient_accumulation_steps: 8; learning_rate: 1.0e-5;num_train_epochs:4; lr_scheduler_type: cosine; warmup_ratio: 0.1; fp16: true; ddp_timeout: 180000000.

C Discussion on Test-Time Scaling

While test-time scaling like (CoT-SC, ToT or O1, R1) has attracted significant research attention, practical deployment in clinical settings requires careful consideration of GPU resource constraints, as hospitals typically have limited computational capacity. To address this, we quantitatively compare the inference efficiency of our method against CoT in Table 6, with measurements reported in GPU-hours.

	zero-shot		one-shot	
	CoT	CF-IR	CoT	CF-IR
Qwen2.5-7B	0.59	0.46↓	0.56	0.69↑
Qwen2.5-14B	1.30	1.26↓	1.40	1.70↑
Qwen2.5-32B	3.88	3.60↓	3.32	4.16↑
Qwen2.5-72B	7.60	7.04↓	5.80	6.84↑

Table 6: Total Inference time of Qwen on CMQCIC-Bench across different methods. The unit is an hour·GPU. The arrow indicates the change in inference time cost from CoT to CF-IR.

Models	CoT	CF-IR
DeepSeek-V3	3h40min	3h29min↓
DeepSeek-R1	10h10min	9h30min↓

Table 7: Inference cost between DeepSeek-R1 and DeepSeek-V3. Experiments in zero-shot setting.

Comparing inference times between chat and reasoning models in Table 7, we observe that while the reasoning model demonstrates better scalability, the performance gains remain marginal in Table 8. In the zero-shot setting, R1 did not demonstrate significant improvement. In contrast, our CF-IR framework outperformed the long cot method.

D Prompt

Here are the zero-shot prompt templates for the data construction, rule representation enhancement, clinical fact-based inferential rule reasoning method, and other prompt-based methods we used in this paper.

For each indicator rule question, we provide a list of facts that should be extracted. The prompt for clinical fact extraction is shown in Figure 7. The main fields are in the Table 9 and Table 10.

Prompt for Rule Enhancement. As mentioned

	Standard	CoT	CF-IR
Qwen2.5-7B	66.49	82.80	82.92
Qwen2.5-14B	78.79	86.49	87.21
Qwen2.5-32B	75.54	82.03	86.11
Qwen2.5-72B	87.77	87.51	92.73
llama3.1-70B	82.54	85.85	85.47
GPT-4o	77.45	88.91	91.84
DeepSeek-V3	82.67	86.83	91.84
DeepSeek-R1	81.01	82.99	92.73

Table 8: Performance on test-time scaling.

[8] https://huggingface.co/openbmb/MiniCPM3-4B

[9] https://github.com/hiyouga/LLaMA-Factory

Fileds Name	Explanation
definition	The definition of the indicator
formula	The calculation formula of the indicator
significance	The medical impact of indicator
other	The relative knowledge or supplement
instruction_standard	The standard prompt for MQCIC with the rule
numerator	The numerator of indicator
denominator	The denominator of indicator
rule	The numerator rule of indicator
facts	The templated clinical facts list
logical_rules	The logical rules lists. Containing natural and symbolic languages.

Table 9: Main fields explanation of indicator file.

Fileds Name	Explanation
unique_id	The unique id of the indicator
patient note	The patient note of the instance
explaination	The explanation of the answer
label	The label of the answer
facts	The list that contains all templated clinical facts with the related original text and answer
logic	The list that contains logical rules with the answer

Table 10: Main fields explanation of data file.

Prompt for Clinical Fact Extraction

Instuction: You are a medical expert. Please extract the {fact} based on the provided electronic medical record. You should determine the value of the fact from the electronic medical record without making any assumptions. Please provide specific evidence of the fact from the original text. If there is no evidence, consider it unable to be determined.
Input: Patient Note
Output: Clinical facts and evidence

Figure 7: The clinical fact extraction prompt.

Prompt for Knowledge Enhancement

Instuction: You are a expert with medical knowledge. Please generate the medical knowledge realted to indicator.
Input: {Indicator Definition}
Output: Relative knowledge

Figure 8: The Prompt Template of Knowledge Enhancement.

in Section 4, we leverage LLMs to transform rules through three steps: Knowledge Enhancement, Rule Decomposition, and Clinical Fact Templaza-tion. As shown in Figure 8, 9 and 10, we leverage GPT-4o first to generate the relative knowledge, logical rules, and templated clinical facts separately.

Prompt for Different Methods. The detail prompt of different method as shown in Figure 11, Figure 12 and Figure 13.

Prompt for Rule Decomposition

Instuction: You are a expert with medical knowledge. Please think step by step. Based on the relative knowledge, break down the {Rule} into sub-ogical rules, which include both natural language and symbolic language components.
Input: {Rule ; Relative Knowledge}
Output: Sub-logical rules

Figure 9: The Prompt Template of Rule Decomposition.

Prompt for Templated Clinical Fact

Instuction: You are a expert with medical knowledge. Please extract all clinical facts that need to be verified from the logical rules. Each clinical fact should be independent of the others, with possible values including Yes/No, numerical values, or characters.
Input: {Logical Rules}
Output: Clinical facts with assumed values

Figure 10: The Prompt of Templated Clinical Fact.

E Additional Analysis Details

E.1 Error Analysis Details

As shown in Table 11 and Table 12 we observed: (1) Although the one-shot method reduces errors across both types in both approaches, the CoT method still results in more Type B errors, which may be due to the differing reasoning paths in the examples. (2) The CF-IR method effectively reduces Type B errors, but when it comes to Type A errors, the issue seems to be more related to the

Prompt Template

Zero-Shot Standard:
English Version
###Instruction: This is an indicator calculation task. You need to evaluate the {patient note} based on the given {rules}. You should provide relevant information from the record as an explanation. Finally, ###output: True/False/Not Sure. \n###Input: **{patient note} {rules}**
Chinese Version
###Instrcution: 这是一个指标计算任务，你需要根据给定的{规则}来对{电子病历}进行判断，你需要给出病历当中相关的信息作为解释。最后输出：True/False/Not Sure.\n###输入：{电子病历}{规则}

Zero-Shot CoT：
English Version
###Instruction: This is an indicator calculation task. You need to evaluate the {patient note} based on the given {rules}. You should provide relevant information from the record as an explanation. Let's think setp by step! Finally, ###output: True/False/Not Sure. \n###Input:**{patient note} {rules}**
Chinese Version
##Instrcution: 这是一个指标计算任务，你需要根据给定的{规则}来对{电子病历}进行判断，你需要给出病历当中相关的信息作为解释。请你一步步思考，给出具体的依据和推理过程。最后输出：True/False/Not Sure.\n###输入：{电子病历}{规则}

Figure 11: The prompt template of standard and cot methods in translated English and Chinese version.

model's intrinsic capabilities, which our method has not been able to enhance or activate effectively.

E.2 Clinical Fact Verification and Inferential Rule

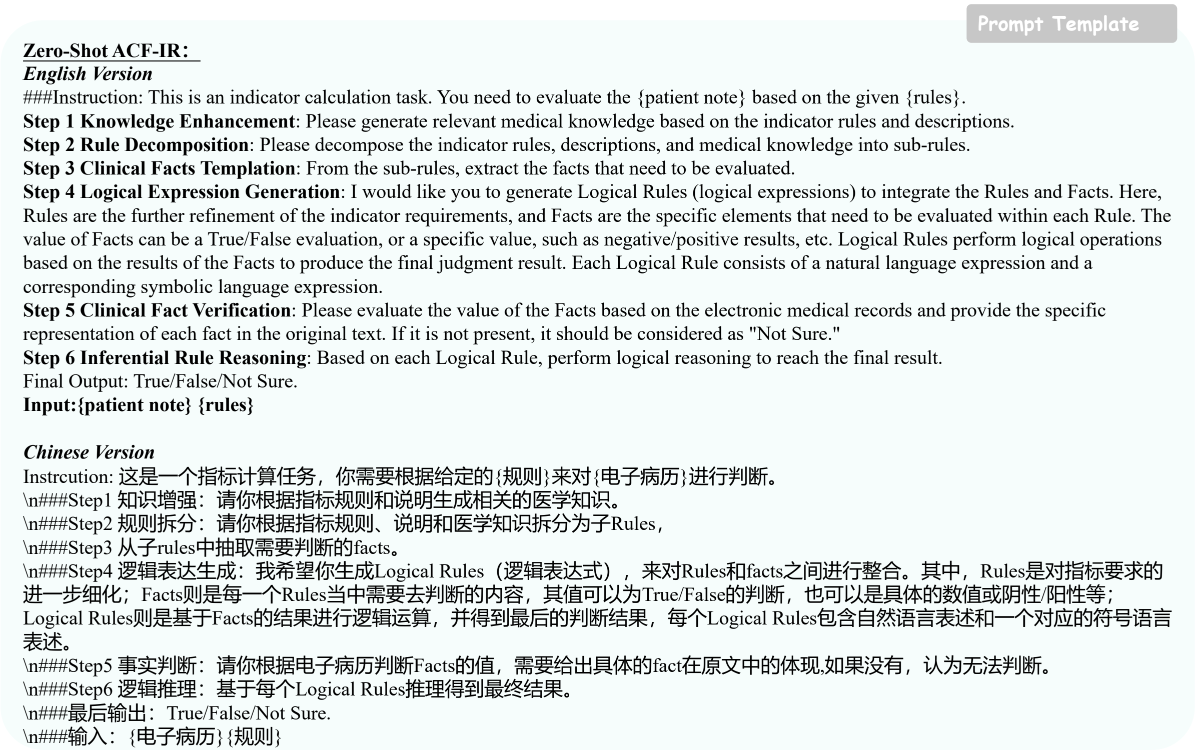
It may be due to the excessively strict scoring criteria that the overall score for clinical fact verification ability is relatively low, but the trend still aligns with expectations. As mentioned in Section 6.3, there is a clear correlation between the model parameters and capabilities in the Apollo and Qwen series. As shown in Table 13, however, models like llama3.1 and HuatuoGPT2, due to differences in the number of parameters, fail to demonstrate this relationship.

		Zero-shot CoT				One-shot CoT			
		Type A	Type B	Type C	Total	Type A	Type B	Type C	Total
General	MiniCPM3-4B	0.21	0.08	0.00	0.28	0.13↓	0.03↓	0.00-	0.16
	Internlm2.5-1.8B	0.35	0.08	0.00	0.44	0.23↓	0.08-	0.00-	0.31
	Internlm2.5-7B	0.15	0.12	0.00	0.27	0.13↓	0.03↓	0.00-	0.16
	Internlm2.5-20B	0.17	0.05	0.00	0.22	0.11↓	0.03↓	0.00-	0.14
	Qwen2.5-0.5B	0.34	0.10	0.00	0.44	0.29↓	0.10-	0.00-	0.39
	Qwen2.5-1.5B	0.21	0.16	0.00	0.37	0.24↑	0.05↓	0.00-	0.29
	Qwen2.5-3B	0.16	0.11	0.00	0.27	0.15↓	0.07↓	0.00-	0.22
	Qwen2.5-7B	0.13	0.04	0.00	0.17	0.12↓	0.03↓	0.00-	0.15
	Qwen2.5-14B	0.10	0.08	0.00	0.18	0.08↓	0.04↓	0.00-	0.12
	Qwen2.5-32B	0.09	0.05	0.00	0.14	0.06↓	0.04↓	0.00-	0.10
	Qwen2.5-72B	0.09	0.03	0.00	0.12	0.07↓	0.02↓	0.00-	0.09
	llama3.1-8B	0.23	0.14	0.00	0.37	0.14↓	0.04↓	0.00-	0.18
	llama3.1-70B	0.10	0.04	0.00	0.14	0.08↓	0.03↓	0.00-	0.11
	GPT-4o	0.08	0.03	0.00	0.11	0.06↓	0.02↓	0.00-	0.08
Medical	HuatuoGPT2-7B	0.17	0.28	0.00	0.46	0.23↑	0.23↓	0.02↑	0.47
	HuatuoGPT2-14B	0.32	0.13	0.01	0.45	0.28↓	0.18↑	0.01-	0.48
	Apollo2-0.5B	0.34	0.19	0.04	0.58	0.30↓	0.15↓	0.02↓	0.46
	Apollo2-1.5B	0.23	0.25	0.00	0.48	0.27↑	0.07↓	0.00-	0.34
	Apollo2-7B	0.17	0.23	0.00	0.40	0.24↑	0.04↓	0.00-	0.28
	Apollo-72B	0.14	0.10	0.00	0.24	0.11↓	0.03↓	0.00-	0.14

Table 11: Error type distribution of LLMs on CMQCIC-Bench dataset. Arrows represent the changes from zero-shot to one-shot.

		Zero-shot CF-IR				One-shot CF-IR			
		Type A	Type B	Type C	Total	Type A	Type B	Type C	Total
General	MiniCPM3-4B	0.27	0.04	0.00	0.31	0.11↓	0.06↑	0.00-	0.17
	Internlm2.5-1.8B	0.39	0.07	0.00	0.46	0.27↓	0.10↑	0.00-	0.36
	Internlm2.5-7B	0.18	0.03	0.00	0.21	0.13↓	0.02↓	0.00-	0.16
	Internlm2.5-20B	0.15	0.04	0.00	0.19	0.10↓	0.01↓	0.00-	0.11
	Qwen2.5-0.5B	0.28	0.18	0.00	0.46	0.28-	0.18-	0.01↑	0.46
	Qwen2.5-1.5B	0.36	0.02	0.00	0.38	0.25↓	0.02-	0.00-	0.27
	Qwen2.5-3B	0.31	0.01	0.00	0.32	0.16↓	0.02↑	0.00-	0.18
	Qwen2.5-7B	0.14	0.03	0.00	0.17	0.08↓	0.01↓	0.00-	0.10
	Qwen2.5-14B	0.12	0.01	0.00	0.13	0.07↓	0.01-	0.00-	0.08
	Qwen2.5-32B	0.12	0.02	0.00	0.14	0.05↓	0.00↓	0.00-	0.05
	Qwen2.5-72B	0.06	0.01	0.00	0.07	0.03↓	0.01-	0.00-	0.04
	llama3.1-8B	0.20	0.02	0.00	0.22	0.12↓	0.01↓	0.00-	0.14
	llama3.1-70B	0.12	0.03	0.00	0.15	0.07↓	0.00↓	0.00-	0.08
	GPT-4o	0.07	0.01	0.01	0.08	0.05↓	0.01-	0.00↑	0.06
Medical	HuatuoGPT2-7B	0.33	0.16	0.01	0.50	0.31↓	0.11↓	0.02↑	0.43
	HuatuoGPT2-14B	0.43	0.09	0.02	0.54	0.46↑	0.08↓	0.04↑	0.57
	Apollo2-0.5B	0.32	0.15	0.12	0.59	0.27↓	0.07↓	0.01↓	0.35
	Apollo2-1.5B	0.30	0.17	0.01	0.49	0.26↓	0.08↓	0.01-	0.35
	Apollo2-7B	0.29	0.10	0.00	0.38	0.14↓	0.21↑	0.01-	0.35
	Apollo-72B	0.26	0.01	0.00	0.27	0.11↓	0.03↑	0.00-	0.14

Table 12: Error type distribution of LLMs on CMQCIC-Bench dataset. Arrows represent the changes from zero-shot to one-shot.



Prompt Template

Zero-Shot ACF-IR：

English Version

###Instruction: This is an indicator calculation task. You need to evaluate the {patient note} based on the given {rules}.

Step 1 Knowledge Enhancement: Please generate relevant medical knowledge based on the indicator rules and descriptions.

Step 2 Rule Decomposition: Please decompose the indicator rules, descriptions, and medical knowledge into sub-rules.

Step 3 Clinical Facts Templation: From the sub-rules, extract the facts that need to be evaluated.

Step 4 Logical Expression Generation: I would like you to generate Logical Rules (logical expressions) to integrate the Rules and Facts. Here, Rules are the further refinement of the indicator requirements, and Facts are the specific elements that need to be evaluated within each Rule. The value of Facts can be a True/False evaluation, or a specific value, such as negative/positive results, etc. Logical Rules perform logical operations based on the results of the Facts to produce the final judgment result. Each Logical Rule consists of a natural language expression and a corresponding symbolic language expression.

Step 5 Clinical Fact Verification: Please evaluate the value of the Facts based on the electronic medical records and provide the specific representation of each fact in the original text. If it is not present, it should be considered as "Not Sure."

Step 6 Inferential Rule Reasoning: Based on each Logical Rule, perform logical reasoning to reach the final result.

Final Output: True/False/Not Sure.

Input:{patient note} {rules}

Chinese Version

Instrcution: 这是一个指标计算任务，你需要根据给定的{规则}来对{电子病历}进行判断。

\n###Step1 知识增强：请你根据指标规则和说明生成相关的医学知识。

\n###Step2 规则拆分：请你根据指标规则、说明和医学知识拆分为子Rules，

\n###Step3 从子rules中抽取需要判断的facts。

\n###Step4 逻辑表达生成：我希望你生成Logical Rules（逻辑表达式），来对Rules和facts之间进行整合。其中，Rules是对指标要求的进一步细化；Facts则是每一个Rules当中需要去判断的内容，其值可以为True/False的判断，也可以是具体的数值或阴性/阳性等；Logical Rules则是基于Facts的结果进行逻辑运算，并得到最后的判断结果，每个Logical Rules包含自然语言表述和一个对应的符号语言表述。

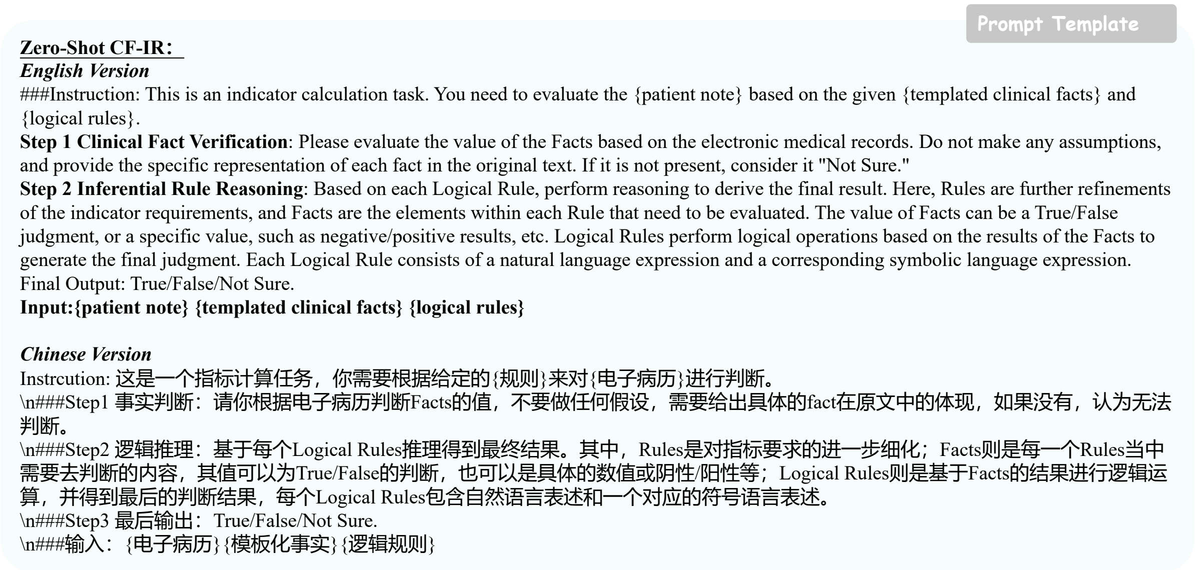
\n###Step5 事实判断：请你根据电子病历判断Facts的值，需要给出具体的fact在原文中的体现,如果没有，认为无法判断。

\n###Step6 逻辑推理：基于每个Logical Rules推理得到最终结果。

\n###最后输出：True/False/Not Sure.

\n###输入：{电子病历}{规则}

Figure 12: The prompt template of ACF-IR method in translated English version and Chinese version.



Prompt Template

Zero-Shot CF-IR：

English Version

###Instruction: This is an indicator calculation task. You need to evaluate the {patient note} based on the given {templated clinical facts} and {logical rules}.

Step 1 Clinical Fact Verification: Please evaluate the value of the Facts based on the electronic medical records. Do not make any assumptions, and provide the specific representation of each fact in the original text. If it is not present, consider it "Not Sure."

Step 2 Inferential Rule Reasoning: Based on each Logical Rule, perform reasoning to derive the final result. Here, Rules are further refinements of the indicator requirements, and Facts are the elements within each Rule that need to be evaluated. The value of Facts can be a True/False judgment, or a specific value, such as negative/positive results, etc. Logical Rules perform logical operations based on the results of the Facts to generate the final judgment. Each Logical Rule consists of a natural language expression and a corresponding symbolic language expression.

Final Output: True/False/Not Sure.

Input:{patient note} {templated clinical facts} {logical rules}

Chinese Version

Instrcution: 这是一个指标计算任务，你需要根据给定的{规则}来对{电子病历}进行判断。

\n###Step1 事实判断：请你根据电子病历判断Facts的值，不要做任何假设，需要给出具体的fact在原文中的体现，如果没有，认为无法判断。

\n###Step2 逻辑推理：基于每个Logical Rules推理得到最终结果。其中，Rules是对指标要求的进一步细化；Facts则是每一个Rules当中需要去判断的内容，其值可以为True/False的判断，也可以是具体的数值或阴性/阳性等；Logical Rules则是基于Facts的结果进行逻辑运算，并得到最后的判断结果，每个Logical Rules包含自然语言表述和一个对应的符号语言表述。

\n###Step3 最后输出：True/False/Not Sure.

\n###输入：{电子病历}{模板化事实}{逻辑规则}

Figure 13: The prompt template of CF-IR method in translated English and Chinese version.

		Clinical Fact Verification		Inferential Rule Reasoning		One-Stage
		Faithfulness	Correctness	NL(ACC)	SY(ACC)	Best(ACC)
General	MiniCPM3-4B	45.14	36.05	67.38	69.68	72.10*
	Qwen2.5-0.5B	45.37	40.1	74.90	75.92	56.05*
	Qwen2.5-1.5B	46.45	43.67	75.66	74.64	66.11*
	Qwen2.5-3B	52.08	50.02	83.69	81.18	76.05*
	Qwen2.5-7B	67.63	53.13	90.44	88.91	82.92*
	Qwen2.5-14B	67.94	62.35	92.35	94.52	87.21*
	Qwen2.5-32B	75.41	69.47	89.80	94.14	86.11*
	Qwen2.5-72B	74.07	77.13	93.63	93.88	92.73*
	llama3.1-8B	47.89	37.88	82.67	80.89	78.34*
	llama3.1-70B	48.59	40.48	85.60	85.98	85.47*
Medical	HuatuoGPT2-7B	13.27	22.93	47.89	44.20	54.26*
	HuatuoGPT2-14B	32.55	35.47	62.42	62.03	55.28*
	Apollo2-0.5B	5.19	10.73	9.80	18.47	49.29*
	Apollo2-1.5B	25.31	29.21	48.53	52.86	55.03*
	Apollo2-7B	33.78	33.93	56.30	42.29	61.91*
	Apollo-72B	42.31	32.07	76.81	71.59	76.24*
Average		45.18	42.16	**71.12**	70.69	70.94*

Table 13: Performance of Clinical Fact Verification and Inferential Rule Reasoning on CMQCIC-Bench. For Clinical Fact Verification, we utilize DeepSeek to assess both faithfulness and correctness. NL represents Natural Language, SY denotes Symbolic Language, and ACC stands for Accuracy. To enable a clearer comparison, we present the best results of the "standard," "CoT," and "CF-IR" approaches in Table 2. All experiments were conducted in the zero-shot setting.