AIDA-SEAT: Towards Reliable AI Doctor Assistant via State-Evaluation-Action Tree Enhanced LLMs in Online Hospital

Lianxin Sun[♢*], **Xiaoying Ying**[♠*], **Guangya Yu**[♢], **Weiyan Zhang**[♢†], **Chenhao Guan**[♢], **Hao He**[♢], **Mingxi Shang**[♡], **Jianhua Li**[♢†], **Chunming Wang**[♠], **Tong Ruan**[♢]
[♢]East China University of Science and Technology, Shanghai, China,
[♠]Renji Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China,
[♡]HealthCloud (Shanghai) Digital Technology Co., Ltd, Shanghai, China
Correspondence: jhli@ecust.edu.cn, weiyanzhang@ecust.edu.cn

Abstract

Artificial intelligence doctor assistants (AIDAs) help streamline clinical decision-making and reduce physician workload. While existing systems primarily utilize Large Language Models (LLMs) or retrieval-augmented generation (RAG), these methods typically retrieve static facts—whether as text passages or structured graphs—lacking the explicit logical pathways essential for multi-step reasoning. In this paper, we propose the AIDA-SEAT framework to provide reliable clinical decision-making support. First, we design the state-evaluation-action tree (SEAT), which covers diagnosis, treatment, and examination. To develop this tree, we refine and transform SEAT collected from medical documents and doctors. Then, we propose an adaptive method to select optimal trees tailored to the current patients' state. Finally, we leverage LLMs to perform state assessment, evaluation, and action execution based on the tree, thereby generating reliable responses. To evaluate the effectiveness of our method, we conducted extensive experiments on a self-built dataset. Our method achieves 1.01% higher than current state-of-the-art (SOTA) baselines across five departments, including common RAG-based methods. Furthermore, analysis of 200 consultation records during deployment on an online hospital revealed that system-assisted responses are 24.16 seconds faster on average than manual ones, improving efficiency by 26.85%.

1 Introduction

Artificial intelligence doctor assistants (AIDAs) are essential in medical dialogue systems, designed to communicate with patients to gather information and make decisions (Liao et al., 2022; Zhong et al., 2022). To build a reliable and trustworthy medical dialogue system, it is crucial to integrate the clinical logic into the AIDA (Shi et al., 2024).

Recently, LLMs like GPT-4 (Achiam et al., 2023) have shown promising capabilities for healthcare, especially advanced medical dialogue systems (Fan et al., 2025). To curate an effective and reliable AIDA for the medical dialogue system, the LLM-based method can be divided into three parts: fine-tuning-, prompt- and RAG- based methods. The fine-tuning approach involves collecting and constructing high-quality medical corpora and datasets, followed by adapting LLMs to the medical domain through continued pre-training (Yang et al., 2024b), instruction tuning (Chen et al., 2024), or reinforcement learning (Dou et al., 2024). While these methods enhance medical knowledge in LLMs, they also introduce two significant drawbacks: catastrophic forgetting (Liu et al., 2024) and high computational cost (Chen et al., 2024). Therefore, prompt-based methods have gained attention by combining the chain-of-thought (CoT) (Wei et al., 2022) and human expert experience (Singhal et al., 2023; Nori et al., 2023). However, these approaches require manual integration of clinical logic. Conversely, RAG dynamically retrieves external medical knowledge to ground LLM responses. While traditional RAG retrieves unstructured text passages (Xiong et al., 2024a,b), recent advancements like Graph-based RAG (Edge et al., 2024; Wu et al., 2025) leverage structured knowledge graphs. Although these approaches effectively mitigate hallucinations (Wang et al., 2025a), they primarily supply static medical facts. Whether retrieving textual snippets or structural graph relations, they fail to explicitly model the step-by-step clinical logic critical for clinician-like diagnostic reasoning. Furthermore, retrieving noisy or irrelevant content can compromise both response quality and efficiency (Arslan et al., 2024).

Therefore, leveraging the Clinical Guidance Tree (CGT) to enable LLM reasoning with clinical logic has shown promising performance (Li et al., 2023). Contrasting with the traditional, regimented frame-

[*] Equal Contribution.
[†] Corresponding Authors.

works of clinical guidance and decision trees (Zhu et al., 2022), this method adopts natural language for a nuanced representation of node content, and executes the decision system with LLM. Building on this idea, we propose the State-Evaluation-Action Tree, which simulates the clinical reasoning of human experts: based on the patient's condition, experts **evaluate** the **state** and determine the appropriate **action**.

In this paper, we introduce the SEAT and explore SEAT-guided LLM generation for an advanced AIDA system. The SEAT integrates root, condition, transition, and action nodes along with their interconnecting relations. Each node is annotated with a natural language description, while each relation carries multiple natural language labels pointing to the next node. To construct SEATs, we first extract raw clinical guidance trees using the tool of Li et al. (2023). We then employ an LLM to decompose the natural language content of each node into structured logical facts, stored in JSON format for easier reasoning. Finally, clinicians refine and customize the SEATs based on their domain expertise. When a patient submits their current information to the AIDA system, it triggers SEAT-guided LLM generation through a three-step process: tree selection, condition node judgment, and response generation. (1) Tree Selection: An LLM evaluates each fact against the patient's current state and selects the optimal tree using a heuristic rule. (2) Condition Node Judgment: The LLM assesses the relations associated with the current condition node, assigns judgments, and identifies the most suitable next node. (3) LLM Generation: The AIDA system adaptively chooses between a standard CoT prompt and a SEAT-guided prompt to generate the final response. The online hospital physician reviews the response and then sent to the patient.

The contributions are summarized as follows:

- We design the state-evaluation-action tree (SEAT) and introduce an SEAT-guided LLM-based AIDA system.
- Experiments on a self-built dataset across five departments demonstrate that our approach outperforms CoT and RAG-based methods, achieving a 1.01% improvement over SOTA baselines.
- Deployment in a real-world online hospital confirms that our method enhances both the efficiency and reliability of the AIDA system.

2 Related Work

Recent studies highlight RAG-based methods that retrieve relevant passages from external knowledge sources—such as clinical guidelines and scientific literature—and inject them into prompts to provide contextual grounding, steering the LLM toward more accurate responses. For example, MedRAG (Xiong et al., 2024a) constructs a large, authoritative medical corpus from PubMed, Wikipedia, and other sources. To enhance performance, i-MedRAG (Xiong et al., 2024b) introduces iterative retrieval, enabling the LLM to dynamically issue follow-up queries based on initial results, significantly improving answer quality for complex clinical questions. Meanwhile, MedCite (Wang et al., 2025b) enhances verifiability by generating responses with explicit citations. To leverage structural connections across knowledge bases, GraphRAG (Edge et al., 2024) further enhances complex reasoning by constructing knowledge graphs, while MedicalGraphRAG (Wu et al., 2025) explores this within the medical domain. However, while traditional RAG approaches rely on unstructured text snippets and graph-based methods utilize structured relational data, both still fall short in capturing the explicit decision-making pathways required for multi-step diagnostic reasoning. To address this limitation, we introduce SEAT, which integrates static medical facts with explicit clinical logic to guide the LLM toward more reliable and coherent reasoning and generation.

3 Methodology

As illustrated in Figure 1, our method comprises two main modules: (1) construction of the SEAT, and (2) SEAT-guided LLM generation.

3.1 Problem Definition

Given the patient's information P_I, which consists of self-report, medical records, and current utterance, the LLM generates a response R using a carefully designed prompt P to assist the doctors. This process can be formally expressed as: $Response = LLM(Prompt, P_I)$

3.2 SEAT Construction

The State-Evaluation-Action Tree (SEAT) is a structured decision framework composed of four distinct node types that facilitate clinical or logical reasoning. The **Root Node** serves as the foundation of the tree, representing the core symptom,

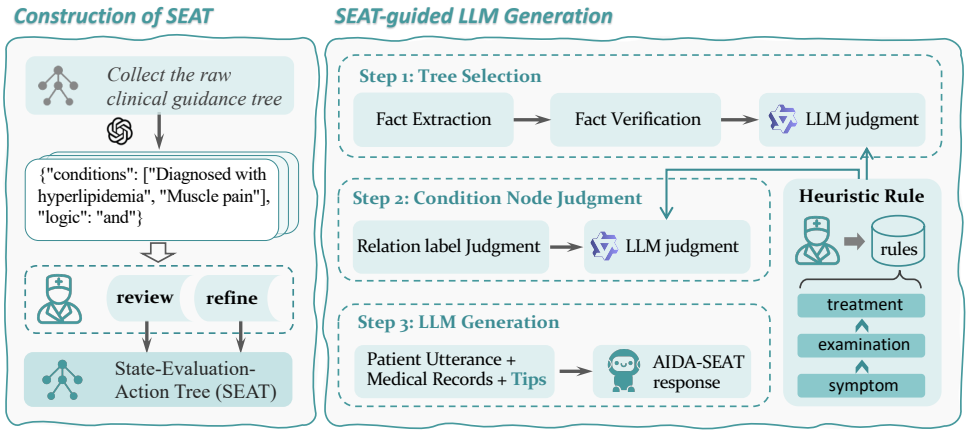

Figure 1: Framework diagram of the AIDA-SEAT

primary disease name, or the initial examination that triggers the diagnostic process. Branching from the root, the **Condition Node** acts as a logical gatekeeper, determining the subsequent path in the decision process based on specific criteria or clinical findings. In cases where the decision flow requires external logic, the **Transition Node** is utilized to redirect the process to another specialized tree. Finally, the **Action Node** marks the termination of the decision path and specifies the ultimate action taken or the final diagnostic conclusion reached. These nodes are interconnected via a network of relations, in which each relation is annotated with multiple labels that specify the precise links between different stages of the tree.

To ensure the effectiveness and authority of the tree, we adopt the method from Li et al. (2023) to collect the raw clinical guidance tree and then use an LLM to convert its natural language into a structured format. For example, the statement "Has hyperlipidemia and experiences muscle pain" is parsed into: {"conditions": ["Diagnosed with hyperlipidemia", "Muscle pain"], "logic": "and"}. Each resulting SEAT is saved as a JSON file for easy configuration. Human experts further review and refine the SEATs.

3.3 SEAT-guided LLM Generation

This section details the SEAT-guided LLM generation framework, which integrates structured clinical logic with large language models through three primary phases: Tree Selection, Condition Node Judgment, and LLM Generation.

Tree Selection. To initiate the process, the system must identify the most relevant decision tree based on the patient information. This selection strategy is designed to balance computational efficiency with clinical accuracy. First, the system performs Fact Extraction and Verification by gathering and deduplicating medical facts from all available root nodes; an LLM then labels each fact as either True or False. Based on these results, the process follows a specific branching logic: if no facts align with any existing tree, the current turn terminates and reverts to a standard prompt; if exactly one tree matches, the decision process continues along that specific path. In cases where multiple trees match, a hybrid approach is employed, combining heuristic rules with LLM judgment centered on patient severity. This heuristic prioritizes trees in the hierarchical order of Treatment, followed by Examination, and finally Symptom. Should multiple trees of the same type persist, the LLM selects the most pertinent one by evaluating the patient's dominant symptom. **Condition Node Judgment.** Once a tree is selected, the LLM evaluates each relation label associated with the current condition node relative to. The model assigns one of three categorical judgments to each label: True, False, or Not Sure. The

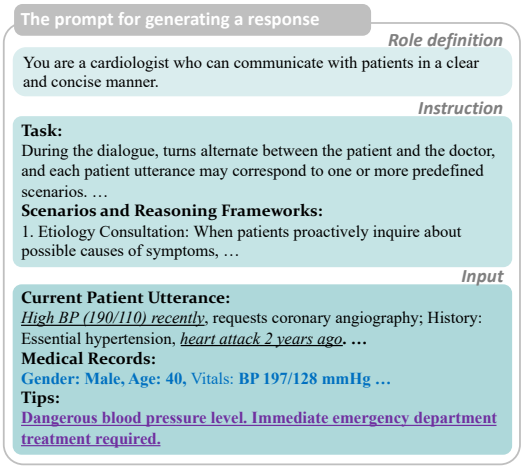

Figure 2: Prompt template for LLM generation

Department	#Test	#Symp.	#Exam.	#Treat.
Cardiology	200	3	4	4
Liver Surgery	200	3	12	1
Gastroenterology	200	14	13	/
General Medicine	200	24	49	4
Reprod. Med.	43	5	8	/

Table 1: Statistics of test datasets. Symp. denotes symptom tree, Exam. denotes examination, Treat. denotes treatment. Reprod and Med. denotes Reproductive Medicine.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.6572	4.7393	4.517	4.2547	4.5421
i-MedRAG	4.6303	4.6150	4.3390	3.9872	4.3929
MedCite	**4.7927**	4.8181	4.4881	4.2300	4.5822
MedRAG	4.7806	4.8182	4.4726	4.1671	4.5596
Ours	4.7922	**4.8399**	**4.5417**	**4.3407**	**4.6286**

Table 2: Overall results across five departments. The abbreviations stand for Med. (Medical Accuracy), Cli. (Clinical Logic), Com. (Communication), and Emp. (Empathy). Avg. S. denotes the average score. **Bold** denotes the best performance. Underline denotes the second performance.

subsequent navigation depends on this assessment: if multiple labels are judged as True, the system applies the same prioritization strategy used during the initial tree selection. If only a single relation label is True, the decision flow proceeds through the tree until it reaches either a transition node leading to another tree or an action node that concludes the current turn, at which point the system proceeds with a SEAT-guided prompt. Conversely, if labels are deemed False or Not Sure, a specific flag is set to True, the current turn terminates, and the system defaults to a standard prompt. This flag mechanism is crucial for maintaining state, as it prevents the redundant re-evaluation of the same condition node in subsequent interaction turns.

LLM Generation. The system adaptively selects between two prompt templates: (1) a **standard CoT prompt**, handcrafted to include *Role Definition*, *Instruction*, and *Input*; (2) an **SEAT-guided prompt**, which augments the standard template by incorporating action node information into the *Tip*. The details are shown in Figure 2.

4 Experiments

We conduct extensive experiments to validate the effectiveness of our proposed method.

4.1 Experimental Setup

Datasets. We construct and employ the following datasets: (1) **Test Sets**: 843 single-turn samples with three types of SEATs from anonymized patient-doctor dialogues across five departments. (2) **Tree Selection Dataset**: 70 cardiology cases annotated with optimal SEATs. (3) **Condition Node Judgment Dataset**: 98 Cardiology questions derived from SEAT nodes, labeled True, False, or Not Sure. (4) **Human Evaluation**: 72 anonymized response pairs assessed by a cardiologist across five dimensions. See Table 1 for detailed statistics.

Baselines and Models. We compare our proposed method with the following baselines: **Chain-of-Thought** (Wei et al., 2022), **MedRAG** (Xiong et al., 2024a), **i-MedRAG** (Xiong et al., 2024b), **MedCite** (Wang et al., 2025b). We choose different representative LLMs such as Qwen series (Yang et al., 2025), Llama3.1-8b-instruct (Grattafiori et al., 2024) and DeepSeek-R1 (Guo et al., 2025) for comprehensive experiments. To balance cost and performance, we chose the **Qwen3-8B** for tree selection and Condition node judgment tasks. Using **Qwen3-8B** as the backbone for generation, we compare our method against baselines. More implementation details in Appendix A and B.

Evaluation Metrics. We employ three distinct sets of metrics tailored to our different experimental objectives. **Model Selection and Condition Node Judgment.** For model selection, we use accuracy. For condition node judgment, we adopt a fine-grained scoring scheme: correct = 1, incorrect = 0, and "uncertain" on a determinable case = 0.5. The final score is the total points earned divided by the maximum possible, scaled to 100. **LLM as Judge for Medical Dialogue System.** We adopt the LLM-as-Judge paradigm and define four clinically grounded dimensions, each scored on a 0–5 scale: **Medical Accuracy**, **Clinical Logic**, **Communication**, **Empathy**. The final evaluation score is the average across all four dimensions. To mitigate model-specific bias (Ye et al., 2025), our main experiments aggregate judgments from three LLMs—DeepSeek-V3.1, Qwen-Max, and Doubao-Seed-1.6-Thinking [1]—while ablation studies use only DeepSeek-V3.1. **Human Evaluation.** We follow the work (Zhu et al., 2025) and evaluate the response in five dimensions ranging from 0-5: **Medical Accuracy**, **Reliability Score**, **Fostering the Relationship**, **Gathering Information**, **Providing Information**. The final score is the average rating across dimensions. More details in Appendix C.

[1] https://console.volcengine.com/

Method	DeepSeek-V3.1				Qwen3-Max				Doubao-Seed-1.6-Thinking				Avg. Score
	Med.	Cli.	Com.	Emp.	Med.	Cli.	Com.	Emp.	Med.	Cli.	Com.	Emp.	
CoT	4.6982	4.7738	4.5089	4.2113	4.7798	4.7500	4.8333	4.7679	3.7054	3.9583	3.9107	3.7887	4.3905
i-MedRAG	4.8333	4.8988	4.2560	3.8482	**4.8810**	4.8214	4.5119	4.3690	**4.3512**	4.3661	3.8750	3.4673	4.3733
MedCite	4.8006	4.8214	4.4762	4.2172	4.8333	4.7976	4.8631	4.7560	4.0357	4.1964	4.1429	3.9554	4.4913
MedRAG	4.8179	4.8262	4.5054	4.2119	4.8333	4.7619	4.8155	4.7143	4.1220	4.3065	4.1696	3.8512	4.4947
Ours	**4.8691**	**4.9345**	**4.6191**	**4.3691**	**4.8810**	**4.8750**	**4.8690**	**4.8274**	4.2440	**4.5119**	**4.4256**	**4.1935**	**4.6349**

Table 3: Main results of Qwen3-8B on test set of Cardiology.

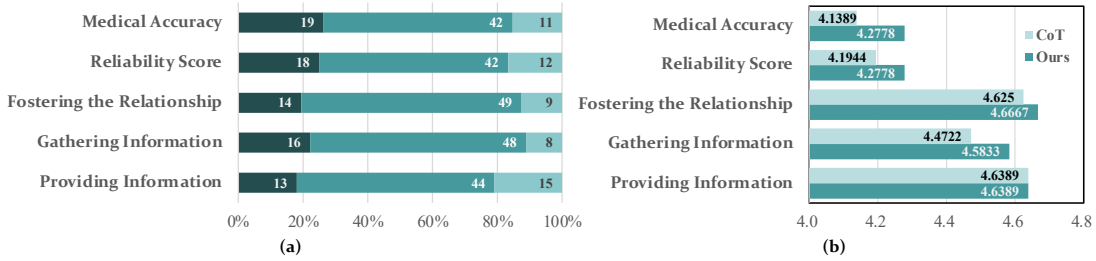

Figure 3: Human evaluation, comparing our SEAT-guided method with the CoT baseline: (a) Pairwise comparison results (win/tie/lose) across all dimensions. (b) Average score comparison across the five dimensions.

Stage1	Stage2	Med.	Cli.	Com.	Emp.
Effectiveness of different models in stage 1					
/	Qwen3-8B	4.6982	4.7738	4.5089	4.2113
Qwen3-0.6B	Qwen3-8B	4.8185	4.8566	4.5714	4.2441
Qwen3-4B	Qwen3-8B	4.7738	4.8869	**4.6577**	4.3630
Qwen3-8B	Qwen3-8B	**4.8691**	**4.9137**	4.6191	**4.3691**
Impact of different models in stage 2					
/	Qwen3-A22B	4.8363	4.8839	**4.9524**	**4.8506**
Qwen3-8B	Qwen3-A22B	4.8537	4.9065	4.8618	4.6911
/	DeepSeek-R1	4.7597	4.8178	4.8527	4.3992
Qwen3-8B	DeepSeek-R1	**4.8622**	**4.9370**	4.7992	4.6063

Table 4: Ablation study on different models at two stages. Qwen3-A22B denotes the Qwen3-235B-A22B.

Model	Tree Sel. (Acc.%)	Node Judg. (Score)
Qwen2.5-7B-Inst.	75.71	71.81
Qwen2.5-3B-Inst.	61.43	67.64
Qwen3-8B	**82.86**	**79.48**
Qwen3-4B	72.86	78.25
Qwen3-0.6B	41.43	53.62
Llama-3.1-8B-Inst.	71.43	73.45

Table 5: Ablation study on model selection of tree selection and Condition node judgment.

4.2 Main Results

Table 2 compares our method with baselines across five departments. Key observations: (1) Using Qwen3-8B, our approach outperforms CoT and RAG-based methods on all four metrics, surpassing MedCite by 1.01% and achieving the highest **Clinical Logic** scores(4.8399). (2) RAG improves **Medical Accuracy** and **Clinical Logic** but often reduces **Communication** and **Empathy**; our method avoids such trade-offs, demonstrating robustness. (3) As shown in Table 3, we evaluate on cardiology data with three judge models. Our method consistently achieves the best performance. DeepSeek-V3.1 offers more discriminative scoring than Qwen3-Max (uniformly high) or Doubao (inconsistent). More results are in Appendix D.

4.3 Detailed Analysis

In this section, we conduct extensive experiments to analyze our proposed method in detail.

Ablation Study. We ablate key components in Table 4. Qwen3-8B is selected as the optimal Stage1 model for tree selection and condition judgment. Fixing Qwen3-8B as the generator: (1) **Stage1**—even Qwen3-0.6B outperforms CoT; Qwen3-8B yields the best results, consistent with Table 5. (2) **Stage2**—SEAT-guided generation improves **Medical Accuracy** and **Clinical Logic** for DeepSeek-R1 and Qwen3-235B-A22B over CoT, occasionally at the expense of **Communication** and **Empathy**. Notably, SEAT enables Qwen3-8B to rival much larger models.

Human Evaluation. Our method outperforms the baseline in four dimensions (Figure 3), particularly in **Medical Accuracy** (19 vs. 11 wins; avg. 4.28 vs. 4.14), **Gathering Information** (16 vs. 8 wins), and **Reliability Score** (4.28 vs. 4.19). Performance is comparable in **Providing Information** (44 ties; avg. 4.64). These results demonstrate our method's superiority in producing clinically reliable responses.

Input	40 years old male; Complaint: **High BP (190/110) recently**, requests coronary angiography; History: Essential hypertension, **heart attack 2 years ago**; Vitals: **BP 197/128 mmHg**; Medication: Amlodipine/Benazepril, Hydrochlorothiazide.
Reference	This level of high blood pressure carries a real risk of an acute heart problem. **You must go to the Emergency Department immediately** to have it stabilized before we can proceed with the angiogram. Also, please understand that this procedure requires hospitalization, so you will need to be admitted.
CoT	... before scheduling this procedure, I suggest you **first make an appointment with the cardiology outpatient department** for a comprehensive evaluation.
MedRAG	... it is recommended that you make an appointment with the cardiology clinic... to **adjust your antihypertensive medication plan and bring your blood pressure into a safer range** before the procedure.
Ours-with incorrect judgment	... before proceeding, I suggest you **first make an appointment with the cardiology outpatient clinic** for a specialist to assess whether this procedure is necessary.
Ours	... this indicates a serious risk that must be addressed immediately. **I recommend you go to the Emergency Department immediately** for urgent evaluation and treatment... your blood pressure **must first be stabilized through emergency treatment** before re-evaluating.

Figure 4: Case Study

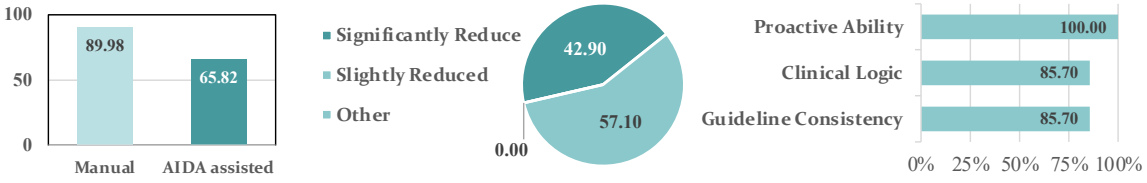

Figure 5: Online evaluation of the AIDA system. (a) Manual vs. AIDA-assisted: average response time. (b) Perceived reduction in physicians' cognitive load. (c) Approval rates for core clinical functionalities.

Figure 6: The user interface of the AIDA's web client.

Case Study. We validate how different knowledge affect LLM outputs (Figure 4). For a patient with severe hypertension (190/110 mmHg), both CoT and RAG suggest BP control and cardiology referral. Our method instead recognizes clinical urgency and recommends immediate ED admission. While baselines reflect general knowledge, our approach yields condition-specific, expertise-driven advice. Introducing an incorrect action node (claiming BP in the ideal range) misled the LLM to omit emergency referral, revealing the critical role of precise tree selection and condition judgment.

5 Application on Online Hospital

Application Deployment. The system has been operational in the hospital since May 21, 2025. As of January 4, 2026, 85 physicians have trialed AIDA, resulting in over 4,700 model invocations and the processing of more than 5,900 images by the multimodal model. Furthermore, the Chief Physician

of the Department of Cardiology handled an average of 234 follow-up consultations per month from January to July 2025 (excluding patients seeking rapid prescription renewals). From August to December 2025, this monthly average increased to 303, representing an increase of approximately 29.48% compared to the previous period. A complete timeline of our phased deployment is outlined in Appendix F. **User Interface.** The application provides responsive web and mobile interfaces for seamless user access across devices. As shown in Figure 6-7, based on the reasoning and responses it generates, doctors can perform the following actions: revise, edit, or send. **Online Evaluations.** Through seamless integration of LLM technology, AIDA significantly enhances physicians' decision-making efficiency. As indicated in Figure 5, AIDA-assisted responses reduced time by 24.16 seconds compared to manual responses on 200 medical consultations. Survey results revealed that (1) all doctors acknowledged AIDA's time-saving benefits in decision-making, (2) all participants affirmed its proactive ability, and (3) 85.7% valued its clinical logic and guideline consistency. Additionally, we analyzed the system's end-to-end inference time, which averages 6.43 seconds. Stage 1 (retrieval) takes 1.30 seconds, Stage 2 (LLM generation) takes 5.12 seconds, and the remaining 0.01 seconds is attributed to inter-component communication and preprocessing overhead. **Application Use and Payoff.** We highlight the practical impact and pivotal role of the AIDA system. AIDA introduces an innovative, effective approach to clinical decision-making, substantially advancing beyond traditional workflows that depend solely on human expertise. By leveraging SEATs, physicians can tailor their own AIDA instances to reflect their clinical reasoning, enabling more natural, logic-driven dialogues. This customization fosters a progressively reliable and trustworthy AI-powered medical dialogue system. More details in Appendix E.

6 Conclusion

In this paper, we construct SEAT and introduce the SEAT-guided LLM generation method to enhance the AIDA system. Experiments demonstrate that our method outperforms CoT and RAG-based methods across five departments from a real-world online hospital. Online evaluations further demonstrate that SEAT-guided LLM improves clinical decision-making efficiency, while its reliability significantly reshapes physicians' cognitive processes.

Limitations

The AIDA-SEAT framework has several limitations. (1) Although an automatic tool was introduced to construct the raw SEAT, its outputs still require review by clinical experts. (2) This study only compares CoT and RAG-based methods; the impact of fine-tuning-based approaches will be explored in future work. (3) While a human evaluation was conducted, the number of experts and departments involved was limited. Future efforts will aim to scale up this evaluation.

Ethical Consideration

We collect the medical dialogue from a real-world online hospital. To safeguard patient privacy, the 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 the dataset. Furthermore, all data usage adheres to ethical guidelines and regulations governing medical information and research.

Acknowledgments

This work is supported by the Shanghai Natural Science Foundation Project under Grant 25ZR1402116.

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*.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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*.

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.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Kangenbei Liao, CHENG ZHONG, Wei Chen, Qianlong Liu, zhongyu wei, Baolin Peng, and Xuanjing Huang. 2022. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning.

Chengyuan Liu, Yangyang Kang, Shihang Wang, Lizhi Qing, Fubang Zhao, Chao Wu, Changlong Sun, Kun Kuang, and Fei Wu. 2024. More than catastrophic forgetting: Integrating general capabilities for domain-specific llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7531–7548.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861.

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.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Qipeng Guo, Xiangkun Hu, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Xuming Hu, and 1 others. 2025a. Survey on factuality in large language models. *ACM Computing Surveys*, 58(1):1–37.

Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, and Minjia Zhang. 2025b. Medcite: Can language models generate verifiable text for medicine? *arXiv preprint arXiv:2506.06605*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

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*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others.

2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Xiechi Zhang, Zetian Ouyang, Linlin Wang, Gerard De Melo, Zhu Cao, Xiaoling Wang, Ya Zhang, Yanfeng Wang, and Liang He. 2025. AutoMedEval: Harnessing language models for automatic medical capability evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6272–6285, Vienna, Austria. Association for Computational Linguistics.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2857.

Wei Zhu, Wenfeng Li, Xiaoling Wang, Wendi Ji, Yuanbin Wu, Jin Chen, Liang Chen, and Buzhou Tang. 2022. Extracting decision trees from medical texts: an overview of the text2dt track in chip2022. In *China Health Information Processing Conference*, pages 89–102. Springer.

A Baseline and Models

Baselines. We compare our proposed method with the following baselines:

- **Chain-of-Thought** (Wei et al., 2022): We handcraft the prompt with human expert guided. The prompt is set in all experiments.
- **MedRAG** (Xiong et al., 2024a): Retrieving the medical fact from a large and authoritative medical corpus to enhance the LLM generation.
- **i-MedRAG** (Xiong et al., 2024b): Through multi-round decomposition and iteration to improve the performance.
- **MedCite** (Wang et al., 2025b): An end-to-end system featuring a novel multi-pass retrieval-citation mechanism that enables LLMs to generate verifiable text.

Models. We choose different representative LLMs such as Qwen2.5-{3b/7b}-instruct (Yang et al., 2024a), Qwen3-{0.6b/4b/8b} series (Yang et al., 2025), Llama3.1-8b-instruct (Grattafiori et al., 2024), DeepSeek-R1[2], and Qwen3-235B-A22B-Thinking-2507 [2] for comprehensive experiments.

B Implementation Details

(1) We conduct all experiments on A800 and use VLLM [3] to accelerate for general LLMs. Specifically, we load the medical LLMs directly. Additionally, we set the max_new_tokens = 8192; repetition_penalty = 1.2; temperature = 0.001. (2) We use the MedCPT (Jin et al., 2023) as the retriever for the baseline method. (3) The medical corpus is compiled from three main sources: (i) PubMedQA (508 MB); (ii) curated expert-consensus documents (203 files, 563 MB); and (iii) a broader collection of clinical guidelines and medical books (5,437 files, 2.4 GB). To align with SEAT, we specifically curate literature and clinical guidelines from these sources. The selected documents are then stored in the Faiss database.

C Evaluation Details

We employ three distinct sets of metrics tailored to our different experimental objectives.

[2]https://bailian.console.aliyun.com/

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

(1) **Model Selection of tree selection and Condition node judgment**. For tree selection, we calculate the metric as accuracy; For Condition node judgment, we employ a refined scoring scheme to better reflect model capability: a correct judgment earns 1 point, an incorrect one 0 points, and an "uncertain" response on a case that admits a definite answer receives 0.5 points. This partial credit recognizes that a cautious "uncertain" prediction is more informative than a confident but incorrect guess. The final score is the total points earned divided by the maximum possible score, as given by the formula:

$$\text{Score} = \frac{\sum_{i=1}^{N} s_i}{N} \times 100 \tag{1}$$

where s_i is the score for the i-th sample and N is the total number of samples.

(2) **LLM as Judge for Medical Dialogue System.** Traditional automatic metrics such as F1 and ROUGE rely heavily on lexical overlap and often overlook semantic nuances. Even embedding-based metrics like BERTScore, despite leveraging pre-trained language models, exhibit weak correlation with human judgments (Zhang et al., 2025). To address this, we adopt the LLM-as-Judge paradigm and define four clinically grounded dimensions, each scored on a 0–5 scale: 1) **Medical Accuracy:** factual correctness of the medical content. 2) **Clinical Logic:** coherence of the response within a clinical reasoning context. 3) **Communication:** clarity and comprehensibility of the response. 4) **Empathy:** expression of care and compassion. The final evaluation score is the average across all four dimensions.

(3) **Detail Criteria for Human Evaluation.** 1) **Medical Accuracy:** Assesses the factual correctness of the diagnosis and other medical information provided by the model. 2) **Reliability Score:** Evaluates whether the proposed diagnosis, treatment plans, or recommended tests are consistent with established medical knowledge and guidelines. 3) **Fostering the Relationship:** Measures the model's ability to build trust, rapport, and a positive connection with the patient. 4) **Gathering Information:** Assesses the model's effectiveness in eliciting relevant and necessary information from the patient. 5) **Providing Information:** Gauges the model's proficiency in delivering information that is clear, understandable, and accurate to the patient. The final reported score is the average rating for each dimension.

D More Experiment Results

As shown in Tables 6–10, our method achieves the best performance across four of five departments (Cardiology, Gastroenterology, General Medicine, and Liver Surgery), with the sole exception being Reproductive Medicine—due to the quality of the corresponding SEAT and supporting documents.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.6982	4.7738	4.5089	4.2113	4.5480
i-MedRAG	4.8333	4.8988	4.2560	3.8482	4.4591
MedCite	4.7935	4.8226	4.4940	4.2173	4.5818
MedRAG	4.8179	4.8262	4.5054	4.2119	4.5903
Ours	**4.8691**	**4.9345**	**4.6191**	**4.3691**	**4.6979**

Table 6: Overall results of Cardiology.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.6212	4.8485	4.5000	**4.4394**	4.6023
i-MedRAG	4.5606	4.5455	4.3030	4.0758	4.3712
MedCite	4.8409	4.8788	4.4318	4.3788	4.6326
MedRAG	**4.8485**	**4.8939**	**4.5909**	4.3030	4.6591
Ours	4.8030	**4.8939**	4.5455	4.4091	**4.6629**

Table 7: Overall results of Liver Surgery.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.6484	4.6813	**4.4615**	4.0604	4.4629
i-MedRAG	4.5275	4.4835	4.3077	3.8407	4.2898
MedCite	**4.7830**	**4.8077**	**4.4615**	4.0577	4.5275
MedRAG	4.7187	4.7830	4.3379	3.9093	4.4372
Ours	4.7363	4.7912	4.4451	**4.1484**	**4.5302**

Table 8: Overall results of Gastroenterology.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.6978	4.7698	**4.6835**	4.5252	4.6691
i-MedRAG	4.6403	4.5827	4.5755	4.4029	4.5503
MedCite	**4.8417**	**4.8489**	4.5827	4.4317	4.6762
MedRAG	4.8129	4.8201	4.6187	4.4604	4.6780
Ours	4.8345	4.8273	4.6475	**4.6043**	**4.7284**

Table 9: Overall results of General Medicine.

	Med.	Cli.	Com.	Emp.	Avg. S.
CoT	4.3906	4.5312	4.1875	4.0312	4.2851
i-MedRAG	4.2500	4.1562	4.0000	3.5625	3.9922
MedCite	4.5312	4.5938	**4.3125**	**4.0938**	**4.3828**
MedRAG	**4.6562**	**4.8125**	4.1875	3.8438	4.3750
Ours	4.5000	4.5625	4.2188	4.0000	4.3203

Table 10: Overall results of Reproductive Medicine.

E Application on Online Hospital

In this section, we introduce the deployment of the AIDA in the Online Hospital and report the application use in detail.

Application Deployment The system has been operational in the hospital since May 21, 2025. A complete timeline of our phased deployment is outlined in Appendix F. In the following, we introduce the deployment into three parts: computational resources, software development, and user interface.

Computational Resources. The system runs on a single server equipped with eight NVIDIA A100 40GB GPUs. We deploy three models tailored to different tasks: **Qwen2.5-VL-72B-AWQ** (Bai et al., 2025), used for extracting patient image information, is distributed across four GPUs to manage its high computational demand. **Qwen3-8B** handles tree selection, condition node judgment, and LLM generation, and runs on the remaining four GPUs.

Software Development. The system adopts a containerized architecture orchestrated by Docker to ensure reproducibility and simplify environment management. Core application logic—handling user interactions and coordinating model inference—is implemented in a custom Python 3.10 container powered by a FastAPI server.

Figure 7: The user interface of the AIDA's mobile client.

F Clinical Deployment Timeline

The clinical deployment of the system adhered to a structured, phased strategy designed to collect

iterative feedback, ensure stability, and validate its practical utility across multiple clinical specialties. The implementation timeline proceeded as follows:

Phase 1: **Initial Pilot** (May 21, 2025): Following a phase of rigorous technical validation, we initiated a pilot study with a single Chief Physician in the Department of Cardiology. This phase focused on iterative usability testing and collecting preliminary expert feedback to guide system refinement.

Phase 2: **Limited Production and Specialty Expansion** (July 29, 2025): After successful initial refinements, the system was integrated into the production workflow of the initial physician. Simultaneously, we expanded pilot testing to the Department of Gastroenterology and the Department of Liver Surgery, engaging one Chief Physician from each to provide preliminary assessments.

Phase 3: **Scaled User Adoption in Cardiology** (August 19, 2025): We expanded the user base within the Department of Cardiology to seven clinicians, including three Chief Physicians, three intermediate-level physicians, and one attending physician. This phase was designed to evaluate the system's performance and utility across a spectrum of clinical seniority.

Phase 4: **Full Departmental Rollout and Broadened Specialty Trials** (September 2025): On September 10, the system was deployed to all physicians in the Department of Cardiology. Subsequently, on September 21, the pilot was further extended to include the Department of Rheumatology, Department of Urology, Department of Bone and Joint Surgery, Department of Reproductive Medicine, Department of Pain Management, and General Medicine Department, with one Chief Physician from each department participating in initial trials and feedback collection, before gradually expanding the trial to more physicians.

This meticulous, phased approach enabled continuous validation of the system's real-world performance through clinical feedback, ensuring its alignment with medical workflows and supporting iterative refinement.