CAMEC: Complexity-Aware Multi-Expert Collaboration for Reliable Chinese Medical Question Answering

Yukang Wu[1] **Xiyuan Jia**[1*] **Jiayi Wu**[2] **Hongchen Yu**[1] **Yuhan Qiu**[1] **Guohua Wu**[1,3,4*]
[1]Hangzhou Dianzi University
[2]Leiden Institute of Advanced Computer Science, Leiden University
[3]Hangzhou Institute for Advanced Study, UCAS
[4]Zhejiang Provincial Key Laboratory for Sensitive Data Security Protection and Confidentiality Management
{wuyukang, jiaxiyuan, wugh}@hdu.edu.cn

Abstract

Large language models (LLMs) are promising for medical question answering (QA) but remain unreliable in Chinese clinical settings due to hallucinations, weak factual grounding, and difficulty handling clinically complex cases. We propose CAMEC (Complexity-Aware Multi-Expert Collaboration), a framework that combines hierarchical medical adaptation with complexity-aware expert routing for reliable Chinese medical QA. We adopt a three-stage LoRA-based supervised fine-tuning pipeline for domain adaptation, instruction following, and clinical reasoning. At inference, CAMEC routes each query by predicted complexity and selectively recruits three experts: an internal chain-of-thought (CoT) expert, a retrieval-augmented expert over a dense medical vector database, and a knowledge graph (KG) expert over a structured medical knowledge base. An LLM-as-a-Judge module evaluates and critiques expert reports, iteratively refining them into a consensus answer. Experiments on four Chinese medical benchmarks show that CAMEC consistently outperforms strong general and medical LLM baselines, achieving 78.86% (CMExam), 84.15% (MedQA-CN), 78.51% (CMMLU-Med), and 74.40% (CMB-exam), with consistent absolute improvements over the previous state-of-the-art HuatuoGPT-o1-7B across all benchmarks. The complexity-aware router reduces expert invocations and inference cost, making CAMEC both highly effective and computationally efficient.

1 Introduction

Large language models (LLMs) have shown remarkable progress across a wide range of natural language processing tasks (Brown et al., 2020; Workshop et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023; Achiam et al., 2023), motivating their deployment in high-stakes domains such as medicine (Thirunavukarasu et al., 2023; Clusmann et al., 2023). Despite Chinese being one of the most widely spoken languages globally, Chinese medical LLMs remain less mature than their English counterparts (Cai et al., 2024; Wang et al., 2025). While recent efforts have improved domain adaptation through specialized medical training (Chen et al., 2023; Yang et al., 2024), medical LLMs still face fundamental challenges in accuracy, interpretability, and evaluation for clinical applications (Yu et al., 2025). In particular, most existing approaches rely on single-model reasoning paradigms and lack mechanisms for multi-perspective validation, structured knowledge grounding, and adaptive quality control, which are critical for reliable clinical decision support.

To address these challenges, recent Chinese medical LLMs have explored three complementary technical directions to improve domain understanding and reliability: single-model domain adaptation, retrieval augmentation, and multi-agent collaboration. Single-model approaches such as HuatuoGPT (Zhang et al., 2023; Chen et al., 2023, 2024) and Zhongjing (Yang et al., 2024) enhance Chinese medical reasoning through large-scale domain-specific training and preference optimization. As illustrated in Figure 1, such single-model paradigms are vulnerable to hallucinations and diagnostic ambiguity in complex or overlapping symptom scenarios.

To improve factual grounding, retrieval-augmented methods (Lewis et al., 2020; Zhao et al., 2025) have been introduced into Chinese medical QA to incorporate external medical knowledge. While effective in reducing factual errors, methods typically operate as isolated pipelines and lack systematic validation or coordination with complementary reasoning processes.

Beyond single-model and retrieval-based approaches, multi-agent frameworks (Tang et al., 2024; Zhou et al., 2025) explore expert collabo-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38427–38457
July 2-7, 2026

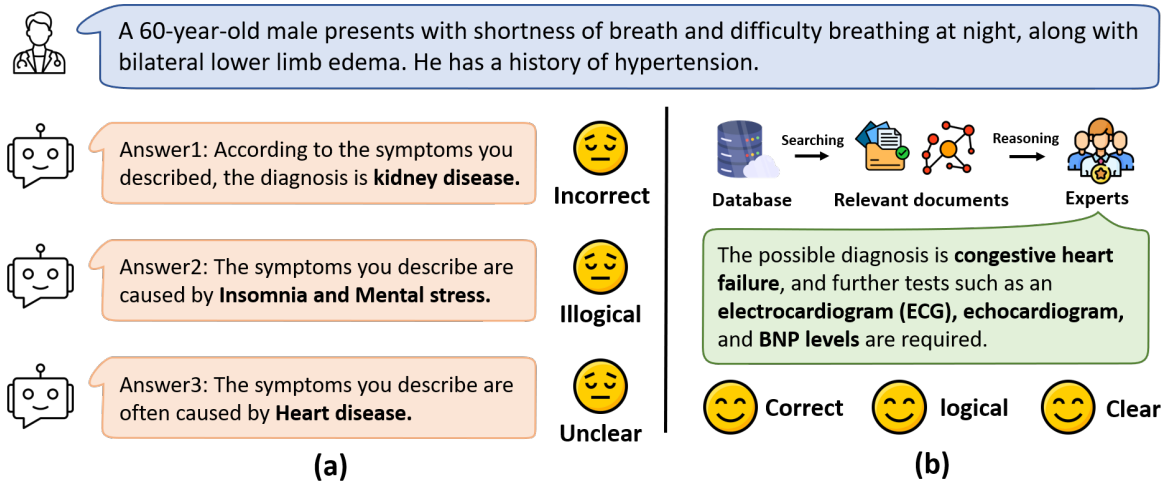

Figure 1: (a) Existing single-model LLMs or naive RAG pipelines may produce incorrect, illogical, or clinically vague diagnoses when symptoms overlap, due to the lack of multi-perspective validation. (b) CAMEC illustrates a multi-expert diagnostic pipeline that integrates external knowledge retrieval and complementary expert reasoning, producing diagnoses that are medically correct, logically consistent, and clinically interpretable.

ration for medical reasoning. Nevertheless, these systems generally adopt static collaboration strategies and do not account for query complexity or provide explicit quality control in high-stakes Chinese medical QA settings.

To address these gaps, we introduce CAMEC (Complexity-Aware Multi-Expert Collaboration), a framework that adaptively coordinates specialized experts based on query complexity for Chinese medical QA. We first establish a strong foundation through hierarchical medical adaptation of Qwen3-8B via three-stage LoRA-based SFT (Domain Adaptation $\rightarrow$ Instruction Following $\rightarrow$ Clinical Reasoning). During inference, a lightweight router predicts query complexity and selectively recruits three specialized experts: E_{CoT} for internal reasoning, E_{RAG} for evidence retrieval, and E_{KG} for knowledge graph grounding. An LLM-as-a-Judge module evaluates expert outputs, triggers refinement when quality thresholds are unmet, and synthesizes high-confidence responses into final decisions.

Our contributions are summarized as follows:

(1) **Complexity-aware adaptive collaboration.** We propose CAMEC, an adaptive multi-expert framework that routes medical queries by predicted complexity to selectively recruit specialists and allocate refinement rounds, improving cost efficiency while maintaining accuracy.

(2) **Multi-expert inference with judge-guided consensus.** We design a multi-expert inference architecture combining internal reasoning, semantic retrieval, and knowledge graph grounding, with a unified LLM-as-a-Judge mechanism for iterative evaluation and synthesis.

(3) **Empirical effectiveness.** Experiments on CMExam, MedQA-CN, CMMLU-Med, and CMB-exam show consistent performance gains over competitive Chinese medical LLM baselines, including HuatuoGPT-o1-7B.

2 Related Work

2.1 LLMs in Medicine

Large medical LLMs have achieved rapid progress in clinical reasoning and medical exam tasks. Med-PaLM and Med-PaLM 2 (Singhal et al., 2023, 2025) improve USMLE performance through medical instruction tuning and alignment with clinician feedback. Building on this, AMIE (Tu et al., 2025) focuses on conversational diagnostic capabilities, while MedFound (Liu et al., 2025) further develops a generalist medical model with a standardized evaluation protocol.

In Chinese medical settings, the HuatuoGPT series (Zhang et al., 2023; Chen et al., 2023, 2024) improves reasoning through large-scale medical corpora and multistage supervised fine-tuning (SFT). DISC-MedLLM (Bao et al., 2023) integrates knowledge graph guidance with human preference alignment for medical dialogue. More recently, ChiMed-GPT (Tian et al., 2024) adopts a full-process training paradigm with pre-training, supervised fine-tuning, and preference optimization to enhance Chinese medical reasoning, while

BenCao (Xie et al., 2025) extends instruction-tuned LLMs to Traditional Chinese Medicine using structured knowledge and expert feedback. Our work follows this line while introducing progressive SFT and complexity-aware multi-expert collaboration.

2.2 RAG in Medical LLMs

Retrieval-augmented generation (RAG) improves factual consistency in medical reasoning by incorporating external biomedical knowledge (Lewis et al., 2020; Asai et al., 2024; Ke et al., 2025). Medical RAG systems address hallucinations in rare-disease and multi-symptom scenarios (Zhao et al., 2025). Hybrid approaches such as KARE (Jiang et al., 2024) and DualRAG (Cheng et al., 2025) integrate KGs to enhance evidence utilization and multi-hop reasoning.

In CAMEC, RAG operates as one of three experts and collaborates with CoT and KG experts under judge supervision. Unlike standard RAG pipelines, our RAG expert is iteratively refined by feedback from a unified judge, improving factual grounding within a collaborative multi-expert setting.

2.3 LLM-based Multi-Agent Collaboration

Multi-agent collaboration has emerged as a way to improve medical reasoning and safety. MedAgents (Tang et al., 2024) uses role-based coordination between diagnostic and reviewer agents. AI Hospital (Fan et al., 2025) simulates clinical interactions using multi-agent architectures, and Tiered Agentic Oversight (Kim et al., 2025) leverages hierarchical oversight for safer decision-making. LLM-as-a-Judge methods (Gu et al., 2025; Li et al., 2025) evaluate and refine model outputs for improved consistency.

CAMEC differs from prior work in two key aspects: it performs parallel expert reasoning rather than sequential agent interactions, and uses a unified judge-driven refinement mechanism to ensure stable and controllable collaboration.

3 Method

We propose CAMEC, a Chinese medical QA framework combining hierarchical supervised fine-tuning, complexity-aware expert routing, multi-expert parallel inference, and judge-guided iterative evaluation. Figure 2 illustrates the complete framework. Given a query q, CAMEC proceeds as:

1. **Hierarchical SFT (training):** starting from Qwen3-8B, we conduct a three-stage LoRA-based SFT pipeline (Domain Adaptation → Instruction Following → Clinical Reasoning).
2. **Routing & parallel generation (inference):** a lightweight router predicts query complexity and activates a subset of experts (CoT/RAG/KG), which generate structured reports in parallel.
3. **Judge-guided evaluation:** A judge scores each report on correctness, completeness, and safety; if no report reaches the acceptance threshold, experts revise with judge feedback for up to three rounds; the judge then synthesizes a final weighted consensus.

3.1 Hierarchical Supervised Fine-Tuning

Training data. We fine-tune Qwen3-8B using several high-quality Chinese medical datasets, including Huatuo26M-Lite (Wang et al., 2025), HuatuoGPT2-SFT-GPT4-140K (Chen et al., 2023), and medical-o1-reasoning-SFT (Chen et al., 2024). Although these datasets are primarily constructed for medical tasks, they may still contain a small amount of noisy or non-medical content. We therefore apply a lightweight filtering step using an LLM to remove samples that are not medically relevant.

Three-stage SFT. We adopt a three-stage hierarchical SFT schedule: (1) Domain Adaptation uses real-world doctor–patient dialogues from Huatuo26M-Lite to align the model with medical terminology and clinical communication patterns. (2) Instruction Following leverages diverse multi-turn consultations from HuatuoGPT2-SFT-GPT4-140K to improve instruction adherence and conversational coherence. (3) Clinical Reasoning employs complex case analyses with chain-of-thought supervision from medical-o1-reasoning-SFT to strengthen diagnostic reasoning.

At each stage, a subset of high-quality data from previous stages is mixed with the current training data to mitigate catastrophic forgetting. We adopt a simple replay strategy, where prior-stage samples are interleaved with current-stage data during training without introducing additional weighting or scheduling. We employ LoRA (Hu et al., 2022) while keeping the base model frozen, which reduces trainable parameters and accelerates training.

3.2 Complexity-Aware Expert Routing

Medical queries exhibit varying diagnostic difficulty. Invoking all experts with multi-round judge refinement for every query is computationally in-

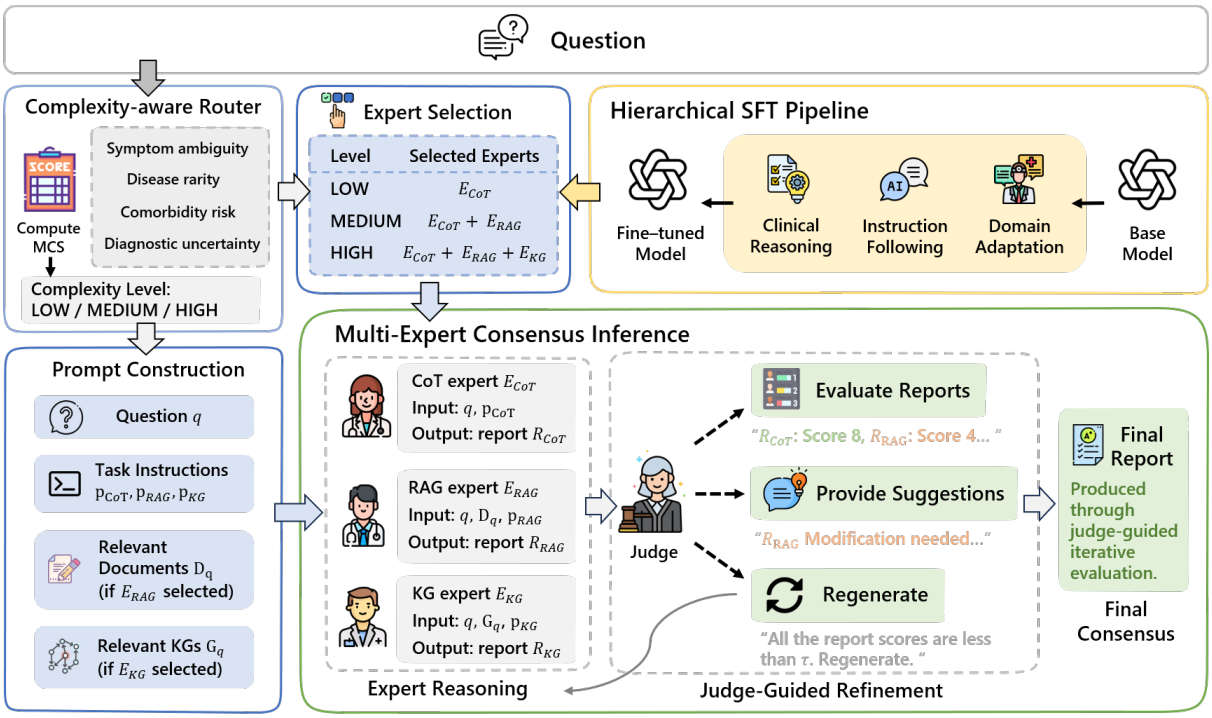

Figure 2: Overview of the CAMEC framework. Given a query, a router predicts its complexity to selectively recruit experts (CoT/RAG/KG); the system retrieves relevant documents or KG facts when needed, constructs expert-specific prompts, generates expert reports in parallel, and uses an LLM-as-a-Judge to iteratively refine them into a final consensus answer.

efficient, especially for straightforward cases. We introduce a lightweight **complexity-aware router** to predict query difficulty and adaptively activate experts.

Medical Complexity Score (MCS). We define a *Medical Complexity Score* (MCS) for each query q as a weighted combination of four clinically motivated dimensions:

$$\mathrm{MCS}(q) = \sum_{i=1}^{4} w_i \cdot f_i(q), \qquad (1)$$

where f_1 = *symptom ambiguity*, f_2 = *disease rarity*, f_3 = *comorbidity risk*, and f_4 = *diagnostic uncertainty*. Each $f_i(q) \in [0, 3]$, and we set $w_i = \{0.3, 0.25, 0.25, 0.2\}$. We partition MCS into **LOW** (<1), **MEDIUM** ($[1, 2)$), and **HIGH** (≥ 2). The prompting template and rubric are provided in Appendix B.1.

Router Training. We sample 8,000 questions from the training split of CMExam (Liu et al., 2023) and MedQA-CN (Jin et al., 2021) (4,000 each) to construct the training set for the complexity router. For each question, we use DeepSeek-R1 (Guo et al., 2025) to assign the four dimension scores and derive the final complexity label. Using these annotated samples (Appendix C.3), we fine-tune a lightweight Qwen3-0.6B (Yang et al., 2025) model as a 3-way classifier (LOW / MEDIUM / HIGH). This trained router enables adaptive expert selection at inference time, reducing computational cost while maintaining diagnostic accuracy.

Adaptive Expert Selection. At inference, the router predicts the complexity level of the question q and applies the following policy:

- **LOW**: E_{CoT} only.
- **MEDIUM**: E_{CoT} + E_{RAG}.
- **HIGH**: E_{CoT} + E_{RAG} + E_{KG}.

This adaptive routing strategy reduces average expert invocations while maintaining high QA accuracy, as we demonstrate empirically in §5.4.

3.3 Multi-Expert Parallel Inference

To enhance the reliability and professionalism of medical reasoning, we design a multi-expert parallel framework comprising three modules: an internal reasoning expert E_{CoT}, a retrieval-augmented expert E_{RAG}, and a knowledge graph expert E_{KG}. The architecture combines intrinsic reasoning, external evidence retrieval, and structured knowledge constraints to generate clinically sound and interpretable answers.

3.3.1 Chain-of-Thought Expert E_{CoT}

Clinical diagnosis relies not only on retrieving facts but also on reasoning about symptoms, signs, disease progression, and risk factors. Given a query q, the internal reasoning expert performs chain-of-thought reasoning purely based on the model's internal knowledge:

$$R_{\text{CoT}} = E_{\text{CoT}}(q, p_{\text{CoT}}) \tag{2}$$

where p_{CoT} is a structured prompt guiding the model to produce differential diagnosis, key evidence, and management suggestions. This expert enhances robustness and interpretability when external evidence is missing.

3.3.2 RAG Expert E_{RAG}

To improve factual grounding, we build a dense retrieval module over a vector store M constructed from Huatuo_encyclopedia_qa (Wang et al., 2025) question–answer pairs. Each concatenated QA pair forms a document entry, and we use Milvus (Wang et al., 2021) as the vector database for efficient similarity search.

For a query q, we first encode it using Qwen3-Embedding-0.6B (Zhang et al., 2025):

$$h_q = f_{\text{emb}}(q) \in \mathbb{R}^d \tag{3}$$

Then we perform IVF-based ANN search in M to obtain the top-k documents:

$$D_q = \text{TopK}\big(\text{Search}(h_q; M, \text{IVF})\big). \tag{4}$$

Given D_q, the RAG expert is prompted with p_{RAG} to read, integrate, and reason over retrieved evidence, explicitly citing slices in its output:

$$R_{\text{RAG}} = E_{\text{RAG}}(q, D_q, p_{\text{RAG}}) \tag{5}$$

This module emphasizes verifiability, factual consistency, and traceable evidence.

3.3.3 Knowledge Graph Expert E_{KG}

Clinical knowledge exhibits strong structure (e.g., *disease→symptom*, *examination→treatment*, *contraindication* links). To avoid inconsistencies from purely unstructured reasoning, we introduce a knowledge graph expert operating over a medical KG $G = (V, R)$. For a query q, entity linking is performed:

$$f_{\text{ent}}(q) \to A \subseteq V \tag{6}$$

Then a local subgraph is expanded around anchored nodes with radius r:

$$G_q = \text{Expand}(G, A, r) \tag{7}$$

The expert performs structured path reasoning and generates a report containing causal paths, compliance checks, conflict warnings, and explicit evidence chains:

$$R_{\text{KG}} = E_{\text{KG}}(q, G_q, p_{\text{KG}}) \tag{8}$$

This module improves explainability, auditability, and safety.

Robustness to Incomplete or Outdated Knowledge. While the above experts rely on both internal and external knowledge, In real-world clinical settings, external knowledge sources such as retrieval databases and knowledge graphs may be incomplete or outdated. To enhance robustness, CAMEC incorporates complementary mechanisms across experts.

First, the internal reasoning expert E_{CoT} provides a fallback when external evidence is missing or unreliable, enabling the system to generate clinically plausible hypotheses based on learned medical knowledge.

Second, the judge module performs cross-expert validation by comparing outputs from E_{CoT}, E_{RAG}, and E_{KG}. When insufficient or inconsistent external knowledge leads to erroneous or conflicting expert outputs, the judge assigns lower scores and triggers iterative refinement.

Finally, the consensus synthesis mechanism prioritizes high-confidence claims supported by multiple experts, reducing the impact of noisy or outdated external knowledge.

3.4 Judge-Guided Iterative Evaluation

Reports from E_{CoT}, E_{RAG}, and E_{KG} may differ or conflict. To obtain a consistent and reliable answer, we employ a unified judge module that evaluates all candidate reports and guides iterative refinement and final synthesis.

Scoring scheme. Inspired by LLM-as-a-Judge (Gu et al., 2025), the judge is instantiated using the same fine-tuned 8B model as the experts, with an evaluation-oriented prompt that scores reports along three dimensions: medical correctness, completeness, and safety (Appendix B.3). For each expert report $R_e^{(t)}$ at iteration t, the judge outputs

dimension-wise scores $(s^{(t)}_{e,\text{corr}}, s^{(t)}_{e,\text{comp}}, s^{(t)}_{e,\text{safe}})$ together with targeted feedback $\delta^{(t)}_e$. These subscores are aggregated into a scalar score:

$$S^{(t)}_e = \alpha s^{(t)}_{e,\text{corr}} + \beta s^{(t)}_{e,\text{comp}} + \gamma s^{(t)}_{e,\text{safe}}, \quad (9)$$

where $\alpha = 0.6$, $\beta = 0.2$, $\gamma = 0.2$ to prioritize medical correctness, and each dimension is scored on a 0–10 scale. This aggregated score is used for threshold-based stopping and expert ranking.

Iterative arbitration. Let $\mathcal{E}_q \subseteq \{E_{\text{CoT}}, E_{\text{RAG}}, E_{\text{KG}}\}$ denote the experts activated for query q. At iteration t, experts generate reports:

$$C_t = \{R^{(t)}_e : e \in \mathcal{E}_q\}. \quad (10)$$

The judge evaluates all candidates and returns:

$$\left(S^{(t)}_e, \delta^{(t)}_e\right) = J(C_t, p_{\text{score}}). \quad (11)$$

We set an acceptance threshold $\tau = 8$ and a maximum of $T_{\max} = 3$ iterations. If $\max S^{(t)}_e < \tau$, experts revise their reports according to judge feedback:

$$R^{(t+1)}_e = \text{Revise}(R^{(t)}_e, \delta^{(t)}_e), \quad (12)$$

where feedback is dimension-aware (e.g., correcting inconsistencies, improving coverage, or mitigating unsafe assertions). Once any report exceeds τ, the iteration stops.

Consensus synthesis. Given the final candidate set $C_T = \{R^{(T)}_e : e \in \mathcal{E}_q\}$ and their aggregated scores $S^{(T)}_e$, the judge synthesizes a consensus report via a weighted, conflict-aware fusion strategy. We define the fusion weights as:

$$w_e = \frac{\exp(S^{(T)}_e)}{\sum_{e' \in \mathcal{E}_q} \exp(S^{(T)}_{e'})}. \quad (13)$$

Each expert report can be viewed as a set of atomic medical claims (e.g., diagnoses, evidence, and recommendations). For any claim a, we compute its aggregated support as:

$$\text{score}(a) = \sum_{e \in \mathcal{E}_q : a \in \mathcal{A}_e} w_e. \quad (14)$$

The fusion procedure retains highly supported claims and resolves conflicts by prioritizing assertions from experts with higher scores. In practice, the judge implements this via structured prompting: it compares expert reports with their associated scores, prioritizes high-confidence content, and resolves contradictions. The final report R_f remains faithful to the expert outputs and does not introduce unsupported medical facts.

4 Experiments

4.1 Datasets and Benchmarks

Training Datasets. We fine-tune Qwen3-8B on three Chinese medical datasets following the three-stage SFT pipeline in §3.1. Dataset statistics and preprocessing are reported in Appendix C.1.

Evaluation Benchmarks. We evaluate on four authoritative Chinese medical benchmarks: CMExam (Liu et al., 2023), CMB-exam (Wang et al., 2024), MedQA-CN (Jin et al., 2021), and CMMLU-Med (Li et al., 2024), which cover factual recall, diagnostic reasoning, and multi-domain medical knowledge. Following prior work, we report accuracy (%) on their official test splits.

Baselines. For CMExam, MedQA-CN, and CMMLU-Med, we compare against general-purpose LLMs (Yi-1.5-9B, GLM-4-9B, Qwen2.5-7B) (Young et al., 2024; GLM et al., 2024; Qwen Team, 2025) and medical-domain models (HuatuoGPT-II-7B, HuatuoGPT-o1-7B) (Chen et al., 2023, 2024), with baseline results from Chen et al. (2024). For CMB-exam, baseline results are taken from Wang et al. (2024), except for Qwen3-8B, which we evaluate under the same benchmark settings. All models are evaluated under identical benchmark settings to ensure fair comparison.

4.2 Implementation Details

Model Fine-tuning. We apply the three-stage LoRA-based SFT schedule described in §3.1, using LoRA rank 16 and maximum sequence length 2048. Full hyperparameters are in Appendix C.2.

Complexity Router. We fine-tune Qwen3-0.6B as a 3-way classifier using 8,000 automatically annotated queries (Appendix C.3). The router activates experts according to predicted complexity: LOW $\rightarrow E_{\text{CoT}}$; MEDIUM $\rightarrow E_{\text{CoT}} + E_{\text{RAG}}$; HIGH $\rightarrow E_{\text{CoT}} + E_{\text{RAG}} + E_{\text{KG}}$.

External Knowledge Sources. For RAG, we build a dense vector index over huatuo_encyclopedia_qa using Qwen3-Embedding-0.6B and Milvus, retrieving k=3 documents per query. For KG, we construct a Neo4j (Robinson et al., 2015) graph with ∼40k entities and ∼290k relations covering diseases, symptoms, drugs, tests, and treatments. Construction and retrieval details are in Appendix D.

Judge Configuration. We set acceptance threshold τ=8, maximum iterations $T_{\max}$=3, decoding temperature 0.3, and context length 4096 tokens. The judge iteratively scores expert reports on [0,10]

Models	CMExam	MedQA-CN	CMMLU-Med
Yi-1.5-9B	68.1	75.8	64.2
GLM-4-9B	70.5	75.2	67.6
Qwen2.5-7B	70.4	71.4	70.5
HuatuoGPT-II-7B	67.4	73.7	58.4
HuatuoGPT-o1-7B	74.1	79.8	74.5
Ours	**78.86**	**84.15**	**78.51**

Table 1: Results on Chinese medical benchmarks. CMMLU-Med indicates that only the medical portion is evaluated. MedQA-CN refers to the Chinese test set of MedQA (MedQA-MCMLE)

Model	Physician	Nurse	Pharmacist	Technician	Disciplines	Graduate Exam	Average
General Models							
Baichuan2-7B-Chat	42.55	51.75	44.59	45.50	43.00	32.56	43.33
Qwen-7B-Chat	48.00	54.25	48.34	48.08	44.87	35.94	46.41
Deepseek-llm-67B-Chat	52.90	61.50	54.28	51.42	51.19	40.63	51.99
Qwen3-8B	61.70	68.88	63.34	59.92	55.69	39.62	58.99
GPT-4	59.90	69.31	52.19	61.50	59.69	54.19	59.46
Yi-34B-Chat	<u>71.10</u>	<u>77.56</u>	<u>73.16</u>	**73.67**	<u>66.56</u>	52.94	<u>69.17</u>
Medical Models							
BianQue-2	4.90	4.19	4.28	3.58	3.31	3.25	3.92
BentSao-7B	21.55	19.94	20.92	22.75	19.56	16.81	20.62
SumSimiao-7B	38.75	44.37	43.81	33.26	37.05	31.34	38.51
IvyGPT-13B	37.70	43.56	40.47	38.08	35.31	32.16	37.88
DISC-MedLLM-13B	42.25	46.28	38.44	33.88	39.10	31.44	38.73
HuatuoGPT-II-7B	64.55	63.75	64.04	62.67	63.08	54.31	62.07
HuatuoGPT-II-13B	67.85	66.12	66.19	64.06	65.40	<u>59.62</u>	64.87
Ours (8B)	**73.75**	**80.56**	**73.91**	<u>71.65</u>	**71.75**	**69.43**	**74.40**

Table 2: Accuracy of general-purpose and medical models on different categories of the CMB-exam dataset. (Bold indicates the best result in each column, and underlined numbers denote the second-best.)

and provides critiques for refinement until an acceptable answer is found.

Infrastructure. Experiments run on a single NVIDIA RTX 4090 (24GB) GPU using PyTorch 2.7.0 and Transformers 4.52.4.

5 Results and Analysis

5.1 Main Result

Table 1 reports performance on three Chinese medical benchmarks. Our model achieves 78.86%, 84.15%, and 78.51% accuracy on CMExam, MedQA-CN, and CMMLU-Med respectively, establishing strong performance among Chinese medical LLMs.

CAMEC consistently outperforms strong general-purpose and medical LLM baselines across all three benchmarks. In particular, it surpasses Qwen2.5-7B by 8.46, 12.75, and 8.01 points on CMExam, MedQA-CN, and CMMLU-Med respectively, highlighting the importance of domain-specific adaptation for medical question answering.

More importantly, CAMEC outperforms the previous state-of-the-art HuatuoGPT-o1-7B by 4.76, 4.35, and 4.01 points on CMExam, MedQA-CN, and CMMLU-Med respectively. These consistent improvements across diverse benchmarks indicate strong generalization capability rather than dataset-specific adaptation, and validate the effectiveness of combining internal reasoning (E_{CoT}), retrieval-augmented evidence (E_{RAG}), and knowledge graph constraints (E_{KG}) with judge-guided consensus.

5.2 Subtask Analysis

Table 2 shows performance on six CMB-exam sub-tasks. Our model achieves 74.40% average accuracy, outperforming the strongest general model Yi-34B-Chat (69.17%) by 5.23 points and the strongest medical model HuatuoGPT-II-13B

Ablation Setting				CMExam
Fine-Tune	Reason Expert	RAG Expert	KG Expert	Accuracy(%)
-	-	-	-	72.30
✓	-	-	-	73.91
-	✓	✓	✓	77.14
✓	-	✓	✓	77.05
✓	✓	-	✓	76.16
✓	✓	✓	-	76.83
✓	✓	✓	✓	**78.86**

Table 3: Ablation study results on the CMExam dataset.

Level	Count	Ratio
LOW	413	12.06%
MEDIUM	1487	43.40%
HIGH	1526	44.54%
TOTAL	**3426**	**100%**

Table 4: Complexity distribution predicted by the router model on MedQA-CN.

Setting	Avg Experts	Acc. (%)	Cost Reduction
Full Experts	3.00	84.36	0%
Ours (Router)	2.32	84.15	-22.51%

Table 5: Comparison of expert usage, accuracy, and inference cost under different settings.

(64.87%) by 9.53 points, demonstrating robust generalization across diverse medical subdomains.

Performance is particularly strong on clinically oriented tasks, with 73.75% on Physician and 80.56% on Nurse exams, while remaining balanced on pharmacology and diagnostic procedures (71–74%). While Yi-34B-Chat slightly outperforms our model on the Technician subtask, this category primarily emphasizes procedural knowledge and operational conventions, where large general-purpose models benefit from scale and surface-level recall. In contrast, our model shows clear advantages on clinically complex tasks. Even on the challenging Graduate Entrance Exam requiring cross-disciplinary reasoning, our model reaches 69.43%, exceeding HuatuoGPT-II-13B by 9.81 points and Yi-34B-Chat by 16.49 points. Overall, these results highlight stronger capability in complex clinical reasoning beyond simple factual or procedural recall.

5.3 Ablation Study

Table 3 reports ablation results on CMExam. To isolate the contribution of individual components, we disable the complexity-aware router for all configurations except the full system, ensuring uniform expert usage across test queries.

The base model without domain adaptation achieves 72.30% accuracy. Hierarchical SFT alone raises it to 73.91%, a gain of 1.61 points, confirming that progressive domain alignment is essential for medical QA.

Multi-expert collaboration without SFT reaches 77.14%, gaining 4.84 points over the base model, demonstrating that external knowledge substantially improves diagnostic decisions even without domain-tuned models. When combining SFT with expert subsets, we observe: (1) SFT + RAG + KG achieves 77.05%, showing that external knowledge sources (RAG and KG) together provide strong factual grounding; (2) SFT + CoT + KG reaches 76.16%, indicating that structured knowledge (KG) complements internal reasoning; (3) SFT + CoT + RAG attains 76.83%, the highest among two-expert combinations, suggesting RAG's strong evidence retrieval capability.

The full model incorporating all three experts and the router achieves 78.86%, corresponding to a 6.56-point improvement over the base model. Adding the third expert consistently improves performance over all two-expert combinations, demonstrating the complementary roles of CoT, RAG, and KG in enhancing reasoning, factual grounding, and structural consistency. These results validate our design choice of multi-expert collaboration with judge-guided synthesis for reliable medical QA.

5.4 Router Efficiency Analysis

Table 4 shows the distribution of complexity levels predicted by our router on MedQA-CN. The router

classifies 12.06% of queries as LOW, 43.40% as MEDIUM, and 44.54% as HIGH, indicating that nearly half of medical queries can be resolved without recruiting all three experts.

Table 5 compares the full expert configuration against our adaptive router. While the full system invokes all three experts for every query (average 3.00), our router reduces the average expert count to 2.32, achieving a 22.51% reduction in inference cost with only a minimal accuracy trade-off (84.36% $\rightarrow$ 84.15%, 0.21 points). This demonstrates that complexity-aware routing effectively balances diagnostic accuracy and computational efficiency, making CAMEC practical for real-world deployment.

6 Conclusion

We present CAMEC, a complexity-aware multi-expert collaboration framework for reliable Chinese medical question answering. CAMEC combines progressive three-stage LoRA-based SFT with a lightweight complexity-aware router and a unified LLM-as-a-Judge to adaptively recruit and refine CoT, RAG, and KG experts. Experiments on four authoritative Chinese medical benchmarks demonstrate consistent improvements over strong baselines with improved inference efficiency. Overall, CAMEC provides an effective and practical solution for robust medical QA in clinically complex settings.

Limitations

CAMEC is evaluated mainly on Chinese medical exam-style benchmarks, which may not fully reflect real-world clinical queries (e.g., noisy patient-generated inputs) or distribution shifts. The RAG and KG experts depend on the coverage, correctness, and timeliness of the underlying retrieval corpus/index and knowledge graph; missing or outdated entries can weaken grounding and completeness. The judge-guided multi-round inference improves reliability but increases latency and computation compared to single-pass generation. Finally, CAMEC does not explicitly quantify uncertainty, so plausible yet incorrect answers may still arise in underspecified cases.

Ethics Statement

This study uses publicly available datasets and knowledge sources and does not involve private patient records or human-subject data collection. CAMEC is intended for research/educational use and must not be treated as clinical advice; any real-world deployment requires qualified clinician oversight. While our multi-expert and judge-guided design aims to reduce hallucinations, residual risks (e.g., unsafe suggestions or biased behavior) may remain and warrant further safety and fairness evaluation.

Acknowledgments

This work was supported in part by the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang under Grant No. 2023C03180, the Research Funds of Hangzhou Institute for Advanced Study, UCAS under Grant No. 2024HIAS-V002, and the Zhejiang Provincial Key Laboratory for Sensitive Data Security Protection and Confidentiality Management under Grant No. 2024E10048.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie

Song, Wenya Xie, Chuyi Kong, and 1 others. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Rong Cheng, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. 2025. Dualrag: A dual-process approach to integrate reasoning and retrieval for multi-hop question answering. *arXiv preprint arXiv:2504.18243*.

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

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*.

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

Yu He Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Chang-Fu Kuo, and 1 others. 2025. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine*, 8(1):187.

Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, and 1 others. 2025. Tiered agentic oversight: A hierarchical multi-agent system for ai safety in healthcare. *arXiv preprint arXiv:2506.12482*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, and 1 others. 2025. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ian Robinson, Jim Webber, and Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, and 1 others. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 international conference on management of data*, pages 2614–2627.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and 1 others. 2024. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205.

Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, Anningzhe Gao, and Benyou Wang. 2025. Huatuo-26M, a large-scale Chinese medical QA dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3828–3848, Albuquerque, New Mexico. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Jiacheng Xie, Yang Yu, Yibo Chen, Hanyao Zhang, Lening Zhao, Jiaxuan He, Lei Jiang, Xiaoting Tang, Guanghui An, and Dong Xu. 2025. Bencao: An instruction-tuned large language model for traditional chinese medicine. *arXiv preprint arXiv:2510.17415*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. 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.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Erlan Yu, Xuehong Chu, Wanwan Zhang, Xiangbin Meng, Yaodong Yang, Xunming Ji, and Chuanjie Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

Yucheng Zhou, Lingran Song, and Jianbing Shen. 2025. MAM: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25319–25333, Vienna, Austria. Association for Computational Linguistics.

A System Overview

A.1 Multi-expert Judge-Guided Evaluation

Algorithm 1 Multi-expert Judge-Guided Evaluation

Input: Question q, Experts $\mathcal{E}$, Initial reports $C_0 = \{R_e^{(0)}\}$, Judge J, Max iterations $T_{\max} = 3$, Threshold $\tau = 8.0$, Weights $(\alpha, \beta, \gamma) = (0.6, 0.2, 0.2)$

Output: Final report R_f

$t \leftarrow 0$

while $t \leq T_{\max}$ **do**

  // Judge evaluates each expert report

  for each $e \in \mathcal{E}$ **do**

    $(s_{e,\text{corr}}^{(t)},\ s_{e,\text{comp}}^{(t)},\ s_{e,\text{safe}}^{(t)},\ \delta_e^{(t)}) \leftarrow J(C_t, p_{\text{score}})$

    $S_e^{(t)} \leftarrow \alpha s_{e,\text{corr}}^{(t)} + \beta s_{e,\text{comp}}^{(t)} + \gamma s_{e,\text{safe}}^{(t)}$

  end for

  // Check acceptance condition

  if $\max_e S_e^{(t)} \geq \tau$ **then**

    // Weighted consensus synthesis

    for each $e \in \mathcal{E}$ **do**

      $w_e \leftarrow \exp(S_e^{(t)}) \,/\, \sum_{e' \in \mathcal{E}} \exp(S_{e'}^{(t)})$

    end for

    $R_f \leftarrow \text{Synthesize}(C_t, \{w_e\})$

    break

  else if $t < T_{\max}$ **then**

    // Dimension-aware revision

    $t \leftarrow t + 1$

    for each $e \in \mathcal{E}$ **do**

      $R_e^{(t)} \leftarrow \text{Revise}(R_e^{(t-1)}, \delta_e^{(t-1)})$

    end for

    $C_t \leftarrow \{R_e^{(t)} : e \in \mathcal{E}\}$

  else

    // Max iterations reached, force synthesis

    for each $e \in \mathcal{E}$ **do**

      $w_e \leftarrow \exp(S_e^{(t)}) \,/\, \sum_{e' \in \mathcal{E}} \exp(S_{e'}^{(t)})$

    end for

    $R_f \leftarrow \text{Synthesize}(C_t, \{w_e\})$

    break

  end if

end while

return R_f

A.2 Case Study

We present three representative cases to demonstrate CAMEC's adaptive multi-expert collaboration: (1) **iterative refinement** when all experts initially fail to meet quality thresholds (MCS=2.6, Table 6), (2) **conflict resolution** when an expert produces clinically invalid output (MCS=2.4, Table 7), and (3) **efficient consensus** under medium complexity with only two experts activated (MCS=1.6, Table 8). Each case illustrates how the judge evaluates expert reports across correctness, completeness, and safety dimensions, provides targeted feedback for iterative improvement, and synthesizes final diagnoses through score-weighted consensus.

【用户问题】
65岁男性，反复胸闷、气短3个月，活动后加重，休息可缓解，既往有高血压病史10年。
【路由预测】
预测复杂度：高（MCS = 2.6）。激活专家：CoT、RAG、KG。
【**RAG**检索证据】
(1) 冠心病的典型症状：气短有活动时加重、休息时减轻的特点。
(2) 高血压是心血管疾病的重要危险因素，长期高血压影响心脏供血。
(3) 心肌缺血常见于老年人的冠状动脉粥样硬化。
【**KG**检索结果】
实体链接：胸闷、气短→ 心肌缺血/ 冠心病/ 心绞痛。
结构化属性：核心症状（胸闷、气短，活动后加重）、推荐检查（心电图、冠状动脉造影）。

【第一轮生成】
【**CoT**专家报告】 (评分: 正确性7/10，完整性6/10，安全性7/10 → 总分**6.8/10**)
患者是一位65岁的男性，反复出现胸闷和气短3个月，活动之后症状加重，休息后可以缓解，同时有多年高血压病史，这种情况考虑可能与心脏供血问题有关，比如【心肌缺血】。建议进一步做心电图等检查明确情况。
Judge反馈：该报告仅给出笼统"心肌缺血"判断，未明确利用"活动诱发、休息缓解"这一关键特征建立到稳定型心绞痛的诊断推理链，也缺乏必要的鉴别诊断说明。
【**RAG**专家报告】 (评分: 正确性8/10，完整性6/10，安全性8/10 → 总分**7.6/10**)
根据症状，患者可能是【冠心病导致的心肌缺血】，结合检索到的证据，老年人易出现此类症状...
Judge反馈：报告仅笼统诊断为"冠心病导致的心肌缺血"，未进一步明确这符合稳定型心绞痛的诊断标准（即劳力诱发、静息缓解的特征性症状）。应整合检索证据建立完整诊断推理链。
【**KG**专家报告】 (评分: 正确性6/10，完整性7/10，安全性7/10 → 总分**6.4/10**)
患者有高血压病史，出现胸闷气短症状，可能是【高血压导致的心脏负担增加】...
Judge反馈：过度关注高血压本身，未整合症状时序特征（活动诱发、休息缓解）得出明确诊断，需补充症状-疾病关联推理。
Judge决策：所有专家评分均低于阈值（τ=8.0），启动迭代优化。

【第二轮生成（基于反馈改进）】
【**CoT**专家报告】 (评分: 正确性9/10，完整性9/10，安全性9/10 → 总分**9.0/10**) ✓
患者表现为典型的劳力性胸闷气短，具有"活动后加重、休息后缓解"的规律，持续时间较短，符合【稳定型心绞痛】的典型表现。结合65岁年龄和10年高血压病史（冠心病主要危险因素），诊断为冠心病可能性大。鉴别诊断需排除肺源性疾病（但患者无咳嗽、咳痰）和胃食管反流（无烧心症状）。建议完善心电图、运动负荷试验或冠脉造影确诊。
【**RAG**专家报告】 (评分: 正确性9/10，完整性8/10，安全性8/10 → 总分**8.6/10**) ✓
根据检索证据，冠心病的典型症状为"气短有活动时加重、休息时减轻"，与患者症状高度一致。患者为65岁男性，合并高血压（重要危险因素），符合【稳定型心绞痛】的高发人群特征。建议及时就医进行心电图和心肌酶检查，必要时行冠状动脉造影明确病变程度。同时需严格控制血压，减少心绞痛发作风险。
【**KG**专家报告】 (评分: 正确性8/10，完整性7/10，安全性8/10 → 总分**7.8/10**)
患者为65岁男性，反复出现胸闷、气短约3个月，症状在活动时加重、休息后缓解。根据检索到的知识图谱，"胸闷+ 气短+ 劳力诱发+ 静息缓解"与心肌缺血/心绞痛/冠心病实体呈高强度关联，同时其既往10年高血压病史进一步提高冠状动脉粥样硬化及心肌缺血发生概率。因此当前最可能诊断为【心肌缺血（冠心病）】。建议进行心电图及冠状动脉相关检查以进一步明确。
Judge评价：诊断方向正确，推理深度与临床完整性仍略弱。

【**Judge**综合评审】
CoT和RAG专家评分均达标（8.5-9.0分），诊断高度一致（冠心病/稳定型心绞痛），推理链完整且充分利用了症状时序特征。KG专家虽有改进但未达阈值，不参与最终融合。
【最终诊断报告】
根据患者提供的临床资料，65岁男性，反复出现胸闷、气短3个月，活动后加重，休息可缓解，既往有高血压病史10年，这些表现符合【稳定型心绞痛】的典型特征。患者症状具有劳力诱发、静息缓解的规律性，结合高血压等危险因素，诊断为冠心病所致稳定型心绞痛可能性大。建议进一步进行心电图、心肌酶谱及冠状动脉造影等检查，及时就医并由心血管专科医生进行评估与治疗。

Table 6: High-complexity case demonstrating judge-guided iterative refinement when none of the experts initially meet the acceptance threshold.

【用户问题】
72岁女性，腰痛放射至双下肢，站立或行走时加重，坐下或前屈时缓解。
【路由预测】
预测复杂度：高（MCS = 2.4）。激活专家：CoT、RAG、KG。
【**RAG**检索证据】
(1) 腰椎间盘突出症常表现为腰痛和下肢放射痛，疼痛可沿臀部、大腿后侧、小腿外侧至足跟。
(2) 腰痛症状：90%以上的患者均有下腰部及腰骶部持久性钝痛，平卧位时疼痛可减轻，站立位及坐位时疼痛加重。
(3) 活动加重、前屈或坐位缓解提示机械性神经根受压，符合腰椎间盘突出的典型特征。
【**KG**检索结果】
实体链接：腰痛→ 腰椎间盘突出、腰椎病、胸腰椎骨折、腰椎结核。
核心症状：腰痛+下肢放射痛+活动诱发+静息缓解→ 腰椎间盘突出症（强关联）。
推荐检查：腰椎MRI/CT、直腿抬高试验、脊柱活动度检查。
治疗策略：保守治疗（卧床休息、理疗、镇痛药）或手术治疗。
【**CoT**专家报告】 (评分: 正确性9/10，完整性9/10，安全性9/10 → 总分**9.0/10**) ✓
患者以腰痛为主，并且疼痛向双下肢放射，同时表现为站立或行走时症状加重，而坐下或身体前屈时明显缓解，这是一种非常典型的机械性神经根受压表现。考虑到患者72岁，年龄相关腰椎退行性改变较为常见，结合双下肢放射痛特点，更加支持【腰椎间盘突出并压迫神经根】的诊断。需要进一步完善腰椎MRI检查明确病变节段及压迫程度，同时排除骨折、感染等其他原因。
【**RAG**专家报告】 (评分: 正确性4/10，完整性5/10，安全性6/10 → 总分**4.6/10**)
患者存在腰痛并向下肢放射，且站立或行走时疼痛加重，平静或坐下时症状缓解。根据搜索到的医学资料，这种情况常见于腰部疾病，比如腰椎疾病引起的神经受压。资料显示，许多患者在类似情况下都会出现下肢放射痛等表现。因此，患者目前可能属于【下肢放射痛】的情况，需要进一步检查以明确原因。
Judge反馈：该专家报告主要停留在症状层面，将“下肢放射痛”错误作为诊断结果输出，缺乏从症状到疾病的推理链，也没有将检索证据与患者具体表现进行结构化整合，因此正确性和完整性较弱。
【**KG**专家报告】 (评分: 正确性9/10，完整性7/10，安全性8/10 → 总分**8.4/10**) ✓
基于知识图谱推理结果，患者当前表现出的“腰痛+ 双下肢放射痛+ 站立/行走加重+ 坐下或前屈缓解”的症状组合，与【腰椎间盘突出症】具有高度匹配度。该疾病与“神经根压迫”存在强关联路径，同时患者年龄较大，也符合退行性椎间盘病变的高发人群特征。结合知识图谱推荐信息，进一步检查应包括腰椎MRI或CT评估压迫程度，并根据病情选择保守治疗或手术干预。整体推理结构清晰且符合临床路径。
【**Judge**综合决策】
RAG专家将症状误当作诊断输出，缺乏从症状到疾病的完整推理链，因此不具备纳入最终融合的条件；而CoT专家（9.0分）与KG专家（8.4分）在诊断结论上高度一致，且推理过程符合临床规范，综上，Judge优先采纳两者意见并进行加权融合，形成最终共识诊断。
【最终诊断报告】
根据患者症状，结合医学知识库与相关文献分析，最可能的诊断为【腰椎间盘突出症】。该病常见于... 尽管腰椎间盘突出是最可能的诊断，但仍需考虑其他可能病因，如胸腰椎骨折、腰椎结核、肾积水等。因此，建议患者进行进一步的影像学检查（如MRI或CT），以明确诊断并制定个体化治疗方案。

Table 7: High-complexity case demonstrating judge-based conflict resolution and reliability gating when an expert outputs symptom–disease hierarchy confusion.

【用户问题】
45岁男性，反复上腹部隐痛3个月，餐后加重，伴反酸、嗳气，无明显体重下降。
【路由预测】
预测复杂度：中等（MCS = 1.6）。激活专家：CoT、RAG。
【**RAG**检索证据】
(1) 慢性胃炎常表现为上腹隐痛、餐后不适、反酸和嗳气。
(2) 胃食管反流病以反酸、烧心、餐后加重为典型表现。
(3) 消化性溃疡可表现为烧心、消化道穿孔、溃疡疼痛等症状。
(4) 胆石症主要症状为腹痛、畏寒、发热及黄疸。
【**CoT**专家报告】 (评分: 正确性9/10，完整性9/10，安全性9/10 → 总分**9.0/10**) ✓
根据病人描述的症状，包括反复上腹部隐痛、餐后加重、反酸、嗳气等，结合病程3个月，这些症状与消化系统疾病有关。其中，反酸和嗳气是【胃食管反流病】的典型症状，而上腹部隐痛和餐后加重则可能与【胃炎或胃溃疡】有关。建议患者到医院进行详细检查，如胃镜等，以明确诊断。同时，应注意避免辛辣食物、饮酒和吸烟等刺激因素。
【**RAG**专家报告】 (评分: 正确性8/10，完整性8/10，安全性8/10 → 总分**8.0/10**) ✓
根据病人的症状描述，上腹部隐痛、餐后加重、伴反酸、嗳气，这些症状与【慢性胃炎或胃溃疡】的表现相符。参考资料中提到，慢性胃炎和胃溃疡可以导致左上腹部不适，包括饭前和饭后疼痛、腹胀、反酸等。建议患者到医院就诊，进行胃镜检查以明确诊断，并根据检查结果进行相应的治疗。平时应注意饮食，避免暴饮暴食和刺激食物，保持规律的饮食习惯。
【**Judge**综合评审】
专家1（9.0分）和专家2（8.0分）诊断高度一致，均超过阈值（τ=8）。Judge综合两个专家意见，优先采纳高分专家1的主要诊断（胃食管反流病），同时保留专家2提出的重要鉴别诊断（慢性胃炎/消化性溃疡）。
【最终诊断报告】
患者提供的病史症状与【胃食管反流病】的典型表现高度一致，尤其是餐后症状加重、反酸、嗳气均为该病常见临床特征。此外，患者无明显体重下降，排除了恶性肿瘤或其他严重器质性疾病的高风险因素，进一步支持胃食管反流病的可能性。考虑到慢性胃炎或胃溃疡也可能表现为类似症状（如餐后腹痛、反酸），不能完全排除这些诊断。建议进行胃镜检查以明确是否存在胃黏膜炎症、溃疡或其他病变。初步诊断建议为：胃食管反流病，待胃镜检查结果以确认或排除其他可能性。同时，建议患者调整饮食结构，避免辛辣、油腻食物，规律进食，避免过饱，减少夜间进食，以缓解症状。如症状持续或加重，应及时就医并完善相关检查。

Table 8: Medium-complexity case demonstrating multi-expert collaboration and judge-guided weighted consensus under partial expert disagreement.

B Prompt Engineering

This section presents the complete prompt templates used in our CAMEC framework: the complexity router prompt (§B.1), expert prompts (§B.2), and judge prompt (§B.3). These prompts define the input–output format, reasoning constraints, and evaluation criteria for each module.

B.1 Router Prompt Template

Tables 9 and 10 present the prompt templates used by the Qwen3-0.6B router to predict Medical Complexity Scores (MCS) and three-way complexity labels (LOW/MEDIUM/HIGH). Table 9 shows the original Chinese prompt used in training and inference, while Table 10 provides an equivalent English version for reproducibility. Both prompts define four scoring dimensions (symptom ambiguity, disease rarity, comorbidity risk, diagnostic uncertainty), as well as the JSON output format described in §3.2.

作为医学专家，请根据医学诊断复杂度模型（Medical Complexity Score, MCS）对以下问题进行评估。

问题：sample[’Question’]

请从以下维度打分（0-3分）：
1. 症状模糊度Symptom Ambiguity：症状是否非特异性、描述是否完整
2. 疾病罕见度Disease Rarity：涉及的疾病是否罕见
3. 并发症风险Comorbidity Risk：是否涉及多系统、有既往病史
4. 诊断不确定性Diagnostic Uncertainty：鉴别诊断数量、需要的检查复杂度

{最终复杂度分级规则(基于加权和MCS)}
MCS = 0.3·Ambiguity + 0.25·Rarity + 0.25·Comorbidity + 0.2·Uncertainty
- LOW (0 ⩽MCS < 1): 单一症状，常见病，无并发症，诊断较明确
- MEDIUM (1 ⩽MCS < 2): 多症状、需要鉴别诊断，一般复杂度
- HIGH (2 ⩽MCS ⩽3): 症状复杂、罕见病、多系统问题、诊断依赖检查

返回严格JSON格式：
{{
"symptom_ambiguity": <score>,
"disease_rarity": <score>,
"comorbidity_risk": <score>,
"diagnostic_uncertainty": <score>,
"overall_level": <"LOW"/"MEDIUM"/"HIGH">,
"reasoning": "<简短解释>"
"Question": sample[’Question’]
}}

Table 9: Prompt Template for Complexity-Aware Router (Chinese Version)

You are a medical expert. Please evaluate the following question based on a Medical Complexity Score (MCS) framework.

Question: sample['Question']

Please assign scores (0–3) for each of the following dimensions:
1. Symptom Ambiguity: Whether the symptoms are non-specific or incompletely described
2. Disease Rarity: Whether the condition involved is rare
3. Comorbidity Risk: Whether multiple systems or prior medical history are involved
4. Diagnostic Uncertainty: The number of differential diagnoses and the complexity of required examinations

Final Complexity Level (based on weighted MCS):
MCS = 0.3 · Ambiguity + 0.25 · Rarity + 0.25 · Comorbidity + 0.2 · Uncertainty
- LOW ($0 \leq$ MCS < 1): Single symptom, common disease, no comorbidity, clear diagnosis
- MEDIUM ($1 \leq$ MCS < 2): Multiple symptoms, requires differential diagnosis, moderate complexity
- HIGH ($2 \leq$ MCS ≤ 3): Complex symptoms, rare diseases, multi-system involvement, diagnosis relies on examinations

Return the result in strict JSON format:
{{
"symptom_ambiguity": <score>,
"disease_rarity": <score>,
"comorbidity_risk": <score>,
"diagnostic_uncertainty": <score>,
"overall_level": "<LOW/MEDIUM/HIGH>",
"reasoning": "<brief explanation>",
"question": "sample['Question']"
}}

Table 10: Prompt Template for Complexity-Aware Router (English Version)

B.2 Expert Prompt

Prompt templates used for the three expert modules (CoT, RAG, and KG) in our multi-expert inference framework are provided in Tables 11 and 12. Table 11 shows the original Chinese prompts used in all experiments, while Table 12 provides equivalent English versions for reproducibility. These templates define the role, input–output format, and reasoning constraints for each expert, and maintain identical structure and requirements across both versions.

Prompt for CoT Expert (p_{CoT}): 你是一名医学推理专家，你的任务是基于题干中出现的所有临床信息（如症状、体征、病程、检查结果等），构建一个清晰、逐步的推理过程，并给出一个由题干信息支持的诊断候选。
请注意：
- 推理需基于题干中出现的信息；
- 不得加入题干未提及的具体医学事实；
- 推理过程应逻辑一致、结构清晰。
【输入】
问题：{Question}
【输出格式】
推理过程：
1. <逐条分析题干中的关键医学信息>
2. <解释每条信息可能指向的临床意义>
3. <说明不同线索之间的逻辑联系>
4. <排除与题干信息不符的方向>
诊断候选：<由题干信息支持的诊断方向>

Prompt for RAG Expert (p_{RAG}): 你是一名循证医学信息整合专家，你的任务是基于患者症状描述与检索到的医学文献证据，提供基于证据的诊断分析。
请注意：
- 优先使用参考资料中与患者症状高度相关的内容作为证据；
- 若检索内容与症状存在矛盾或不相关，应明确指出并依赖患者信息；
- 引用参考资料时需标注来源，确保论断可追溯；
- 结合医学推理能力，确保诊断逻辑一致性。
【输入】
患者问题：Question
参考资料（检索节选）：context
【输出格式】
证据分析：<评估参考资料质量，提取与患者症状相关的核心证据>
诊断结论：<最符合证据与临床逻辑的诊断，包含推理依据>

Prompt for KG Expert (p_{KG}): 你是一名医学知识图谱推理专家，你的任务是基于题干内容与提供的知识图谱结构，从结构化知识约束与因果路径角度分析该单选题。
请注意：
- 优先利用知识图谱中与题干匹配的疾病-症状、检查-治疗关系路径；
- 若图谱覆盖不完整，可结合医学推理补充，但需明确标注图谱支持程度；
- 识别图谱与题干的冲突或不一致，并在推理中说明；
- 确保推理链的临床合理性优先于单纯的图谱匹配。
【输入】
问题：{Question}
知识图谱（子图）：{graph_info}
【输出格式】
关键路径分析（1-2句）：<从知识图谱中识别出的关键因果路径或关联关系>
诊断候选：<基于图谱结构支持的诊断方向>

Table 11: Prompt Template for Specialized Experts (Chinese Version)

Prompt for CoT Expert (p_{CoT}): You are a medical reasoning expert. Your task is to construct a clear and step-by-step reasoning process based solely on the clinical information provided in the question (e.g., symptoms, signs, disease course, and examination results), and propose a diagnosis candidate supported by the given information.
Please note:
- The reasoning must strictly rely on the information explicitly stated in the question;
- Do not introduce external or unstated medical facts;
- The reasoning process should be logically consistent and well-structured.

Input:
Question: {Question}
Output Format:
Reasoning Process:
1. <Analyze key clinical information in the question>
2. <Explain the clinical implications of each piece of information>
3. <Describe the logical relationships among different clues>
4. <Rule out inconsistent or unsupported possibilities>

Diagnosis Candidate: <Diagnosis supported by the given information>

Prompt for RAG Expert (p_{RAG}): You are an evidence-based medical analysis expert. Your task is to provide a diagnosis grounded in both the patient's symptoms and the retrieved medical evidence.
Please note:
- Prioritize evidence from the references that is highly relevant to the patient's symptoms;
- If the retrieved content is irrelevant or contradictory, explicitly identify this and rely on the patient information instead;
- Clearly cite evidence sources to ensure traceability;
- Maintain logical consistency in clinical reasoning.

Input:
Patient Question: {Question}
Retrieved Evidence: {context}
Output Format:
Evidence Analysis: <Evaluate the quality of references and extract relevant evidence>
Diagnosis Conclusion: <Diagnosis supported by evidence and clinical reasoning>

Prompt for KG Expert (p_{KG}): You are a medical knowledge graph reasoning expert. Your task is to analyze the question using structured medical knowledge and reasoning over the provided knowledge graph.
Please note:
- Prioritize disease–symptom and examination–treatment relations from the knowledge graph;
- If the graph coverage is incomplete, you may supplement with medical reasoning, but clearly indicate the level of graph support;
- Identify and explain any inconsistencies between the graph and the question;
- Ensure clinical plausibility takes precedence over simple graph matching.

Input:
Question: {Question}
Knowledge Graph (subgraph): {graph_info}
Output Format:
Key Path Analysis (1–2 sentences): <Key causal or relational paths identified from the graph>
Diagnosis Candidate: <Diagnosis supported by the graph structure>

Table 12: Prompt Template for Specialized Experts (English Version)

B.3 Judge Prompt

Tables 13 and 14 present the prompt templates for the LLM-as-a-Judge module. Table 13 shows the original Chinese prompt used in all experiments, while Table 14 provides an equivalent English version for reproducibility. The judge evaluates expert reports on three dimensions (correctness, completeness, safety), provides dimension-wise feedback for iterative refinement, and synthesizes a final consensus report without introducing new medical facts, maintaining identical evaluation criteria and constraints across both versions.

你是一名医学报告评估与整合专家，负责对激活的专家提供的诊断报告进行评价与整合，且不能生成任何新的医学知识。

【任务说明】

1. 对每位专家的报告分别进行三维评分（正确性、完整性、安全性），并计算加权总分。
2. 若有专家得分均低于阈值（<8），在feedback中给出改进建议，供专家修订。
3. 若至少一位专家得分达到阈值（⩾8），综合高分专家内容生成最终诊断报告：
   - 内容只能来自专家文本，不引入新的医学事实或推理；
   - 可进行重写、合并、去重、冲突处理和结构化整理；
   - 保留高分专家判断，删除矛盾或错误内容。

【输入】

问题：{Question}

激活专家及其报告：{expert_reports}

（注：根据问题复杂度，可能激活CoT、RAG、KG中的1-3位专家）

【评分要点】

评分基于以下三个维度：

（1）正确性（correctness）：诊断与患者症状的一致性，是否存在事实错误或逻辑矛盾；

（2）完整性（completeness）：是否充分覆盖关键临床信息，推理链是否完整；

（3）安全性（safety）：是否存在过度断言、潜在风险或临床不安全表述。

综合评分计算：$S_e = 0.6 \times s_{e,\text{corr}} + 0.2 \times s_{e,\text{comp}} + 0.2 \times s_{e,\text{safe}}$

【输出格式（JSON）】

```
{
  "dimension_scores": {
    "expert1": {"correctness": X1, "completeness": Y1, "safety": Z1},
    "expert2": {"correctness": X2, "completeness": Y2, "safety": Z2},
    "expert3": {"correctness": X3, "completeness": Y3, "safety": Z3}
  },
  "aggregated_scores": {"expert1": X, "expert2": Y, "expert3": Z},
  "feedback": {
    "expert1": "<综合反馈>",
    "expert2": "<综合反馈>",
    "expert3": "<综合反馈>"
  },
  "decision": "refine" | "synthesize",
  "synthesized_report": "<仅当decision=synthesize时输出>"
}
```

说明：当decision=refine时，返回feedback供专家修订；当decision=synthesize时，返回基于专家内容综合的最终报告。

Table 13: Prompt template for the LLM-as-a-Judge module, including three-dimensional evaluation, iterative feedback generation, and consensus synthesis (Chinese Version)

You are a medical report evaluation and integration expert. Your role is to assess and synthesize the diagnostic reports provided by activated experts. You must not introduce any new medical knowledge beyond the provided expert outputs.

Task Description

1. Evaluate each expert report along three dimensions (correctness, completeness, safety) and compute a weighted aggregate score.
2. If all expert scores fall below the threshold (< 8), provide feedback for each expert to guide refinement.
3. If at least one expert meets the threshold (≥ 8), synthesize a final diagnostic report by integrating high-quality expert outputs:
   - The content must be strictly derived from expert reports without introducing new medical facts or reasoning;
   - You may rewrite, merge, deduplicate, resolve conflicts, and restructure the content;
   - Retain high-quality conclusions and remove inconsistent or incorrect statements.

Input

Question: {Question}
Activated Experts and Reports: {expert_reports}
(Note: Depending on question complexity, 1–3 experts among CoT, RAG, and KG may be activated.)

Scoring Criteria

Each expert is evaluated along three dimensions:
(1) Correctness: Consistency between the diagnosis and patient symptoms; absence of factual errors or logical contradictions;
(2) Completeness: Coverage of key clinical information and coherence of the reasoning chain;
(3) Safety: Absence of overconfident claims, potential risks, or clinically unsafe statements.
The aggregated score is computed as:
$S_e = 0.6 \times s_{e,\text{corr}} + 0.2 \times s_{e,\text{comp}} + 0.2 \times s_{e,\text{safe}}$

Output Format (JSON)

```
{
  "dimension_scores": {
    "expert1": {"correctness": X1, "completeness": Y1, "safety": Z1},
    "expert2": {"correctness": X2, "completeness": Y2, "safety": Z2},
    "expert3": {"correctness": X3, "completeness": Y3, "safety": Z3}
  },
  "aggregated_scores": {"expert1": X, "expert2": Y, "expert3": Z},
  "feedback": {
    "expert1": "<feedback>",
    "expert2": "<feedback>",
    "expert3": "<feedback>"
  },
  "decision": "refine" | "synthesize",
  "synthesized_report": "<only when decision = synthesize>"
}
```

Note: If the decision is "refine", return feedback for expert revision; if the decision is "synthesize", return the final integrated report based solely on expert outputs.

Table 14: Prompt Template for the LLM-as-a-Judge Module, including multi-dimensional evaluation, iterative feedback, and consensus-based synthesis (English Version)

C Training Configuration

C.1 Training Data Statistics and Preprocessing

We fine-tune Qwen3-8B using three Chinese medical datasets aligned with our three-stage SFT schedule: Huatuo26M-Lite (Wang et al., 2025) (Domain Adaptation), HuatuoGPT2-SFT-GPT4-140K (Chen et al., 2023) (Instruction Following), and medical-o1-reasoning-SFT (Chen et al., 2024) (Clinical Reasoning). We retain only Chinese samples and filter noisy/non-medical instances.

Preprocessing. We (i) keep native Chinese samples, (ii) remove empty/garbled entries, and (iii) discard clearly non-medical content. We also ensure no overlap with evaluation benchmarks by applying a conservative filtering strategy to remove samples that may overlap with the evaluation sets, thereby reducing the risk of potential data leakage.

Dataset	Stage	#Samples	Avg. Tokens
Huatuo26M-Lite	Domain Adapt.	177703	297.48
HuatuoGPT2-SFT-GPT4-140K	Instr. Follow.	142248	672.94
medical-o1-reasoning-SFT	Clinical Reason.	20171	1049.96

Table 15: Training data statistics after preprocessing.

Huatuo26M-Lite {
"instruction": "嗓子干疼，尤其是吞咽拭感觉疼得厉害，是什么原因，去医院检查说是典型的扁桃体炎，但吃药后疼痛没有减轻，想问问到底是什么病，应该怎么治疗？",
"input": "",
"output": "根据症状描述，可能是扁桃体炎引起的喉咙疼痛，但是吃药后疼痛没有减轻，可能需要进一步检查确定病因。此外，建议患者注意饮食调理，少吃辛辣油炸食物，多吃清热食物。"}
HuatuoGPT2-SFT-GPT4-140K {
"instruction": "我因为和朋友最近喝酒，在某次饭局中出现了急性胃出血，想问问急性胃出血需要住院接受治疗吗？",
"input": "",
"output": "急性胃出血是一种严重的医疗状况，需要立即就医。这种情况可能由于胃溃疡、胃炎、胃癌或者长期过量饮酒等因素引起。症状可能包括呕吐鲜血或咖啡色物质，黑色粘稠的大便，以及可能伴随的虚弱、头晕、心跳加速等。如果你怀疑自己有急性胃出血，应立即去医院急诊部门。医生可能会进行胃镜检查以确定出血的原因和位置，然后进行相应的治疗。治疗可能包括药物治疗，比如质子泵抑制剂（如奥美拉唑）来减少胃酸分泌，或者可能需要内镜下的手术治疗，如止血或者移除溃疡等。因此，对于你的问题，急性胃出血确实需要住院接受治疗。这是因为这种状况可能导致严重的并发症，如休克、贫血或者胃穿孔等，需要在医生的密切监控下进行治疗。同时，医生也会建议你改变生活习惯，比如戒酒、戒烟、避免使用非甾体抗炎药物等，以防止再次出血。"}
medical-o1-reasoning-SFT {
"instruction": "38岁女性右手被开水烫伤10分钟，局部红润无水泡，减轻疼痛应该采取什么措施？",
"input": "哦，38岁的女性手上被开水烫了一下，这肯定很疼。看来时间还不久，只有10分钟，得赶紧处理。嗯，局部红润但没有起水泡，那应该不算太严重。通常，这意味着只是表皮的浅表烧伤，可能是一度烧伤。如果是这样，首先想到的就是赶紧给它降温。用流动的冷水冲洗是个不错的选择。这样做的目的是为了给皮肤降温，缓解疼痛，也能减少进一步的损伤。不过，不能用冰块，虽说冰冷得快，但直接用冰可能会冻坏皮肤。还有，不能涂抹油类的东西，听说有的人喜欢抹牙膏或酱油这类东西，但这样热量反而不容易散发出来。所以，最好的做法是在流动的冷水下冲洗或泡着手，大约15到20分钟，这样她的疼痛应该会有所缓解。这样处理完全符合我们日常听到的急救常识，也是医生通常建议的。这下可以放心，这样应该能帮她减轻疼痛和防止更严重的后果。",
"output": "对于38岁女性的烫伤情况的正确处理步骤如下：1. **降温处理**：在烫伤后立即用流动的冷水冲洗受伤部位是非常有效的方法。持续用冷水冲洗15到20分钟，可以有效降温，缓解疼痛，并减少进一步的组织损害。要注意避免使用冰块直接接触皮肤，因为冰块可能会导致冻伤。2. **避免使用不当治疗**：请勿在烫伤处涂抹牙膏、酱油或其他民间偏方，因为这些物质不能有效散热，反而可能导致感染或其他并发症。3. **保持清洁**：冷却后，轻轻拍干受伤部位，避免揉搓。保持烫伤区域清洁和干燥，避免感染。4. **观察伤情**：如果烫伤面积较大，或有起水泡、创面变痛等情况发生，建议及时就医，以便专业医疗人员进行进一步处理。通过按照这些步骤来处理，这位女士可以在一定程度上缓解疼痛，并降低烫伤恶化的风险。"}

Table 16: Representative samples from the three-stage hierarchical SFT pipeline.

C.2 Training Hyperparameters

Hierarchical SFT. We fine-tune Qwen3-8B with LoRA in three stages. Table 17 reports key hyperparameters.

Hyperparameter	Value
Model	Qwen3-8B
LoRA rank	16
LoRA dropout	0.05
Max sequence length	2048
Batch size	2
Gradient accumulation	4
Precision	bfloat16
Optimizer	AdamW
Scheduler	cosine
Epochs (each stage)	1
Learning rates	1.5e-5 / 1e-5 / 5e-6

Table 17: Key hyperparameters for hierarchical SFT.

Router fine-tuning. We fine-tune Qwen3-0.6B with LoRA as a 3-way classifier (LOW/MEDIUM/HIGH) over 8,000 labeled queries. The model is trained with cross-entropy loss, and Table 18 reports the hyperparameters.

Hyperparameter	Value
Model	Qwen3-0.6B
LoRA rank	8
LoRA dropout	0.05
Max sequence length	512
Batch size	4
Gradient accumulation	4
Precision	bfloat16
Optimizer	AdamW
Loss	Cross-entropy
Epochs	3
Learning rate	1e-4

Table 18: Hyperparameters for training the complexity router.

C.3 Router Training Data Examples

Table 19 shows representative samples from the 8,000 automatically annotated training instances for the complexity-aware router (§3.2). Each sample includes: (1) the original medical question, (2) four dimensional scores (0–3) assessing symptom ambiguity, disease rarity, comorbidity risk, and diagnostic uncertainty, (3) natural language reasoning explaining the scoring rationale, and (4) the derived overall complexity label (LOW/MEDIUM/HIGH). These annotations were generated by prompting DeepSeek-R1 to analyze each question along our complexity assessment framework.

Example 1:{
"symptom_ambiguity": 2,
"disease_rarity": 1,
"comorbidity_risk": 3,
"diagnostic_uncertainty": 3,
"overall_level": "HIGH",
"reasoning": "患者表现出失眠、被害妄想、拒食拒药等精神症状，可能与风湿性心脏病相关的精神障碍或药物副作用有关，症状具有一定的非特异性；风湿性心脏病并不罕见，但合并精神症状时诊断复杂度增加；患者有基础心脏病史，可能涉及神经系统、代谢或心理多系统问题，并可能存在治疗矛盾；需鉴别精神疾病（如精神分裂症）、药物反应、代谢紊乱等多种可能性，需结合实验室、影像学及精神评估，检查复杂度高。",
"Question": "女性，30岁，因风湿性心脏病住院，经治疗效果不好。近1周来失眠，认为医生和护士以及同室病友都在背地里议论她，说她的病治不好了，不用治了。服了主管刘医生给的药以后感到难受，怀疑有毒，为此紧张、拒食、拒药，要求立即出院。根据此临床症状。考虑其诊断可能是（）A.分裂症样精神障碍　B.癔病　C.心脏病所致精神障碍　D.偏执性精神病"
}
Example 2:{
"symptom_ambiguity": 2,
"disease_rarity": 0,
"comorbidity_risk": 3,
"diagnostic_uncertainty": 1,
"overall_level": "MEDIUM",
"reasoning": "该患者为61岁女性，体检发现重度骨质疏松，症状可能隐匿或非特异性（如背痛、身高变矮等），描述相对完整。骨质疏松症常见，不属罕见病。但患者合并多种慢性疾病（食管裂孔疝、高血压、2型糖尿病）并长期服用多种药物，涉及多系统并发症风险，尤其需关注药物对骨代谢的潜在影响（如PPI类药物与骨密度下降的关系）。骨质疏松诊断已明确，鉴别诊断和检查复杂度较低，但健康教育内容需个体化评估，存在一定的诊断不确定性。",
"Question": "患者，女，61岁，体检发现重度骨质疏松，既往患有食管裂孔疝、高血压、2型糖尿病。长期服用奥美拉唑肠溶片、厄贝沙坦-氢氯噻嗪片，氨氯地平片、阿卡波糖片、阿司匹林肠溶片。实验室检查结果：肝、肾功能正常。针对该患者骨质疏松症健康教育的说法，错误的是（）A.推荐低盐、富含钙和蛋白质的均衡膳食　B.戒烟、限酒，少饮咖啡和碳酸饮料　C.为降低骨折风险，建议少动或制动　D.每1～2年复查一次骨密度　E.每日上臂暴露日光浴15～20分钟"
}
Example 3:{
"symptom_ambiguity": 2,
"disease_rarity": 0,
"comorbidity_risk": 0,
"diagnostic_uncertainty": 0,
"overall_level": "LOW",
"reasoning": "该问题属于基础营养学知识，不涉及具体症状、罕见病或多系统并发症。问题明确，无需复杂鉴别诊断。",
"Question": "不能为机体提供能量的营养素是（）A.糖类　B.淀粉类　C.蛋白质类　D.维生素类"
}

Table 19: Representative training samples for the complexity-aware router. Each sample includes four dimensional scores (symptom ambiguity, disease rarity, comorbidity risk, diagnostic uncertainty), LLM-generated reasoning, and the derived complexity label (LOW/MEDIUM/HIGH).

D External Knowledge Source Configuration

D.1 RAG Corpus Construction

We construct a dense retrieval corpus from the huatuo_encyclopedia_qa dataset. After preprocessing (removing non-medical entries, deduplication, and filtering incomplete answers), the final corpus contains 364,420 QA pairs with an average of 760.65 tokens per concatenated document.

Each QA pair is concatenated as "question + answer" and encoded using Qwen3-Embedding-0.6B (1024-dim). Vectors are stored in Milvus v2.5.12 with an IVF_FLAT index (nlist=1024, inner-product metric). At inference, we retrieve k=3 nearest neighbors with nprobe=16. Table 20 summarizes corpus statistics, and Table 21 shows representative examples.

D.2 Knowledge Graph Construction

We construct a Chinese medical KG from publicly available medical knowledge sources. After preprocessing (removing incomplete entries and deduplication), the final KG contains 40,276 entities (including 8,808 diseases, 5,998 symptoms, 3,828 drugs, and other medical concepts) and 291,503 relations across 11 relation types, stored in Neo4j v5.18.1.

Subgraph Retrieval. For each query, we perform entity linking to identify mentioned medical terms, then retrieve relevant subgraphs using predefined Cypher templates. We define 16 templates covering common clinical intents (symptom inquiry, treatment recommendation, examination, etiology, etc.). For example, for a treatment inquiry about diabetes:

```
MATCH (d:Disease {name:''})
      -[:cure_way]->(t:Treatment)
RETURN t.name
```

Example entry. A representative disease entity from the source dataset:

> **name:** 肺泡蛋白质沉积症
> **symptom:** 紫绀, 胸痛, 呼吸困难, 乏力
> **department:** 内科, 呼吸内科
> **check:** 胸部CT, 肺活检
> **cure_way:** 支气管肺泡灌洗

Statistic	Value
Total QA pairs	364,420
Avg. tokens per document (Q+A)	760.65
Document format	“question + answer” concatenation
Embedding model	Qwen3-Embedding-0.6B
Index type	Milvus IVF_FLAT (nlist=1024)

Table 20: Statistics of the RAG corpus used for external evidence retrieval.

Example 1:{
"questions": "口干的治疗方案是什么?",
"answers": "口干症的治疗包括病因治疗和对症治疗。对因治疗在明确病因的情况下是最有效的，如药物性口干，通过调整药物及其剂量，可缓解口干。对唾液消耗增加而产生的口干，可通过消除张口呼吸等原因来解决。如果是由于唾液腺实质破坏所引起的口感，如头颈部恶性肿瘤放疗后、舍格伦综合征，目前主要通过对症治疗来缓解口干，减少并发症。"
}
Example 2:{
"questions": "酒精性脂肪肝什么症状",
"answers": "酒精性脂肪肝是一种长期酗酒引起的肝病，是一种酒精性肝病。病人有很长的饮酒史，通常超过5年。临床症状是非特异性的，可以是无症状的，也可以是右上象限疼痛、食欲不振、疲劳、体重减轻等。酒精性脂肪肝的临床表现与肝脏脂肪浸润程度成正比，肝脏脂肪过多清除后症状消失。临床上，肝肿大是最常见的症状，其次是肝痛和触痛..."
}

Table 21: Representative examples from the RAG corpus (`huatuo_encyclopedia_qa`) used for evidence retrieval.