Hierarchical Agent Reflection for Aligning LLM Reasoning with Clinical Diagnostic Processes

Anonymous ACL submission

Abstract

Medical diagnosis is a complex, iterative process that relies heavily on clinicians' reasoning and judgment. Traditional models, while able to provide consistent diagnostic results, fail to replicate the reasoning process of clinicians, making their outputs difficult to understand and justify. In this paper, we address this limitation by first generating clinical notes that capture the clinician's diagnostic reasoning. These notes are then used to train a large language model, allowing it to mimic the step-by-step reasoning employed by clinicians during diagnosis. Our method introduces a hierarchical agent reflection mechanism to generate clinical notes, which deconstructs the diagnostic process into key stages, each handled by specialized agents. This structured approach not only improves the accuracy and reliability of the generated clinical notes but also ensures that the model's reasoning aligns with human clinical practice. Experimental results show that models trained on this data outperform both general-purpose large language models and domain-specific medical models in diagnostic tasks. The proposed method enhances diagnostic transparency and interpretability, offering a valuable tool for AI-assisted clinical decision-making.

1 Introduction

Timely and accurate diagnosis is foundational to good clinical practice and an essential first step to achieving optimal patient outcomes (Singh et al., 2019). With the continuous advancement of modern medical technologies, particularly the rise of artificial intelligence and large language models (LLMs), the medical diagnostic process is undergoing transformative change (Zhang et al., 2025; Buess et al., 2025). Research is increasingly focused on leveraging these technologies to assist clinicians in achieving more accurate and efficient diagnoses (McDuff et al., 2025; Maleki Varnosfaderani and Forouzanfar, 2024). Recent studies

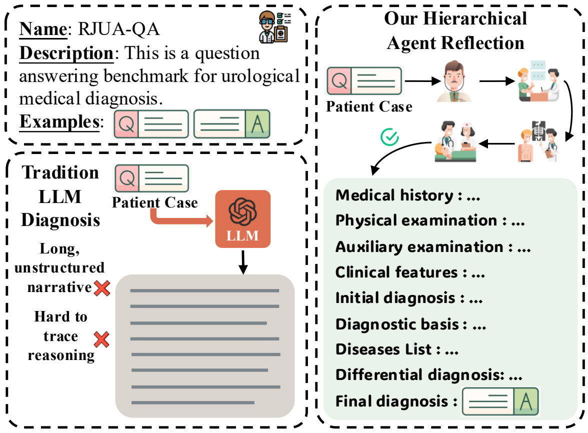

Figure 1: Previous models often produce diagnostic outputs as long, unstructured narratives, making it difficult to trace their reasoning process. In contrast, our method first generates clinical notes that document the clinician's reasoning process. These notes are then used to train the model, enabling it to reason in a manner similar to clinicians.

have demonstrated that LLMs, when operating autonomously, can outperform clinicians in certain diagnostic tasks, underscoring the significant potential of these models (Goh et al., 2024; Brodeur et al., 2024).

However, current LLM-based diagnostic systems primarily offer static responses to clinician inquiries, lacking active engagement in the clinical reasoning process (Goh et al., 2024; Almansoori et al., 2025). This limitation restricts their effectiveness as collaborative tools in medical diagnosis, as they do not engage in the dynamic and iterative reasoning processes that clinicians rely on (Vally et al., 2023). Despite the rapid advancements of large models in the medical field and their high accuracy on static medical evaluation benchmarks, they have yet to demonstrate optimal performance in the domain of medical diagnosis (Kelly et al., 2019; Reese et al., 2024; Hager et al., 2024). In particular, diagnostic reasoning in medicine involves nuanced, non-linear decision-making based on a combination of clinical intuition, patient history, and test results (Ball et al., 2015; Vanstone et al., 2019; Stolper et al., 2021). To be truly effective in medical settings, LLMs must not only process vast amounts of data but also replicate the dynamic, step-by-step reasoning that clinicians employ during diagnosis (Kwon et al., 2024; Wu et al., 2025; Fan et al., 2025).

To bridge this gap and harness the full potential of LLMs, it is essential to align their diagnostic reasoning with clinical reasoning (Savage et al., 2024; Wang and Liu, 2025). This alignment can be achieved by fine-tuning the models using clinical notes, which encapsulate the detailed diagnostic processes of clinicians (Wang et al., 2024).

To do so, we propose a hierarchical agent reflection mechanism that integrates knowledge-enhancement techniques. We deconstruct the diagnostic process and design agents to simulate the multiple steps a clinician would take when diagnosing with clinical notes. The resulting clinical notes are then used for further training of the model, ensuring that the LLM's diagnostic reasoning resonates with that of clinicians (see Fig. 1). Our framework is designed with a hierarchy of specialized agents, consisting of three foundational agents and one supervisory agent: (1) **Information Collection Agent** – Extracts and summarizes relevant patient data. (2) **Preliminary Diagnosis Agent** – Conducts iterative reasoning to generates an preliminary diagnostic hypothesis. (3) **Differential Diagnosis Agent** – Conducts iterative reasoning to refine the differential diagnosis. (4) **Coordinator Agent** – as the supervisory agent, Oversees and integrates the reasoning outputs of other agents. Our contributions are as follows:

- Simulation of Clinician Reasoning: We introduce a pioneering approach to explicitly simulate clinicians' diagnostic reasoning trajectories using clinical notes, teaching the model the diagnostic thinking of doctors.
- Hierarchical Agent Reflection: We develop an innovative hierarchical agent reflection framework, which enhances clinical note generation through structured iterative refinement. This framework significantly improves the accuracy and reliability of the generated data.
- Empirical Validation: Our experimental results demonstrate that models trained on datasets generated by our method significantly outperform both general-purpose large language models and domain-specific medical models in diagnostic tasks. Ablation studies further confirm the effectiveness of the hierarchical agent reflection mechanism.
- Enhanced Diagnostic Transparency: The model produces diagnostic pathways that are both interpretable and traceable, effectively aligning with clinicians' reasoning processes. This transparency enhances trust in AI-assisted diagnostics, making it a reliable tool for clinical applications.

2 Related Works

2.1 Medical Large Language Models

In recent years, the application of large language models in the medical field has become a major research focus (Singhal et al., 2023; Thirunavukarasu et al., 2023; Han et al., 2023; Kim et al.; Saab et al., 2024; Truhn et al., 2024; Christophe et al., 2024; Zhou et al., 2023). These models enhance LLM capabilities in medicine through various approaches. For instance, models like BioMedLM (Bolton et al.), OphGLM (Gao et al., 2023), and GatorTronGPT (Peng et al., 2023) absorb extensive medical knowledge during pre-training, enabling strong performance across a range of medical tasks. Given the time and cost associated with developing specialized medical LLMs from scratch, models like Med-Gemini (Yang et al., 2024), Med42 (Christophe et al., 2024), MedAlpaca (Han et al., 2023), and MedPaLM-2 (Singhal et al., 2025) opt to build on robust general-purpose base models, fine-tuning them with different strategies to meet the specific needs of the medical domain and ultimately transforming them into specialized medical LLMs.

Furthermore, certain models have improved their medical capabilities by implementing preference alignment techniques. For example, HuatuoGPT-o1 (Chen et al., 2024b) significantly enhanced its medical reasoning abilities by utilizing verifiable medical reasoning datasets and reinforcement learning. MedFound aligned itself with standard clinical practices by introducing a unified preference alignment framework, while Baichuan-M1 improved its diagnostic capabilities through reinforcement learning and pairwise data optimization.

While fine-tuning reduces computational resources compared to pre-training, it still requires additional model training and high-quality datasets, which can be resource-intensive. In contrast, prompt engineering offers a more efficient method to adapt base models to specific use cases without altering model parameters. Techniques like few-shot learning, in-context learning, chain-of-thought prompting (Wei et al., 2022), and retrieval-augmented generation (RAG) (Lewis et al., 2020) are commonly used. Given the critical importance of accuracy in medical applications, RAG is particularly effective for providing models with reliable information. Models such as Oncology-GPT-4(Ferber et al., 2024), MedRAG (Xiong et al., 2024), and MedGraphRAG (Wu et al., a) enhance overall performance by incorporating external, trustworthy sources of information into the answer generation process.

Our approach combines the strengths of these techniques with the capabilities of general models. We find relying solely on LLMs for direct question-answering may not sufficiently meet the demands of medical diagnosis. Therefore, we leverage generated patient record data to focus on more complex medical diagnostics, enabling the model to handle intricate medical issues with greater accuracy.

2.2 Multi-Agent Collaboration

A growing body of research demonstrates that collaborative frameworks involving multiple LLM agents can effectively address the limitations of individual models when tackling complex tasks, resulting in more efficient and precise execution across domains such as finance, coding, literature, and mathematics (Li et al., 2023; Wu et al., b; Huot et al., 2024; Hong et al., 2023; Han et al., 2024; Zhang et al., 2023). In the medical field, which is closely tied to everyday life, multi-agent collaboration frameworks are increasingly being recognized for their potential. By leveraging collaboration between different LLM agents, tasks like diagnosis, treatment planning, and patient management can be more effectively handled.

For example, MedAgents (Tang et al., 2023), the first multi-agent framework proposed in the medical domain, has demonstrated exceptional performance in extracting and utilizing medical expertise from LLMs while improving their reasoning capabilities. Agent Hospital (Li et al.), by creating a hospital simulation environment with evolving medical agents, has achieved ongoing improvements in clinician agent performance, both in simulated and real-world settings, thereby laying the groundwork for the use of LLM-driven agent technology in medical applications. Inspired by clinicians' decision-making processes, MDAgents (Kim et al., 2024) has developed an adaptive medical decision-making framework that uses LLMs to simulate hierarchical diagnostic procedures, ranging from individual clinicians to collaborative clinical teams. This has opened new possibilities for enhancing LLM-assisted medical diagnostic systems and advancing automated clinical reasoning.

Building on the success of multi-agent collaboration frameworks, we propose a dual-agent reflection and correction mechanism, augmented by knowledge-enhancement techniques, to further improve the accuracy of generated clinical notes.

3 Method

In this section, we provide a detailed explanation of our method for generating clinical notes using a hierarchical agent reflection mechanism, along with a knowledge enhancement strategy. First, we describe the process of constructing standardized templates for clinical notes. Following this, we outline the specific functions of each agent, as well as the reflection and correction mechanisms and the strategies for knowledge enhancement. Finally, we discuss the generation of high-quality clinical notes, which are used to train models, thereby improving their ability to effectively utilize these notes for medical diagnosis.

3.1 Standardized Clinical Note

For clinicians, creating high-quality clinical notes is essential for ensuring thorough patient care and accurate diagnoses (Standard, 2012; Demsash et al., 2023). These notes provide critical documentation of patient symptoms, preliminary diagnoses, and the rationale behind the differential diagnosis. During the diagnostic process, clinicians typically start with the patient's description of symptoms, conducting further examinations and tests to gather comprehensive information that forms the patient's case profile. This information is meticulously documented in clinical notes, which clinicians use for comprehensive assessment, leading to preliminary diagnoses and corresponding rationale (Ball et al., 2015; Gale and Martin Gale, 2022; Vally et al., 2023). Subsequently, clinicians apply their distinctive differential diagnostic reasoning to verify the preliminary diagnosis and ultimately confirm the disease.

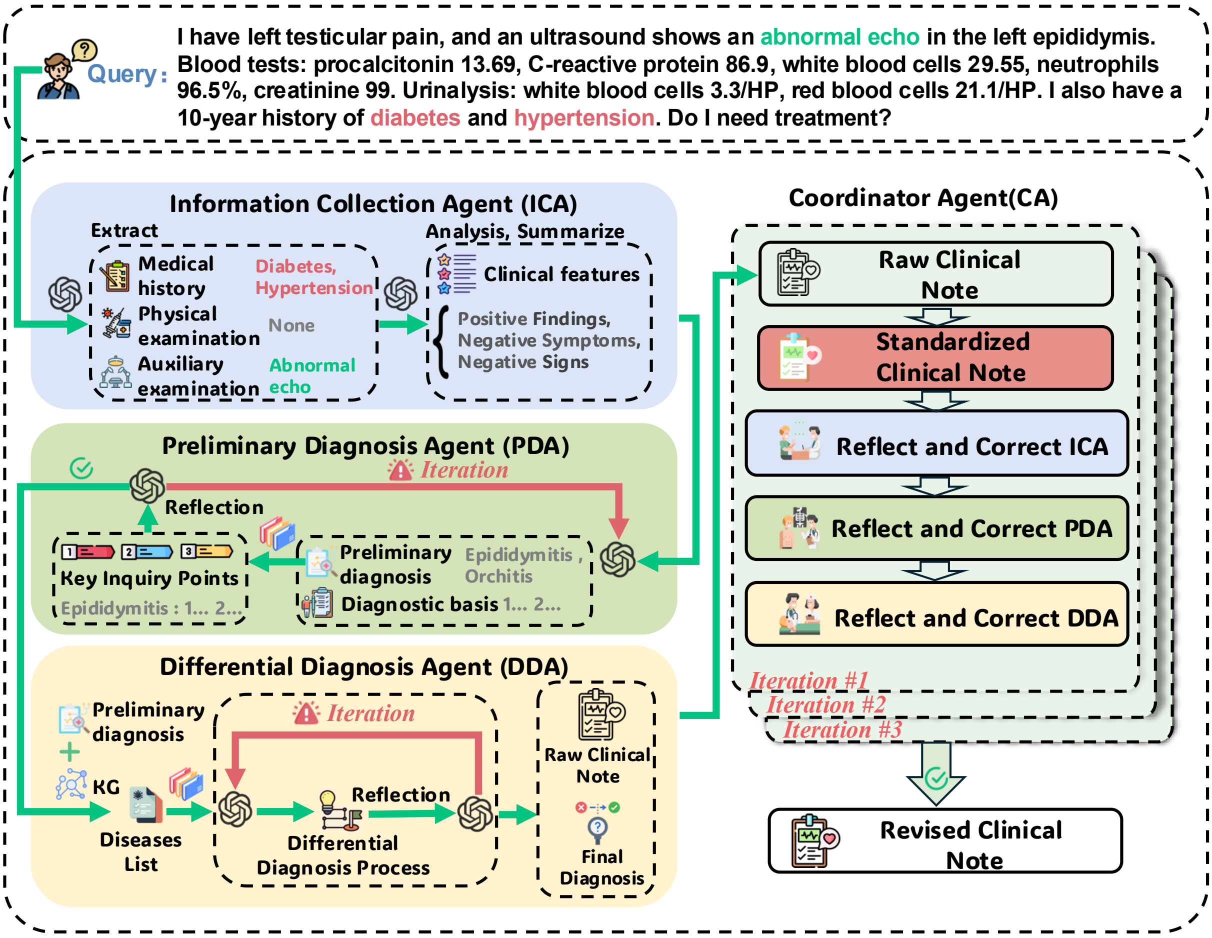

Figure 2: The overview of our proposed framework Hierarchical agent reflection. refers to the knowledge base we collected as mentioned in Section 3.1, and refers to the LLM.

To simulate this complex diagnostic process and construct high-quality generative clinical note data, we first collected all 433 diseases from the dataset (Jin et al., 2021; Lyu et al., 2023) and compiled the corresponding key inquiry points from medical textbooks. which served as the knowledge base used in our approach. Subsequently, we utilized the powerful generative capabilities of DeepSeek-R1 (Guo et al., 2025) to generate initial clinical note templates for each disease. For detailed information on the prompts used to generate the medical record templates, please refer to Appendix B.1. After generating the initial templates with DeepSeek-R1, we invited three clinical physicians to annotate and revise each disease's clinical note individually. Ultimately, this process resulted in a standardized disease course note for each condition. These disease templates not only encompass the key symptoms of each condition but also systematically reflect the typical diagnostic processes and associated medical reasoning logic, aiming to provide a comprehensive and accurate description of the characteristics of each disease. Ultimately, these templates were integrated into our methodology as the core enhancement module within the pipeline, laying the foundation for fully simulating diagnostic scenarios.

3.2 Hierarchical Agent Reflection

In our hierarchical agent reflection process, we configured four agents: the Information Collection Agent (ICA), the Preliminary Diagnosis Agent (PDA), the Differential Diagnosis Agent (DDA) and the Coordinator Agent(CA). We used Knowledge-enhanced methods to assist them in intra-agent and inter-agent reflection and correction. We provided these agents with a knowledge base, covering 433 diseases, collected in 3.1, along with a detailed map of diseases and their differential diagnoses. Specific information about the knowledge graph and the knowledge base can be found in Appendix C.1. The prompts used by each agent can be found in Appendix B. The pseudo code of Generating the Clinical Notes can be found at Appendix A.

Information Collection Agent. ICA is responsible for recording the patient's basic information, which includes their medical history, physical examination, and auxiliary tests. Subsequently, the ICA conducts a comprehensive analysis, synthesis, and organization of this information to document the characteristics of the case.

Preliminary Diagnosis Agent. Based on the case characteristics recorded by the ICA, the Preliminary Diagnosis Agent provides an initial diagnosis and its diagnostic basis. Subsequently, the PDA retrieves diagnostic key points related to the initial diagnosis from the knowledge base and uses these to reflect on the initial diagnostic process. It then evaluates the accuracy of the initial diagnosis, and if deemed inaccurate, performs iterative diagnostic corrections.

Differential Diagnosis Agent. The DDA first retrieves a list of diseases requiring differential diagnosis exclusion from the provided knowledge graph, based on the initial diagnosis provided by the PDA. It then acquires the diagnostic key points for each of these diseases from the knowledge base and performs differential diagnosis for each disease using these points and the patient's case characteristics. Finally, the DDA reflects on the reasonableness of the entire differential diagnosis process; if found to be unreasonable, it conducts iterative corrections.

Coordinator Agent. In hierarchical agent reflection, the Coordinator Agent operates at a higher level than the ICA, PDA, and DDA. Specifically, the CA first receives the raw clinical notes from the CA. It then uses the final diagnosis provided by the DDA to match this note with standardized clinical notes in the knowledge base, obtaining the standardized notes for the corresponding diseases. The CA then reflects on and evaluates whether the outputs of the ICA, PDA, and DDA align with the standards by comparing the raw clinical note with the matched standardized clinical notes. If significant discrepancies are found between an agent's output and the standardized clinical notes, the CA identifies potential errors in that agent's process and notifies it of the reasons for reflection. Conversely, if the raw clinical note is deemed reasonable, the CA integrates and outputs a verified complete clinical note. Throughout this process, the CA leverages knowledge augmentation and the hierarchical agent reflection mechanism to enhance the accuracy of the generated clinical notes.

The combination of self-reflection in the ICA, PDA, and DDA agents, along with supervisory feedback from the CA agent, enhances accuracy of the generated clinical notes. Self-reflection allows each agent to independently refine its reasoning and detect errors, while the CA agent provides additional oversight to ensure the final output aligns with clinical standards. This dual-layer feedback system improves error detection, enables better generalization across scenarios, and supports continuous adaptation, ultimately leading to more reliable and accurate clinical decision-making.

3.3 Enhance LLM Medical Diagnosis with Clinical Notes

Our Raw dataset is $\mathcal{D}_{\text{Raw}} = \{x_i, y_i\}_{i=1}^{|\mathcal{D}_{\text{Raw}}|}$, where x_i denotes a patient's question, and y_i denotes the original answer without diagnostic logic. After using our hierarchical agent reflection framework, the data format becomes: $\mathcal{D}_{\text{note}} = \{x_i, (y_{i1}, y_{i2}, y_{i3}) \to y_{i4} \to (y_{i5}, y_{i6}) \to (y_{i7}, y_{i8}, y_{i9})\}_{i=1}^{|\mathcal{D}_{\text{note}}|}$ where x_i denotes a patient's question, while y_{i1} through y_{i9} represent various components of the clinical note: y_{i1} is the medical history, y_{i2} the physical examination, y_{i3} auxiliary examination, y_{i4} clinical features, y_{i5} initial diagnosis, y_{i6} diagnostic basis, y_{i7} disease list, y_{i8} differential diagnosis process, and y_{i9} the final diagnosis. After extracting information from x_i, we obtain (y_{i1}, y_{i2}, y_{i3}), which are then further organized and summarized to derive y_{i4}. Then generating the initial diagnosis and diagnostic basis (y_{i5}, y_{i6}). Finally, the process results in a detailed differential diagnosis and the final diagnosis (y_{i7}, y_{i8}, y_{i9}).

In the standard post-training setup, pre-trained language models are fine-tuned via supervision to better follow instructions or specific formats (Ouyang et al., 2022; Zhou et al., 2024; Fan et al., 2024). We use SFT to train the model to generate clinical notes step by step, enabling it to reason using prior knowledge for patient info collection, preliminary diagnosis, and differential diagnosis. We randomly sample the prefix (which can be empty) and supervise the model to reason before responding by optimizing the following objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)}\left[\sum_{t=1}^{T} \log p_\theta\big(y_t \mid x \oplus y_{<t}\big)\right]. \quad (1)$$

	RJUA-QA F1	UDDD F1	Medbullets-5op Acc.	Medbullets-4op Acc.	JMED Acc.
∼ 7-8B Large Language Models					
LLaMA3.1-8B-Instruct	28.33 ± 0.16	58.17 ± 0.54	30.63 ± 0.50	45.56 ± 3.56	38.70 ± 0.24
Qwen2.5-7B-Instruct	36.39 ± 0.25	65.84 ± 0.34	37.23 ± 0.19	42.97 ± 0.50	57.50 ± 0.01
DeepSeek-R1-Distill-Qwen-7B	34.67 ± 0.90	60.19 ± 0.10	26.73 ± 1.32	28.79 ± 2.39	40.17 ± 0.56
Huatuo-o1-Qwen-7B	43.71 ± 1.72	72.51 ± 0.25	50.33 ± 2.82	52.69 ± 2.53	**60.28 ± 0.13**
Huatuo-o1-LLaMA-8B	34.17 ± 1.13	60.19 ± 0.10	45.07 ± 1.39	52.49 ± 1.60	57.39 ± 0.62
MedFound-7B	16.48 ± 5.33	33.08 ± 6.44	35.19 ± 2.37	29.67 ± 3.21	30.45 ± 1.69
MedFound-LLaMA3-8B-finetuned	29.64 ± 1.35	53.51 ± 1.87	18.07 ± 2.31	25.00 ± 5.77	25.64 ± 1.89
LLaMA3.1-8B-Instruct	44.71 ± 0.70	74.16 ± 0.51	50.76 ± 1.14	57.47 ± 1.75	56.37 ± 0.09
Qwen2.5-7B-Instruct	**46.50 ± 0.06**	**74.75 ± 0.02**	**53.35 ± 1.14**	**60.17 ± 1.42**	59.24 ± 0.37
> 10B Large Language Models					
GPT-3.5-Turbo	28.92 ± 0.56	55.21 ± 0.66	35.71 ± 0.57	42.75 ± 0.68	47.35 ± 0.17
GPT-4-turbo	31.14 ± 0.02	59.95 ± 0.31	58.23 ± 0.18	65.26 ± 0.97	56.24 ± 1.34
GPT-4o	33.98 ± 0.05	65.32 ± 0.64	69.48 ± 0.56	75.00 ± 0.65	64.60 ± 1.27
DeepSeek-V3	37.34 ± 0.01	66.58 ± 0.09	56.71 ± 0.49	61.69 ± 0.65	64.65 ± 1.58
HuatuoGPT2-13B	33.13 ± 0.79	59.70 ± 0.99	37.77 ± 1.35	37.23 ± 3.47	40.21 ± 2.15
Baichuan-M1-14B	50.01 ± 1.05	75.60 ± 1.62	55.52 ± 1.17	61.58 ± 0.49	67.25 ± 0.29
LLaMA3.1-70B-Instruct	35.54 ± 0.67	66.65 ± 1.05	57.58 ± 1.05	64.29 ± 0.33	55.90 ± 1.48
Qwen2.5-72B-Instruct	38.54 ± 0.65	66.08 ± 1.07	54.76 ± 0.49	62.88 ± 0.99	66.70 ± 1.89
Huatuo-o1-LLaMA-70B	38.11 ± 0.96	68.07 ± 0.17	68.83 ± 0.65	73.38 ± 1.72	64.67 ± 1.18
Citrus1.0-llama-70B	36.13 ± 0.34	59.59 ± 0.54	66.23 ± 0.31	**78.57 ± 0.26**	**68.40 ± 0.00**
LLaMA3.1-70B-Instruct	44.93 ± 0.42	73.25 ± 0.39	70.24 ± 0.65	74.78 ± 1.21	65.24 ± 0.28
Qwen2.5-72B-Instruct	**50.61 ± 2.10**	**77.29 ± 0.91**	**71.21 ± 1.14**	76.52 ± 0.65	67.79 ± 0.62

Table 1: Main Results on Medical Benchmarks. LLMs with are specifically trained for the medical domain, and indicates LLMs training for our clinical note dataset. The **bold** highlights the best scores, and underlines indicate the second-best.

4 Experiments

4.1 Experimental Setup

Training Data We construct a Chinese clinical note dataset containing 2K notes and an English clinical notes dataset containing 10K notes respectively from the training sets of RJUA-QA (Lyu et al., 2023) and MedQA (Jin et al., 2021) by applying our hierarchical agent reflection framework.

Model Training After obtaining the dataset of clinical notes generated by our framework, we trained the LLM using LLaMA-Factory (Zheng et al., 2024), a widely-used library for LLM training. We conducted all experiments on eight NVIDIA A100 (80G) GPUs. Specifically, we fine-tuned the model using LoRA (Hu et al., 2021) with the DeepSpeed (Rasley et al., 2020) library and Zero Redundancy Optimizer (ZeRO) (Rajbhandari et al., 2020) Stage 2. For SFT, we set the epoch to 3, the learning rate to 5e-5, and the context length to 4096.

Baselines We utilized the generated clinical note data for finetuning of the model and compared the results with two types of LLMs: **1) General LLMs**: the GPT series (Achiam et al., 2023), Qwen-2.5 (Team, 2024), LLaMA-3.1 (Dubey et al., 2024) and DeepSeek-V3 (DeepSeek-AI, 2025); and **2) Medical-Specific LLMs:** Huatuo series models (Chen et al., 2024b), MedFound (Liu et al., 2025), Citrus (Wang et al., 2025b), and Baichuan-M1 (Wang et al., 2025a).

Benchmarks We evaluate on the standard medical diagnostic benchmarks: including the *RJUA-QA*(test set) (Lyu et al., 2023) and *Urological Disease Diagnosis Dataset(UDDD)*, both of which are Chinese medical diagnosis datasets, using the F1 score to assess diagnostic accuracy. Additionally, we evaluated *Medbullets* (Chen et al., 2024a) and *JMED* (Wang et al., 2025b), both of which are single-choice medical diagnosis datasets, using accuracy as the metric for assessing diagnostic performance. To enhance the reliability of the experimental results, we ran every evaluation 3 times and averaged the results and variance.

4.2 Experimental Results

Main Results We evaluated various LLMs on medical benchmarks, as shown in Table 1. The results indicate that foundational models, which have not undergone enhanced training with spe-

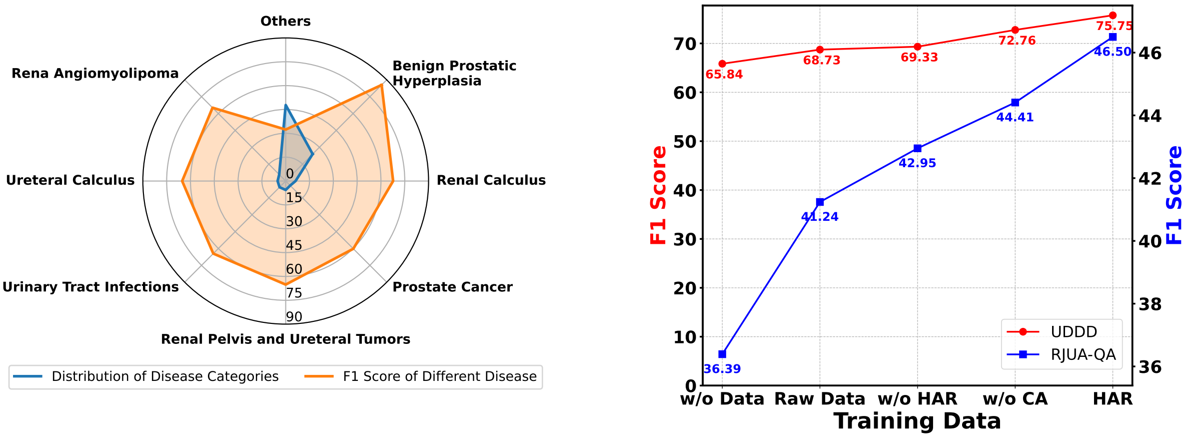

Figure 3: Diagnostic ability for different diseases (left) and ablation results on training data (right).

cialized medical knowledge, perform rather poorly in the medical diagnosis domain. This is evident in models such as Qwen2.5-7B and LLaMa3.1-8B. Even when the parameter scale of these models is increased, the improvements in performance remain quite limited. For currently popular models, such as GPT-4o (OpenAI, 2024) and DeepSeek-V3 (DeepSeek-AI, 2025), their performance on medical diagnosis datasets remains inadequate. This further highlights that solely relying on general capabilities cannot achieve optimal results in the medical field. In contrast, the Citrus model (Wang et al., 2025b), the Huatuo series models, and the Baichuan-M1 model (Wang et al., 2025a) demonstrate more significant diagnostic capabilities in the field of medical diagnosis.

After undergoing SFT on datasets generated by our method, the models consistently outperformed their original Qwen and LLaMA baselines across all five benchmark test sets. Notably, the fine-tuned models achieved SOTA results on three of these datasets, surpassing other domain-specialized models, including Huatuo-o1, Baichuan-M1, and Citrus, all of which were explicitly optimized for medical tasks. These results underscore the effectiveness of our dataset in enhancing model performance for clinical NLP applications.

4.3 Ablations Study

In this section, we thoroughly explore the impact of variations in disease distribution on diagnostic performance, while providing a detailed evaluation of the performance of individual components within our framework during the data generation process.

Disease Distribution on Diagnostic Performance In real clinical settings, disease distribution often exhibits significant imbalances. For instance, in the field of urology, the prevalence of benign prostatic hyperplasia (BPH) exceeds 50% among men over the age of 50, whereas the annual incidence of rare diseases such as urethral gland carcinoma is less than one per million. Fig. 3(left) illustrates our experimental results based on the RJUA dataset (Lyu et al., 2023), where we evaluated the diagnostic performance of the Qwen2.5-7B model, fine-tuned on our dataset, under varying disease distributions. The results show that for prevalent diseases (e.g., BPH), the model achieves a diagnostic accuracy of over 60%. However, its accuracy declines when diagnosing rare diseases.

Performance of Individual Components The Fig. 3(right) presents the results of ablation experiments on the Qwen2.5-7B model for diagnostic tasks after fine-tuning with different training corpora. Among them, (1) **w/o Data** means no training data is used, (2) **Raw Data** refers to the fine-tuning data consisting of the original RJAU-QA training set (Lyu et al., 2023). (3) **w/o HAR** denotes the dataset generated without hierarchical agent reflection, involving only the ICA, PDA, and DDA without reflection. (4) **w/o CA** indicates the dataset generated by removing the CA for upper-level reflection. (5) **Com. HAR** represents the dataset generated through the complete hierarchical agent reflection framework. Without fine-tuning on medical data, the base model demonstrates poor diagnostic performance. After fine-tuning using the original RJUA-QA training set, diagnostic accuracy shows improvement. Replacing the training data with preliminary clinical notes enables the model to simulate a doctor's diagnostic logic, further enhancing diagnostic accuracy. Incorporating disease knowledge into the agent improves the quality of the generated clinical note data, leading to additional gains in diagnostic performance. Finally, the integration of a hierarchical agent reflection framework and standardized clinical note templates results in significant advancements in the model's diagnostic capabilities.

5 Analysis

5.1 Automated Evaluation

To comprehensively evaluate the quality of the generated clinical notes, we designed an automated scoring framework, with detailed descriptions provided in Appendix D.1. Each clinical note was assigned a maximum total score of 40, with individual sections assessed using a LLM. By setting appropriate score thresholds, we filtered and curated a high-quality clinical notes dataset. Following the methodology proposed in Section 3, we utilized different LLMs to generate 100 clinical notes for each model. The quality evaluation results of the clinical notes generated by different models are illustrated in Appendix D.2. The results indicate that GPT-4o (OpenAI, 2024) achieves the highest quality in generating clinical notes.

5.2 Impact of Reflection Iterations

In our method, both the PDA and the DDA involve multiple rounds of reflective iterations. Therefore, we analyzed the impact of the number of reflective iterations (N) for PDA and DDA on the quality of clinical note generation. The experiments were conducted with $N \in \{1, 2, 3, 4, 5, 10, 20\}$, generating 50 clinical notes for each configuration. The notes were evaluated using the pipeline detailed in Appendix D.1. Results D.3 indicated a significant positive correlation between the number of reflective iterations and the quality scores of the generated notes—more reflective iterations effectively improved the alignment of the generated content with clinical standards. However, when $N > 5$, the quality improvement exhibited a diminishing marginal return. Considering computational efficiency and economic costs, we ultimately selected N = 5 as the optimal configuration.

5.3 Expert Evaluation

In the inherently rigorous field of medicine, expert evaluation is indispensable. To ensure a comprehensive assessment, we invited three physicians with varying levels of expertise: a urology specialist, a doctoral candidate in oncology, and a doctoral candidate in urology. A random sample of 50 generated clinical notes was selected from the dataset, and the experts independently scored them using the evaluation criteria outlined in Appendix D.1. The detailed scoring results are summarized in Appendix D.4. The average score of the 50 clinical notes evaluated by experts was 25.3, which closely aligns with the scores obtained through our automated evaluation using the LLM.

5.4 Case Study

To understand the differences in diagnostic accuracy and transparency between the model fine-tuned using the hierarchical agent reflection framework and other medical or base models, we manually compared their diagnostic results on the same medical issue (see Appendix **??**). The base model is disorganized and fails to utilize patient information adequately. The HuatuoGPT-o1 model shows medical knowledge errors, such as not recognizing that Mirabegron is a β3-adrenergic receptor agonist used for overactive bladder symptoms. The Baichuan-M1 model struggles to differentiate similar urinary incontinence diseases. In contrast, our fine-tuned model delivers a clear diagnostic process that better aligns with the clinical reasoning of clinicians. This is achieved by using our hierarchical agent reflection framework during the generation of the clinical note training dataset, which injects the model with the correct inquiry points related to diseases, enabling it to recognize specialized medical knowledge. Furthermore, our model effectively employs differential diagnosis techniques to exclude similar diseases.

6 Conclusion and Future Work

In this paper, we propose a hierarchical agent reflection framework to generate high-quality clinical notes. By training LLMs with clinical notes that reflect the reasoning processes of clinicians, we aim to enhance the model's ability to engage in medical reasoning and improve diagnostic accuracy. The model's output not only mirrors the reasoning clinicians use in diagnosis but also assists them by offering a similar thought process during clinical decision-making. Experimental results demonstrate that simulating clinicians' use of clinical notes for diagnosis significantly boosts the model's diagnostic performance. Moving forward, we plan to extend the framework to cover rare diseases and refine the model's reasoning capabilities for even greater diagnostic accuracy.

Limitations

In this paper, we aim to develop a hierarchical agent reflection framework that narrows the gap between model-based diagnostic processes and the diagnostic logic used by clinicians by generating high-quality clinical note data. Despite our best efforts, certain limitations remain. First, our current work is limited to text-based medical diagnoses, while the medical field often involves a wealth of multimodal information that aids in diagnosis. Second, when it comes to rare and complex diseases, our framework lacks the capability to compose the discussion section of challenging cases. We plan to address these limitations in future work.

References

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

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*.

John R Ball, Bryan T Miller, and Erin P Balogh. 2015. Improving diagnosis in health care.

E Bolton, A Venigalla, M Yasunaga, D Hall, B Xiong, T Lee, R Daneshjou, J Frankle, P Liang, M Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text, arxiv, 2024. *arXiv preprint arXiv:2403.18421*.

Peter G Brodeur, Thomas A Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian Haimovich, Jason A Freed, et al. 2024. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*.

Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. 2025. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine. *arXiv preprint arXiv:2502.09242*.

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

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

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Addisalem Workie Demsash, Sisay Yitayih Kassie, Abiy Tasew Dubale, Alex Ayenew Chereka, Habtamu Setegn Ngusie, Mekonnen Kenate Hunde, Milkias Dugassa Emanu, Adamu Ambachew Shibabaw, and Agmasie Damtew Walle. 2023. Health professionals' routine practice documentation and its associated factors in a resource-limited setting: a cross-sectional study. *BMJ health & care informatics*, 30(1):e100699.

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

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. 2024. Reformatted alignment. *arXiv preprint arXiv:2402.12219*.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*.

Dyke Ferber, Isabella C Wiest, Georg Wölflein, Matthias P Ebert, Gernot Beutel, Jan-Niklas Eckardt, Daniel Truhn, Christoph Springfeld, Dirk Jäger, and Jakob Nikolas Kather. 2024. Gpt-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI*, 1(6):AIcs2300235.

Martin S Gale and BDS Martin Gale. 2022. Diagnosis: fundamental principles and methods. *Cureus*, 14(9).

Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. 2023. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.

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

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

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

Xuewen Han, Neng Wang, Shangkun Che, Hongyang Yang, Kunpeng Zhang, and Sean Xin Xu. 2024. Enhancing investment analysis: Optimizing ai-agent collaboration in financial research. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 538–546.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

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

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

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.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. arxiv 2024. *arXiv preprint arXiv:2401.06866*.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18417–18425.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

J Li, S Wang, M Zhang, W Li, Y Lai, X Kang, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents [internet]. arxiv; 2024 [cited 2024 may 10].

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. 2025. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, pages 1–11.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, et al. 2023. Rjua-qa: A comprehensive qa dataset for urology. *arXiv preprint arXiv:2312.09785*.

Shiva Maleki Varnosfaderani and Mohamad Forouzanfar. 2024. The role of ai in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering*, 11(4):337.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

OpenAI. 2024. Hello gpt-4o. *OpenAI*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1).

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Justin T Reese, Daniel Danis, J Harry Caufield, Tudor Groza, Elena Casiraghi, Giorgio Valentini, Christopher J Mungall, and Peter N Robinson. 2024. On the limitations of large language models in clinical diagnosis. *medRxiv*, pages 2023–07.

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

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Hardeep Singh, Mark L Graber, and Timothy P Hofer. 2019. Measures to improve diagnostic safety in clinical practice. *Journal of patient safety*, 15(4):311–316.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 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, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Quality Improvement Guide Standard. 2012. Australian commission on safety and quality in health care.

Erik Stolper, Paul Van Royen, Edmund Jack, Jeroen Uleman, and Marcel Olde Rikkert. 2021. Embracing complexity with systems thinking in general practitioners' clinical reasoning helps handling uncertainty. *Journal of Evaluation in Clinical Practice*, 27(5):1175–1181.

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

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Daniel Truhn, Jan-Niklas Eckardt, Dyke Ferber, and Jakob Nikolas Kather. 2024. Large language models and multimodal foundation models for precision oncology. *NPJ Precision Oncology*, 8(1):72.

Zunaid Ismail Vally, Razia AG Khammissa, Gal Feller, Raoul Ballyram, Michaela Beetge, and Liviu Feller. 2023. Errors in clinical diagnosis: a narrative review. *Journal of International Medical Research*, 51(8):03000605231162798.

Meredith Vanstone, Sandra Monteiro, Eamon Colvin, Geoff Norman, Jonathan Sherbino, Matthew Sibbald, Kelly Dore, and Amanda Peters. 2019. Experienced physician descriptions of intuition in clinical reasoning: a typology. *Diagnosis*, 6(3):259–268.

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

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *arXiv preprint arXiv:2408.01933*.

Guangyu Wang and Xiaohong Liu. 2025. Medical large language model for diagnostic reasoning across specialties.

Guoxin Wang, Minyu Gao, Shuai Yang, Ya Zhang, Lizhi He, Liang Huang, Hanlin Xiao, Yexuan Zhang, Wanyue Li, Lu Chen, et al. 2025b. Citrus: Leveraging expert cognitive pathways in a medical language model for advanced medical decision support. *arXiv preprint arXiv:2502.18274*.

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

J Wu, J Zhu, and Y Qi. a. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. arxiv 2024. *arXiv preprint arXiv:2408.04187*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Q Wu, G Bansal, J Zhang, Y Wu, B Li, E Zhu, L Jiang, X Zhang, S Zhang, J Liu, et al. b. Autogen: Enabling next-gen llm applications via multi-agent conversation,(2023). *arXiv preprint arXiv:2308.08155*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

Kuo Zhang, Xiangbin Meng, Xiangyu Yan, Jiaming Ji, Jingqian Liu, Hua Xu, Heng Zhang, Da Liu, Jingjia Wang, Xuliang Wang, et al. 2025. Revolutionizing health care: The transformative impact of large language models in medicine. *Journal of Medical Internet Research*, 27:e59069.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

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

A Algorithm

In this algorithm, the MH_i, PE_i andAE_i respectively denote Medical History, Physical Examination, and Auxiliary Examination , while CF stands for Clinical Features. The $\mathcal{D}$ and $\mathcal{B}$ represent Preliminary Diagnosis and Diagnostic Basis, respectively, and $\mathcal{K}$ refers to key inquiry points associated with a disease, sourced from the knowledge base. The $\mathrm{Dia_{flag}}$ and $\mathrm{Dia_{error}}$ are used to indicate the correctness of the preliminary diagnosis reflection and areas for improvement in subsequent iterations should errors occur. The $\mathrm{DisList}$ represents the list of diseases that require differential diagnosis to be excluded based on the preliminary diagnosis results. The process of differential diagnosis is denoted by Diff, while $\mathrm{Diff_{flag}}$ and $\mathrm{Diff_{error}}$ indicate the correctness of the differential diagnosis reflection and the necessary improvements for future iterations in case of errors. Finally, $\mathrm{T_{flag}}$ and $\mathrm{T_{error}}$ signify the correctness evaluation after CA reflection and highlight the aspects that need to be communicated to other agents for improvement if errors are detected.

Algorithm 1 Generate the Clinical Notes

Input: Question Q, Knowledge Graph KG, Clinical Note Template T, Disease Knowledge DK
{**Initialization**}
Initialize Information Collection Agent ICA, Preliminary Diagnosis Agent PDA,
Differential Diagnosis Agent DDA and Coordinator Agent CA, Maximum Attempts N
{**Generate the Progress Notes**}
for try count $i = 0$ to $N-1$ and $\mathrm{T_{flag}}$ is $False$ **do**
  $MH_i, PE_i, AE_i \leftarrow$ Extraction of $\mathrm{ICA(Q)}$
  $CF_i \leftarrow$ Summarization of $\mathrm{ICA}(MH_i, PE_i, AE_i)$
  $\mathrm{Dia_{error}} \leftarrow$ None, $\mathrm{Diff_{error}} \leftarrow$ None
  for $\mathrm{Dia_{flag}}$ is $False$ and $j = 0$ to $N-1$ **do**
    $\mathcal{D}_j, \mathcal{B}_j \leftarrow$ Diagnosis of $\mathrm{PDA}(CC_i, \mathrm{Dia_{error}})$
    $\mathcal{K}_j \leftarrow$ Retrieve of $\mathrm{PDA}(\mathcal{D}_j, \mathrm{DK})$
    $\mathrm{Dia_{flag}}, \mathrm{Dia_{error}} \leftarrow$ Reflection of $\mathrm{PDA}(\mathcal{D}_j, \mathcal{B}_j, \mathcal{K}_j)$
  end for
  for $\mathrm{Diff_{flag}}$ is $False$ and $j = 0$ to $N-1$ **do**
    $\mathrm{DisList}_j \leftarrow$ Differential Diagnosis List of $\mathrm{DDA}(\mathcal{D}_j, \mathrm{KG})$
    $\mathcal{K}_j \leftarrow$ Retrieve of $\mathrm{DDA}(\mathrm{DisList}_j, \mathrm{DK})$
    $\mathrm{Diff}_j \leftarrow$ Differential Process of $\mathrm{PDA}(\mathrm{DisList}_j, \mathcal{K}_j, \mathrm{Diff_{error}})$
    $\mathrm{Diff_{flag}}, \mathrm{Diff_{error}} \leftarrow$ Reflection of $\mathrm{DDA}(\mathrm{DisList}_j, \mathrm{Diff}_j, \mathcal{K}_j)$
  end for
  $\mathrm{RawNote_i} \leftarrow$ Output Raw Clinical Note(ICA, PDA, DDA)
  $\mathcal{T}_i \leftarrow$ Retrieval Standardized Clinical Note Template of $\mathrm{CA}(\mathcal{D}_j, \mathrm{T})$
  $\mathrm{T_{flag}}, \mathrm{T_{error}} \leftarrow$ Reflection of $\mathrm{CA}(\mathrm{ICA}, \mathcal{T}_i)$
  if $\mathrm{T_{flag}}$ is $False$ **then**
    ICA $\leftarrow$ Corrective of $\mathrm{ICA(T_{error})}$
  end if
  $\mathrm{T_{flag}}, \mathrm{T_{error}} \leftarrow$ Reflection of $\mathrm{CA}(\mathrm{PDA}, \mathcal{T}_i)$
  if $\mathrm{T_{flag}}$ is $False$ **then**
    PDA $\leftarrow$ Corrective of $\mathrm{PDA(T_{error})}$
  end if
  $\mathrm{T_{flag}}, \mathrm{T_{error}} \leftarrow$ Reflection of $\mathrm{CA}(\mathrm{DDA}, \mathcal{T}_i)$
  if $\mathrm{T_{flag}}$ is $False$ **then**
    DDA $\leftarrow$ Corrective of $\mathrm{DDA(T_{error})}$
  end if
end for
Return Revised Clinical Note $\mathrm{(ICA, PDA, DDA)}$

B Prompt Templates

B.1 Generate Raw Clinical Note

Generate Raw Clinical Note Prompt

You are an experienced medical expert skilled in drafting standardized medical course records based on diseases and key consultation points. Please use the provided disease information and corresponding consultation points, along with the given template and supplied knowledge, to compose a standardized medical course record for this disease.

Below is the knowledge to this disease:

`{{disease}}`

`{{Diagnostic key points}}`

Below is the template for the clinical note:

Medical history:\n\n Physical examination:\n\n Auxiliary examination:\n\n Case characteristics:\n\n Initial diagnosis:\n\n Diagnostic basis:\n\n Diseases List:\n\n Differential diagnosis process:\n\n Final diagnosis:

B.2 Information Collection Agent Setting

Patient Information Extraction Prompt

You are an experienced clinical note specialist, adept at extracting the medical history, physical examination, and auxiliary examination information from data provided by patient. Please use the information provided by the patient to systematically consider and itemize the medical history, physical examination, and auxiliary examinations. If certain data are not provided, mark the corresponding section as 'None' without making additional assumptions.

Below is the patient's question:

`{{question}}`

Analysis and Summarize Prompt

You are an experienced medical analysis expert, skilled in comprehensively analyzing, summarizing, and organizing a patient's medical history, physical examination, and auxiliary examination to document the patient's clinical features. Please carefully review the patient's issues and itemize the clinical features, including positive findings and negative symptoms and signs relevant for differential diagnosis. Be sure to use only the provided information, without referencing external data.

Below is the medical history, physical examination, and auxiliary examination to this patient:

`{{Medical history}}`

`{{Physical examination}}`

`{{Auxiliary examination}}`

Below is the patient's question:

`{{question}}`

B.3 Preliminary Diagnosis Agent Setting

Make Preliminary Diagnosis Prompt

You are an experienced clinical diagnosis expert, skilled in making preliminary diagnoses and analyses based on provided patient clinical features. Please provide a preliminary diagnosis based on the patient's case features and detail the diagnostic basis point by point.

Below is the clinical features to this patient:

`{{Clinical features}}`

Below is the patient's question:

`{{question}}`

Reflect Preliminary Diagnosis Prompt

You are an experienced clinical review expert, skilled in evaluating the diagnostic validity of clinical notes based on key inquiry points for diseases. Please thoroughly review the key inquiry points of the preliminary diagnosis provided and assess whether the preliminary diagnosis and diagnostic basis in the clinical note align with these points.
If deemed unreasonable, output the result as a JSON-formatted Dict{"flag": false, "diagnosis_error": Str(Reasons for diagnostic errors)}. 1011

Below is the preliminary diagnosis and diagnostic basis:
`{{Preliminary Diagnosis}}`
`{{Diagnostic Basis}}`

Below is the key inquiry points:
`{{key inquiry points}}`

B.4 Differential Diagnosis Agent Setting 1012

Differential Diagnosis Prompt

You are an experienced differential diagnosis expert, skilled in systematically analyzing key inquiry points to rule out diseases. Please carefully review the inquiry points of the diseases requiring differentiation and conduct a step-by-step differential diagnosis based on the patient's clinical note.
Document the differential diagnosis process point by point and output it in JSON format as Dict{"diff_process": Str(differential diagnosis process)}. 1013

Below is the list of diseases to be ruled out through differential diagnosis:
`{{Diseases List}}`

Below is the key inquiry points to these diseases:
`{{key inquiry points}}`

Reflect Differential Diagnosis Process Prompt

You are an experienced clinical differential diagnosis expert, skilled in reflecting on and evaluating the rationality of differential diagnosis processes. Please reflect on the differential diagnosis process and assess whether the differentiation for each disease is reasonable.
If it is reasonable, output in JSON format as Dict{"flag":true, "Final_Diagnosis": Str(final diagnosis)}.
Otherwise, output in JSON format as Dict{"flag":false, "diff_error": Str(Diseases requiring rediagnosis)}. 1014

Below is the list of diseases to be ruled out through differential diagnosis, along with the corresponding diagnostic process.
`{{Diseases List}}`
`{{Differential Diagnosis Process}}`

B.5 Coordinator Agent Setting 1015

Reflect and Correct ICA Prompt

You are an experienced expert in reviewing clinical notes, skilled in comparing raw clinical note with a given standardized template. Now, please compare the obtained raw clinical note with the given standardized clinical note template. The part that needs to be analyzed is the medical history, physical examination, auxiliary examination, and clinical features. If you find any part to be unreasonable, provide suggestions for improvement, and output in JSON format as Dict{"flag":false, "ICA_error": Str(suggestions for improvement)}. 1016

Below is the raw clinical note.
`{{Raw Clinical Note}}`

Below is a standardized template for a standardized clinical note of the final diagnosis.
`{{Standardized Clinical Note}}`

Reflect and Correct PDA Prompt

You are an experienced expert in reviewing clinical notes, skilled in comparing raw clinical note with a given standardized template. Now, please compare the obtained raw clinical note with the given standardized clinical note template. The part that needs to be analyzed is the preliminary diagnosis and diagnostic basis. If you think this part is unreasonable, please give suggestions for improvement. , and output in JSON format as Dict{"flag":false, "PDA_error": Str(suggestions for improvement)}.

Below is the raw clinical note.
`{{Raw Clinical Note}}`

Below is a standardized template for a standardized clinical note of the final diagnosis.
`{{Standardized Clinical Note}}`

Reflect and Correct DDA Prompt

You are an experienced expert in reviewing clinical notes, skilled in comparing raw clinical note with a given standardized template. Now, please compare the obtained raw clinical note with the given standardized clinical note template. The part that needs to be analyzed is the diseases list and differential diagnosis process. If you think this part is unreasonable, please give suggestions for improvement. , and output in JSON format as Dict{"flag":false, "DDA_error": Str(suggestions for improvement)}.

Below is the raw clinical note.
`{{Raw Clinical Note}}`

Below is a standardized template for a standardized clinical note of the final diagnosis.
`{{Standardized Clinical Note}}`

C Knowledge Base and Knowledge Graph

C.1 Knowledge Graph

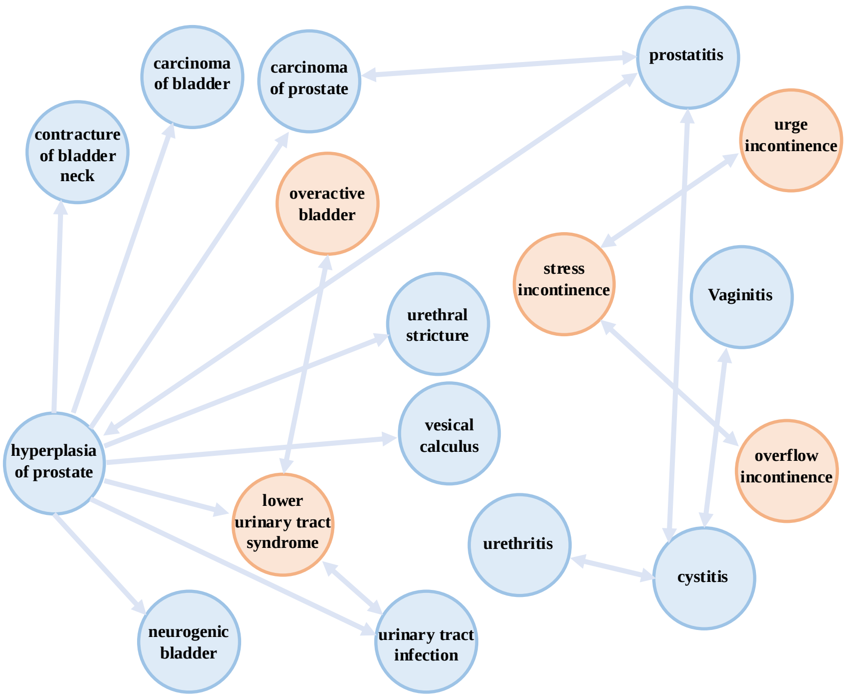

Figure 4: A knowledge graph of diseases and those requiring differential diagnosis, with ◯ refers to the diseases used in this paper.

C.2 Knowledge Base

Standardized Clinical Note

Medical history : 1. The 68-year-old patient experiences urinary leakage when coughing, sneezing, or during urgency. 2. She used Mirabegron for one month, with symptom improvement during treatment, but symptoms recurred within two days after she stopped the medication. 3. She underwent coronary intervention two months ago.
Physical examination : None
Auxiliary examination : 1. In urinalysis, the microscopic white blood cell count was 27.7/HPF two months ago and 2.1/HPF in the most recent analysis.
Clinical features : 1. The patient is a 68-year-old female who has recently experienced frequent nighttime urination and incontinence, with normal urination frequency during the day but requiring three trips at night. 2. She experiences urinary leakage when coughing, sneezing, and during urgency. 3. She used Mirabegron for one month, which improved symptoms, but they recurred after discontinuation. 4. She underwent coronary intervention two months ago. 5. Urinalysis showed a high white blood cell count of 27.7/HPF two months ago, which has since decreased to normal levels at 2.1/HPF in the most recent analysis.
Initial diagnosis : stress incontinence, overactive bladder
Diagnostic basis : 1. The patient experiences urinary leakage during coughing and sneezing, which is indicative of typical stress urinary incontinence. 2. The patient exhibits urgency and increased nighttime urination, consistent with overactive bladder, but lacks other symptoms such as frequency and dysuria. The effectiveness of Mirabegron, a medication primarily used for overactive bladder, further supports this diagnosis.
Diseases List : urge incontinence, overflow incontinence, lower urinary tract syndrome
Differential diagnosis process : 1. The patient experiences urinary leakage during urgency without symptoms like frequency or dysuria, and shows improvement with Mirabegron, allowing us to preliminarily rule out urge incontinence. 2. Overflow incontinence is often caused by lower urinary tract obstruction, such as prostatic hyperplasia. This patient has no relevant history, and the white blood cell count in the urinalysis has returned to normal, largely excluding this possibility. 3. Lower urinary tract syndrome encompasses various symptoms like frequency, urgency, and dysuria. The patient only exhibits urgency and leakage, and responds well to Mirabegron, which does not strongly align with the characteristics of lower urinary tract syndrome.
Final diagnosis : stress incontinence, overactive bladder

D Evaluation

D.1 Automated evaluation

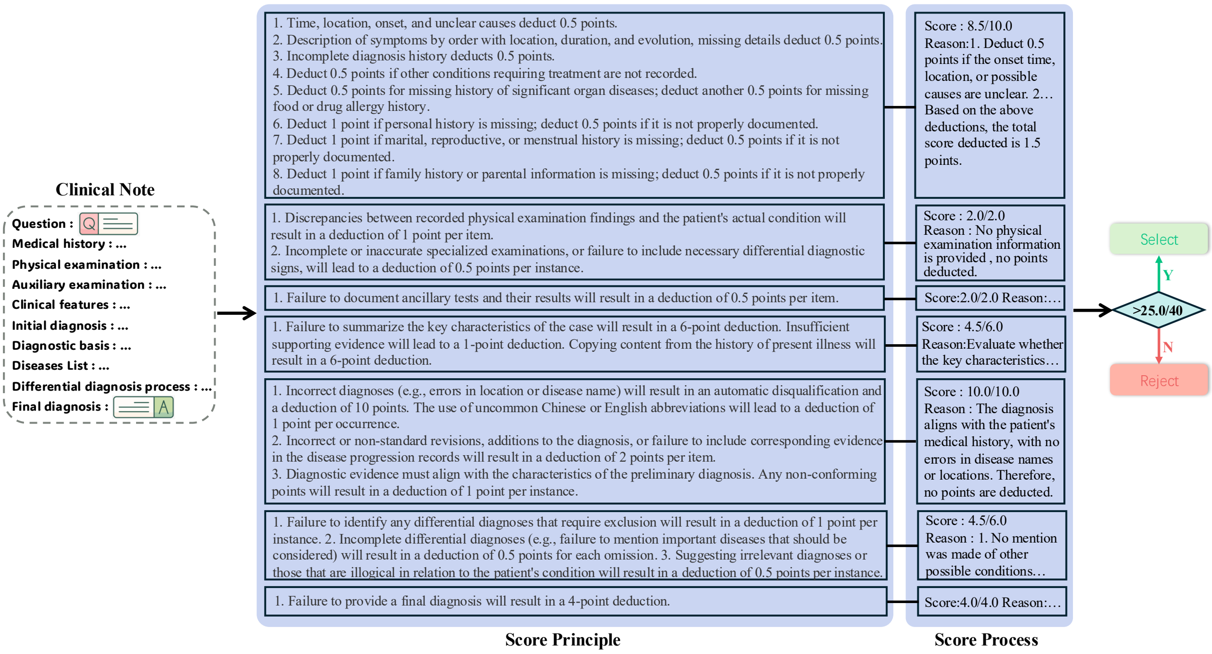

Figure 5: The process of scoring each part of the clinical note and filtering based on the scores.

D.2 Quality Scores by Section Across LLMs

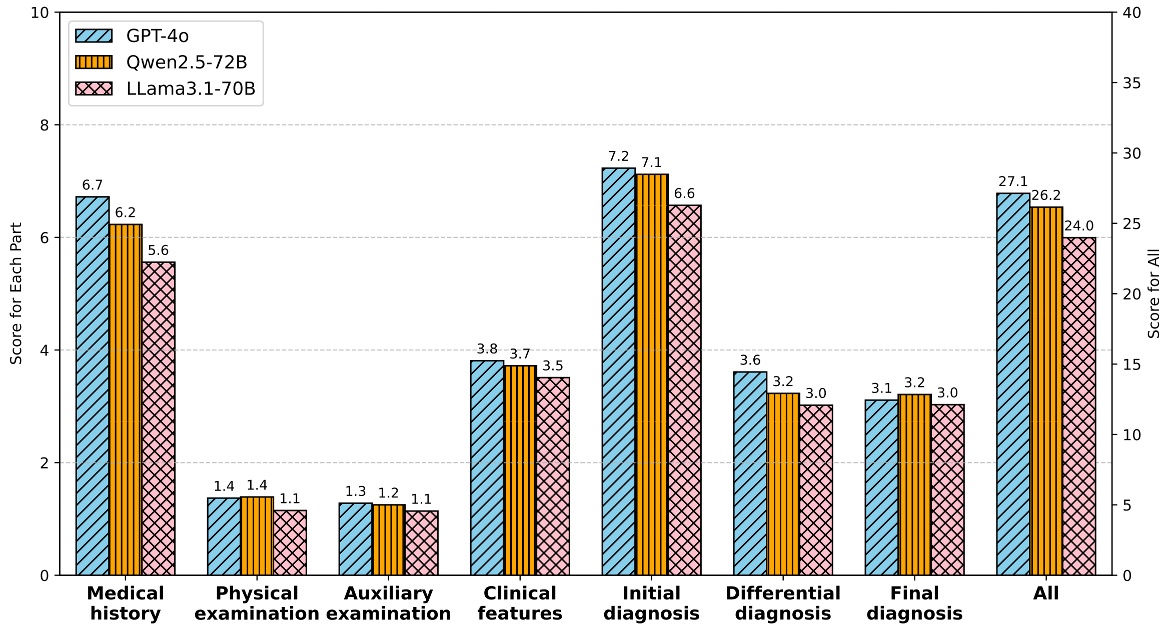

Figure 6: The quality scores for different sections of the clinical notes generated by various LLMs.

D.3 Impact of Reflection Iterations

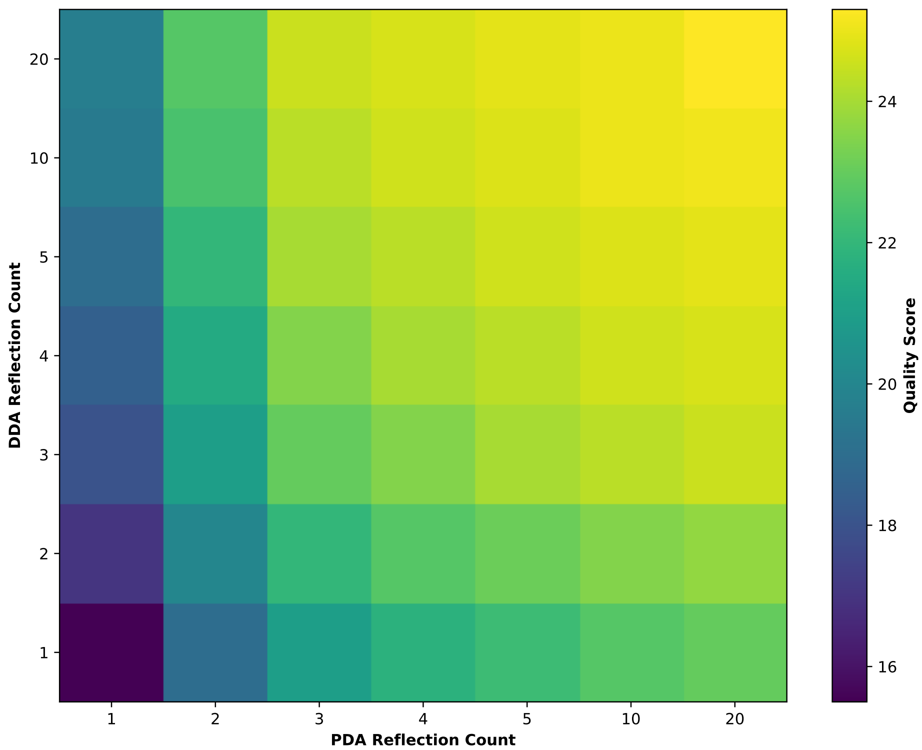

Figure 7: The impact of the number of reflection iterations in DDA and PDA on the quality of clinical notes.

D.4 Expert evaluation

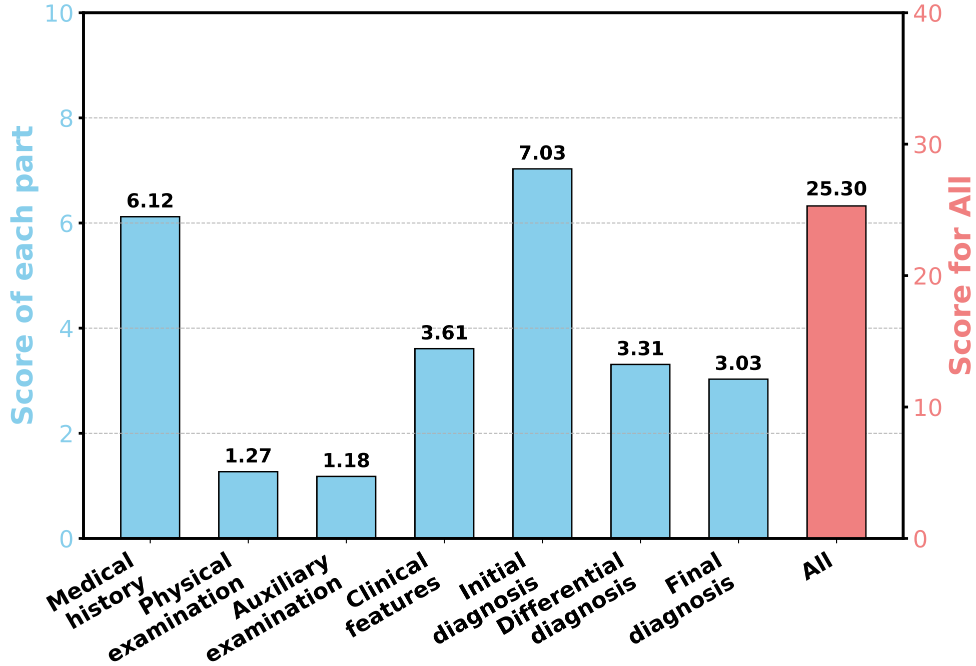

Figure 8: Medical experts scored each section of 50 clinical notes.

D.5 Case Study

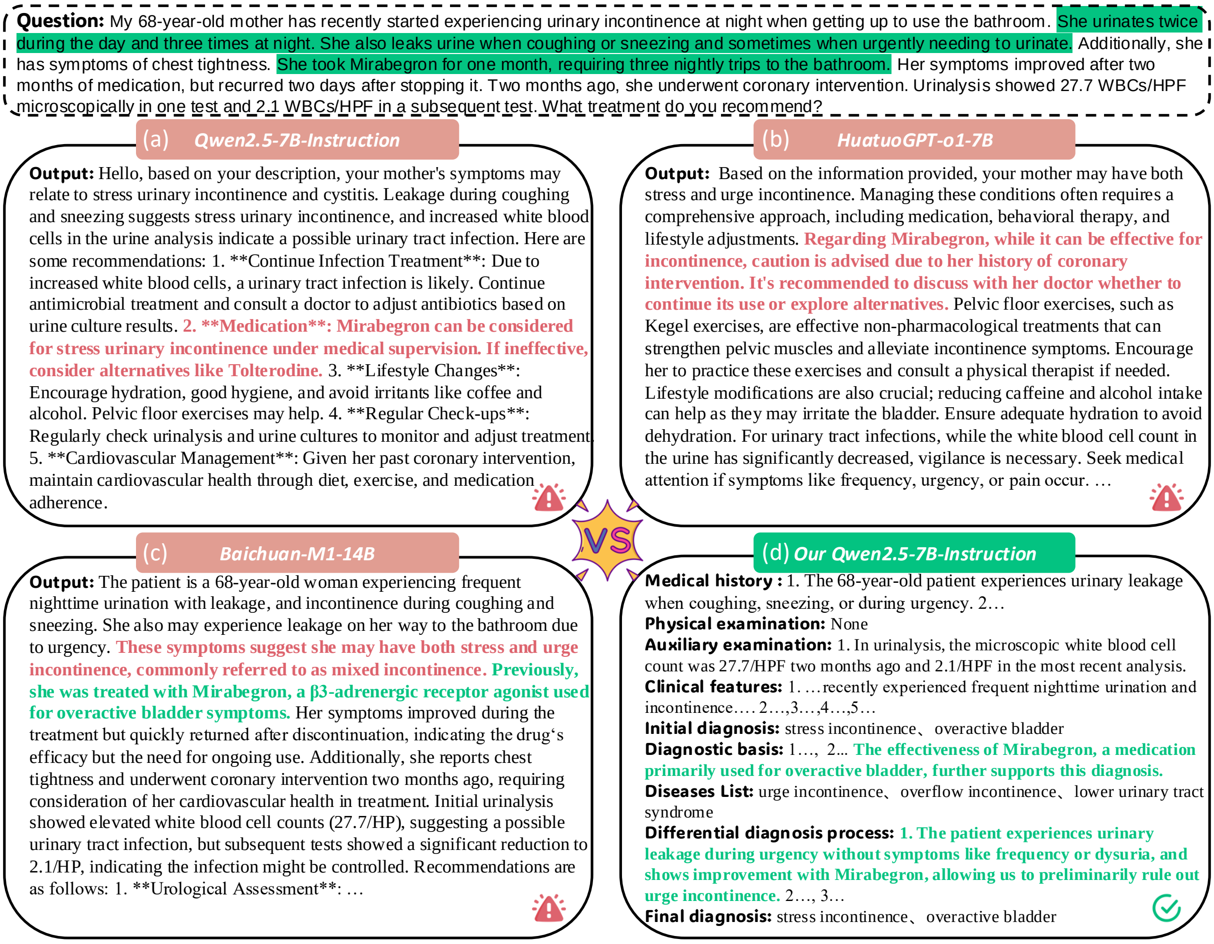

Figure 9: Case study on RJUA-QA. We examined a patient case requiring the diagnosis of two diseases, with key symptoms highlighted for emphasis. Panel (a) displays the zero-shot diagnostic result from the base model, Qwen2.5-7B-Instruction. Panel (b) shows the output from Huatuo-o1-7B, with its reasoning process omitted for brevity. Panel (c) presents the diagnostic result from Baichuan-M1-14B. Panel (d) illustrates the diagnostic outcome from the Qwen2.5-7B-Instruction model after fine-tuning with our high-quality clinical note data. Sections marked in red indicate errors in the model's responses, while those in green highlight areas where the model accurately used key symptoms to diagnose diseases.