Bringing Two-Turn Reasoning Into Pathological Diagnosis

Anonymous ACL submission

Abstract

In real-world pathology, diagnosis sometimes involves a two-stage reasoning process, an initial differential diagnosis with preliminary evidence, followed by a definitive diagnosis after further examinations. Existing research rarely reflects this workflow, treating diagnosis as a one-turn task. This work explicitly models the diagnostic process in pathology as a continuous two-turn dialogue with large language models (LLMs). To bridge the evidence gap between stages, we propose a Retrieval-Augmented Generation-based Examination Simulation (RAGES) method to simulate follow-up examination results requested in the first dialogue based on existing records and external knowledge. We curate a high-quality training dataset of initial and follow-up consultations and evaluate LLMs in the two-turn consultation across another multilingual dataset. Our experiments show that (1) LLMs significantly improve diagnostic accuracy with additional evidence, (2) our model outperforms or matches larger and reasoning-enhanced baselines, and (3) RAGES generates more plausible results than pure LLM generation.

1 Introduction

Multi-turn consultation is central to real-world medical diagnosis. Medical experts often begin with preliminary clinical evidence and iteratively refine their hypotheses through additional examinations and expert reasoning. This process, known as the hypothetico-deductive method, typically starts with a list of potential diseases, called the differential diagnosis, and converges to a definitive conclusion once sufficient further evidence is obtained, as illustrated in Fig. 1.

In pathology, this reasoning pattern is especially structured and sometimes manifests as a two-turn process. Initially, pathologists examine hematoxylin and eosin (H&E)-stained slides to assess tissue architecture and cellular features, combining these findings with clinical history to propose differential diagnoses and recommend further tests (e.g., immunohistochemistry, molecular tests, and whole-genome sequencing). After receiving the test results, they conduct a more detailed follow-up analysis and conclude with a final diagnosis.

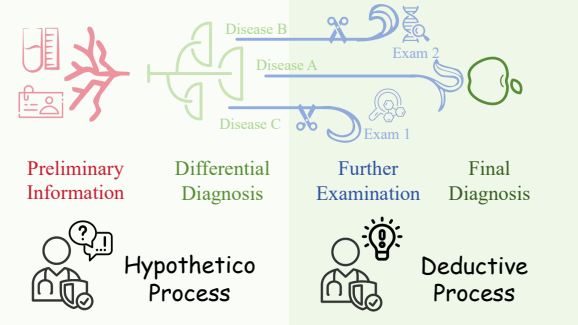

Figure 1: The hypothetico-deductive method in medical diagnosis. Medical experts first root a differential diagnosis in preliminary information. Later, they prune the branches, arriving at a final diagnosis with evidence from further examinations.

Despite its clinical importance, prior LLM-based diagnostic research has focused mainly on single-turn, multiple-choice tasks that assume access to complete information. This bypasses the core challenge of early-stage differential reasoning and undermines the authenticity of simulated diagnostic workflows. Moreover, evaluating open-ended differential diagnoses is inherently difficult due to their subjectivity. Even pathologists may derive different disease suspects and further testing items based on the same case. This variability further discourages exploration of this space.

To address these challenges, we explicitly model the diagnostic process in pathology as a two-turn reasoning workflow, i.e., generating a differential diagnosis in the initial consultation and refining it into a definitive diagnosis based on appended test results in the follow-up. To support this, we intro-

duce the Retrieval-Augmented Generation-based Examination Simulation (RAGES), which generates realistic follow-up evidence by combining the reuse of original records, retrieval from a curated knowledge base, and LLM-based generation.

Using RAGES, we curate high-quality supervised fine-tuning (SFT) data and train models capable of interactive diagnostic reasoning. We also propose automatic evaluation metrics for both diagnostic stages using powerful LLMs as judges. Our findings confirm that LLMs benefit significantly from additional clinical evidence, and our model outperforms or rivals larger and reasoning-enhanced baselines. Experiments show that RAGES can produce more plausible results than solely depending on LLMs' generation. This work brings a structured, two-turn reasoning workflow, which is closer to realistic pathological diagnosis, offering an early exploration of interactive AI diagnosis in pathology.

2 Related Work

2.1 Reasoning Capabilities of LLMs in Medical Diagnosis

Early studies showed that pretrained LLMs encode rich clinical knowledge and can answer medical questions effectively, e.g., Flan-PaLM (Singhal et al., 2023) and MedFound (Liu et al., 2025b). Prompting techniques such as chain-of-thought (CoT) have proven effective for inducing reasoning (Wei et al., 2023; Besta et al., 2024; Yao et al., 2023). In medicine, structured prompting enhances diagnostic accuracy (Nori et al., 2023; Savage et al., 2023; Kwon et al., 2024; Savage et al., 2024).

With the advent of OpenAI's o1 model (Jaech et al., 2024), the focus shifted to the native reasoning capability of LLMs on medical tasks. Nori et al. (2024) evaluated o1-preview on medical challenge problems and found it dramatically outperforms previous models with prompting. Sandmann et al. (2025) and Tordjman et al. (2025) both evaluated DeepSeek-R1, an open-source reasoning model, on medical tasks and clinical reasoning, demonstrating the potential of reasoning models. Building on this new paradigm, recent work has introduced medical LLMs and frameworks designed for stepwise reasoning. HuatuoGPT-o1 (Chen et al., 2024) is a medical LLM trained via verifiable reasoning steps, including exploiting complex reasoning trajectories and reinforcement learning with verifier-based rewards. Huang et al. (2025) focused on inference-time scaling of reasoning in the medical domain. With a learned process reward model, MedS^3 (Jiang et al., 2025a) learned to reason about medical problems.

2.2 Multi-Turn Diagnosis

Clinical diagnosis is inherently iterative, involving hypothesis formation, information gathering, and refinement. Several studies simulated multi-turn doctor–patient interactions (Bao et al., 2023; Chen et al., 2023; Li et al., 2023; Toma et al., 2023; Liu et al., 2025c), including systems like AMIE (Tu et al., 2024), AI Hospital (Fan et al., 2025), and MedAgentSim (Almansoori et al., 2025). APP (Zhu and Wu, 2025) explored a patient-centered multi-turn consultation approach to enable online consultations. MedAgentBench (Jiang et al., 2025b) and MMD-Eval (Liu et al., 2025a) provided realistic simulation environments grounded in structured patient data.

Other efforts focus on simulating the sequential diagnosis stages conducted by medical experts. Sun et al. (2024) observed that most LLM-based studies treat diagnosis as a one-shot question and answer with all information provided. They therefore proposed a two-planner system for differential diagnosis and final prediction. Likewise, MAC (Chen et al., 2025) explicitly models two consultation stages, the primary with limited data and the follow-up with complete data, and engages a multi-disciplinary treatment simulation. However, few studies address continuous multi-turn reasoning with additive evidence in a single-session LLM setting.

3 Method

3.1 Two-Turn Reasoning Workflow in Pathological Diagnosis

When pathologists make a diagnosis, they begin by examining H&E-stained slides under a microscope, which highlight cell structures. Combining these microscopic observations with the patient's clinical history, they establish a preliminary differential diagnosis consisting of several likely diseases. Based on this differential, pathologists then order targeted tests, such as immunohistochemistry for protein markers, or molecular and genetic analyses, to distinguish between diseases with similar morphological features. Once the test results are available, they follow a deductive reasoning process to arrive at a final diagnosis, provided there is

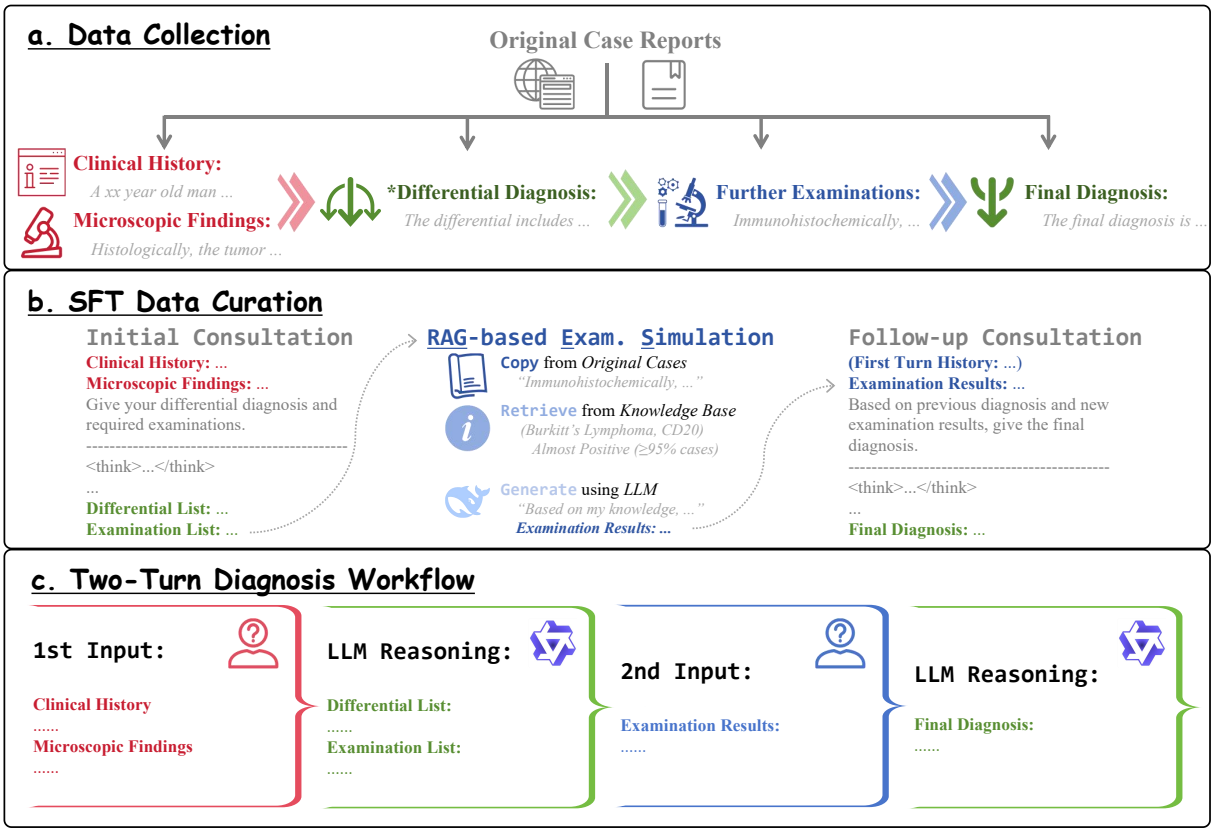

Figure 2: An overview of this work. (a) We collect raw case reports from open-source websites and journals and arrange them into a two-turn consultation form. (b) We create two-turn supervised fine-tuning data with examination results generated by the proposed RAG-based examination simulation strategy. (c) The proposed two-turn reasoning workflow of diagnosis.

sufficient supporting evidence.

As illustrated in Fig. 2 (c), we model this diagnostic workflow as a two-turn interactive process with LLMs. In the initial turn, we provide the patient's clinical history and findings from the H&E slides to LLMs, prompting them to generate a reasoning process that includes candidate diseases and recommended further examinations. In the follow-up turn, we supply the LLMs with the newly acquired test results and prompt it to deliver a final diagnosis.

3.2 Data Collection

Detailed case reports form the foundation of the proposed two-turn diagnostic process. However, datasets containing pathology-specific cases remain scarce. To address this gap, we curated cases from publicly available sources, including websites and academic journals. Due to data usage restrictions, some of these cases can only be used for evaluation purposes. We will discuss them later in Section 4.2. This section focuses on the data sources used during the SFT stage.

- DakaPath[1] is a Chinese platform for pathological teaching and communication. Besides the plentiful knowledge, DakaPath has a special section called Micro Lecture, which provides expert explanations over hundreds of real cases. We collected 373 raw explanations.
- Chinese Journal of Pathology reports on advanced scientific research achievements and pathological diagnosis experience as case discussions. Originally, we filtered out 653 cases.

While original case reports describe the complete diagnostic workflow, they typically lack explicit stage boundaries. Additionally, many include follow-up discussions that pertain to the post-diagnosis stage. To construct data suitable for interaction with LLMs, we employ powerful LLMs to extract key information from the original cases while simultaneously filtering out unsuitable examples.

As shown in Fig. 2 (a), we prompt LLMs to explicitly extract five components: clinical history,

[1] https://www.dakapath.com

microscopic findings, differential diagnosis, further examinations and their results, and the final diagnosis. The first two components are used as input for the first consultation. Although the differential diagnosis is informative, we exclude it at this stage to avoid constraining the LLMs' reasoning in the first turn. The examination results serve as the most dependable source for constructing the second-turn input, which will be described in the next section. The final diagnosis is treated as ground truth and used to validate the generated SFT data. We use GPT-4 for automatic information extraction and manually verify the quality of each output.

Algorithm 1: RAGES

Input: Case report $\mathcal{C}$, requested exams E, structured knowledge base $\mathcal{K}$
Output: Simulated examination outputs $\mathcal{E}$

```
# Split original case
```
$(\mathcal{E}_{\text{gt}}, \mathcal{D}_{\text{gt}}) \leftarrow \text{PreSplit}(\mathcal{C})$;
```
# Reuse overlapping real results
```
$\mathcal{E}_{\text{direct}} \leftarrow \text{MatchOverlap}(\mathcal{E}_{\text{gt}}, E)$;
```
# Retrieve disease-exam mappings
```
$\text{Candidates} \leftarrow \text{EmbedAndSearch}(\mathcal{K}, \mathcal{D}_{\text{gt}})$;
$\text{BestMatch} \leftarrow \text{SelectHighestSimilarity}(\text{Candidates})$;
$\mathcal{E}_{\text{retrieved}} \leftarrow \text{GetMappings}(\mathcal{K}, \text{BestMatch})$;
```
# Generate final results via LLM
```
$\mathcal{E}_{\text{gen}} \leftarrow \text{LLMGenerate}(E, \mathcal{E}_{\text{direct}}, \mathcal{E}_{\text{retrieved}}, \mathcal{D}_{\text{gt}})$;
$\mathcal{E} \leftarrow \mathcal{E}_{\text{gen}}$;
return $\mathcal{E}$

RAGES Prompt

Based on the given information, after careful consideration, infer the possible result of each examination item. The given information includes the final diagnosis, examination items, existing results (if any), and relevant knowledge (if any). Specifically, you need to:

0. Only focus on the content that can produce definitive results.
1. First, check the "Existing Results" and record results that overlap with the examination.
2. Then, check the "Relevant Knowledge". First, determine whether the relevant knowledge pertains to the same disease as described in the "Final Diagnosis". If it is the same disease, then, based on this knowledge, infer the results of the remaining examination items.
3. Retrieve your own knowledge and speculate on the results of the remaining items.
4. Output the above results in the specified format. The format is as follows:
ExamRes: {"Item 1": ("Result 1", Confidence Level 1), "Item 2": ("Result 2", Confidence Level 2)}

Figure 3: An illustration of RAGES prompt.

3.3 RAG-based Examination Simulation

Before initiating the simulation of two-turn diagnosis, we need to design a strategy to obtain the required examination results proposed by the response in the initial consultation. In real scenarios, these results can be acquired from pathologists' practice to facilitate a multi-turn interaction. It is, however, less practical to consult the laboratory during SFT. Therefore, we propose the retrieval-augmented generation-based examination simulation (RAGES), a method designed to produce plausible yet grounded examination outcomes without real-time laboratory access. RAGES operates in three key stages, as in Algorithm 1.

Reuse of Existing Results. For each case, we first separate the original case report and match the examinations ordered by the model with those already performed in the record. If overlap exists, the corresponding results are directly reused with complete confidence, as they originate from verified laboratory data. However, considering the probable difference between the real and generated differential list, this stage typically contributes to only a proportion of the results requested by the LLMs.

Retrieval from Structured Knowledge. We retrieve results from a curated database comprising over 24,000 mappings between 1,629 diseases and 465 immunohistochemistry (IHC) markers to cover examinations not present in the case report. Each mapping captures statistical associations indicating the likelihood of a test result given a disease, e.g., "almost positive, with $\geq$95% positive cases". We embed the final diagnosis from the case using a sentence transformer and search for similar disease entries in the database. The closest match is selected, and its mappings are retrieved to augment the following result generation process. These results are statistically grounded but may be ambiguous in rare or conflicting cases.

LLM-based Generation. We query a powerful LLM with the case context and retrieved knowledge to generate plausible outputs for any examinations not covered in prior steps. An example of the prompt is shown in Fig. 3. These outputs may draw upon the LLM's internal knowledge or reasoning capabilities. However, since they lack direct empirical backing, their reliability is lower than that of reused or retrieved results. Notably, in practice, we randomly drop some results with extremely low confidence and tell the models in the second turn that these examinations are unavailable.

This design prioritizes high-quality, explainable and complete results, reducing noise from hallucinations or overconfident LLM generations.

3.4 Activating Two-Turn Reasoning Capability

Following Huang et al. (2025), we create an SFT dataset generated by powerful reasoning models to activate existing LLMs' reasoning on two-turn diagnosis. The whole process is illustrated in Fig. 2 (b).

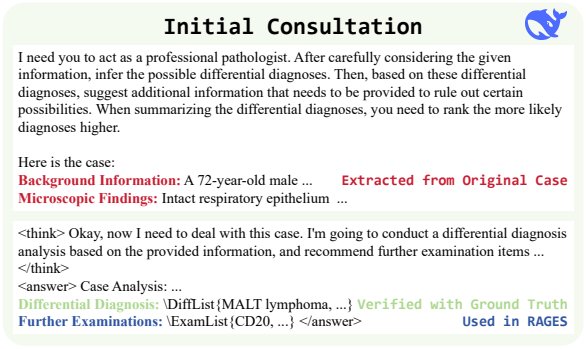

Figure 4: The SFT data in the initial consultation turn.

In the **initial consultation**, LLMs take clinical history and histological findings as input and think about possible differential diagnoses and appended examinations. Instead of directly giving the differential list in the original cases as a guide, we adopt a post-verification strategy. We only provide the input and generate it several times for each case. After gathering these trials, we use another LLM to judge whether the ground truth diagnosis appears in these trials, and retain the positive ones as SFT data. We choose this more complex strategy for two reasons. First, we do not expect LLMs to be bound by the original text, since different pathologists might derive different disease suspects from the same case. Second, when a differential list is provided, LLMs might have hallucinations like direct references to the original results. A shortened sample of training data in the first consultation is illustrated in Fig. 4.

Based on the verified trials, we use the RAGES method to simulate the appended results. After collecting sufficient appended results, we can start the **follow-up consultation**. We provide LLMs with the response in the first turn and the acquired further examination results, and ask them to produce a final diagnosis after careful thinking. Also, we employ the post-verification strategy and retain as SFT data those trajectories that propose the true diagnosis. A shortened sample of training data in the second consultation is illustrated in Fig. 5. Notably, we explicitly provide the history only in the data-generating stage. In the SFT and evaluation stage, we do not repeatedly give the existing information.

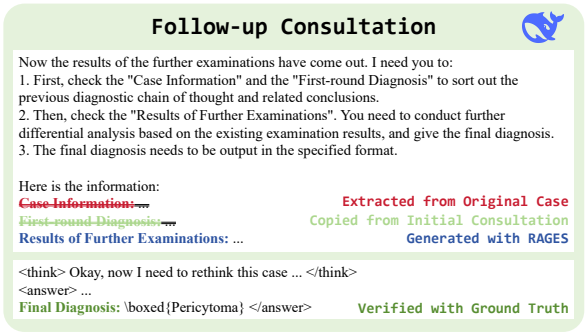

Figure 5: The SFT data in the follow-up turn. Notably, the case information and history of the initial consultation are only provided during the data-generating process and are not exposed to LLMs in dialogue.

4 Experiment

4.1 Implementation

We use DeepSeek-R1 (DeepSeek-AI, 2025) to perform RAGES and generate SFT data, because R1 offers a transparent reasoning process and consistently strong performance. A total of 925 training samples related to initial consultations and 623 follow-up samples are constructed. For SFT, we adopt Qwen2.5-32B-Instruct (Yang et al., 2024) as the base model. We apply parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2022) and enable bf16 precision to optimize training with our curated dataset. The training workflow is implemented using Llama-Factory (Zheng et al., 2024), and evaluation is conducted with vLLM (Kwon et al., 2023). We use LoRA with default hyperparameters, as $\alpha = 16$ and $r = 8$. The initial learning rate is set to 5×10^{-5} with cosine decay, and training is run for 15 epochs. The fine-tuning is carried out on 8 A100 GPUs, and the entire process completes in approximately 12 hours.

4.2 Evaluation Configurations

4.2.1 Evaluation Datasets

To ensure fair comparisons, we collect evaluation data from other sources, including publicly available English-language cases from Pathology Outlines and the Hans Popper Hepatopathology Society (HPHS), as well as in-house Chinese-language cases used for resident training at Hospital X. **Pathology Outlines**[2] is a comprehensive platform for pathology communication. It offers extensive knowledge across subspecialties and weekly case studies. We collected 483 cases published before

[2] https://www.pathologyoutlines.com

December 2024. **Hans Popper Hepatopathology Society (HPHS)**[3] is a hepatopathology-focused community that presents a noteworthy case every 3 to 4 months. We included 37 cases published up to September 2024. The **in-house dataset** originates from internal examination materials used for resident education at Hospital X. It includes 339 cases covering 11 major domains, including the endocrine system, reproductive system, respiratory system, central nervous system, skin, bone and soft tissue, etc.

We manually extract relevant diagnostic information from case reports to ensure dataset quality and compile two evaluation subsets.

1. The **Public English Dataset (EN)** consists of 110 English-language cases, 100 from Pathology Outlines and 10 from HPHS. Due to distribution constraints, we open-source only the URLs of the cases in this dataset.

2. The **In-house Chinese Dataset (CN)** comprises 276 Chinese-language cases sourced from Hospital X.

Since our model is trained exclusively on Chinese data, the English dataset is a relatively unbiased and open-source benchmark. However, it primarily features complex and atypical cases, which are less representative of routine diagnostic scenarios. Therefore, we also evaluate on the Chinese dataset to simulate a more realistic yet sufficiently challenging clinical setting.

4.2.2 Evaluation Metrics

For the initial consultation stage (*Initial*), where the model proposes potential disease candidates, we evaluate whether the ground truth final diagnosis appears in the differential list. If so, it is considered a hit, and the hit rate is used to quantify the differential accuracy (*DiffAcc*). For the follow-up consultation stage (*Follow-up*), where the model refines its decision and outputs a precise diagnosis, we check whether the ground truth diagnosis appears as the top-ranked candidate. In this case, we use the hit-at-one rate to measure the diagnostic accuracy (*DxAcc*). Unless otherwise specified, *DiffAcc* is reported for the initial consultation turn, and *DxAcc* for the follow-up turn.

To ensure a comprehensive and objective evaluation, we rely on three strong LLMs, GPT-4o (*4o*) (OpenAI et al., 2024), DeepSeek-R1 (*R1*), and Qwen2.5-Max (*QM*), to assist in verification. We also report the average score across these three models for overall performance.

[3] https://hanspopperhepatopathologysociety.org

5 Results Analysis

5.1 More Information, More Accurate Diagnosis

Before conducting a comprehensive comparison of two-turn diagnostic performance, we aim to demonstrate that LLMs produce more accurate diagnoses when provided with additional evidence. To this end, we compare diagnostic accuracy after the initial consultation and again after the follow-up stage, evaluating whether access to more information improves the precision of the model's predictions.

In addition to our SFT model (*Ours-32B*), we evaluate three baselines: the original Qwen2.5-32B-Instruct (*Qwen-32B*), a reasoning variant, *QwQ-32B* (QwenTeam, 2025), and a larger Qwen2.5-72B-Instruct (*Qwen-72B*).

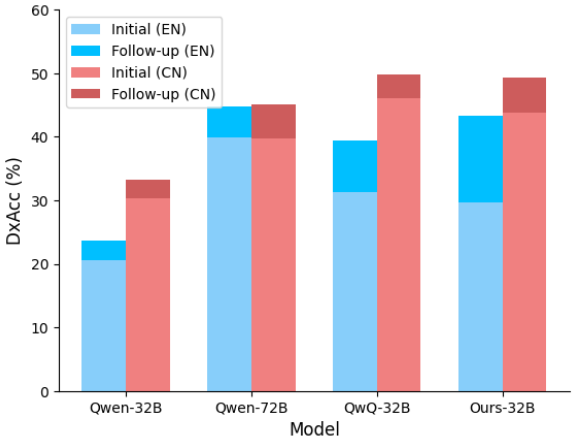

Figure 6: The comparison of diagnosis accuracy in the initial and follow-up consultation.

As shown in Fig. 6, all models achieve higher diagnostic accuracy across both datasets when given follow-up information. While the outcome may seem intuitive, this experiment provides a critical foundation for our study, validating the necessity of incremental evidence in LLM-assisted diagnosis and motivating the subsequent analyses.

5.2 Performances in Two-Turn Diagnosis

We compare our model with the three baseline models introduced earlier. Tables 1 and 2 present the overall performance across both the initial and follow-up consultation stages.

During the initial consultation, our model outperforms the original Qwen-32B, with average accu-

Table 1: The differential accuracy (DiffAcc) of different models in the initial consultation (Yellow: chat models; Red: reasoning models). **Bold** figures suggest the best performance, and the underlined are the second best.

Model	Public English Dataset				In-House Chinese Dataset				Overall
	4o	R1	QM	Avg.	4o	R1	QM	Avg.	
Qwen-32B	42.7	32.7	43.6	39.7	51.8	42.4	57.7	49.6	46.8
Qwen-72B	59.1	47.3	53.6	53.3	66.7	58.3	67.8	64.3	61.2
QwQ-32B	60.0	**48.2**	**61.8**	**56.7**	69.9	58.3	68.5	65.6	63.1
Ours-32B	**62.7**	46.4	57.3	55.5	**71.7**	**58.7**	**69.6**	**66.7**	**63.5**

Table 2: The diagnosis accuracy (DxAcc) of different models in the follow-up consultation (Yellow: chat models; Red: reasoning models). **Bold** figures suggest the best performance, and the underlined are the second best.

Model	Public English Dataset				In-House Chinese Dataset				Overall
	4o	R1	QM	Avg.	4o	R1	QM	Avg.	
Qwen-32B	23.6	16.4	30.9	23.6	38.0	23.9	37.7	33.2	30.5
Qwen-72B	**50.0**	33.6	**50.9**	**44.8**	49.3	35.9	50.0	45.1	45.0
QwQ-32B	40.9	34.5	42.7	39.4	**53.6**	**43.8**	52.2	**49.9**	46.9
Ours-32B	46.4	**38.2**	45.5	43.4	52.2	42.0	**53.6**	49.3	**47.6**

racy improvements of 17.1% on the Chinese dataset and 14.2% on the English dataset. In the follow-up stage, the gains remain substantial at 16.1% and 19.8%, respectively. When compared with the reasoning-augmented QwQ-32B and the larger Qwen-72B, our model demonstrates comparable or even superior performance. Considering the overall performance of both datasets, our approach achieves the best overall results among all the models evaluated. These findings highlight the effectiveness of reasoning in enhancing differential and diagnostic accuracy across multilingual settings.

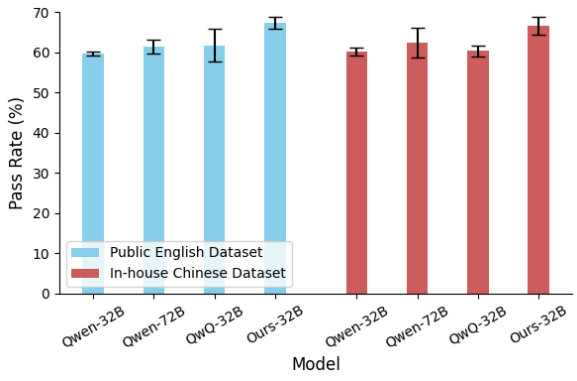

Figure 7: The plausibility of examinations requested by different models.

5.3 The Plausibility of Examinations

To assess the plausibility of the follow-up examinations suggested by each model during the initial consultation stage, we employ GPT-4o as an external evaluator. Specifically, GPT-4o is asked to judge whether the proposed examination items are appropriate, given the differential diagnosis list generated by the model. To ensure stability and reduce variance, we repeat the evaluation three times for each response and compute the model's pass rate as the final metric, i.e., the proportion of times the suggestions are deemed plausible by GPT-4o.

The results are shown in Fig. 7. As illustrated, our model exhibits a clear advantage in plausibility, which can be attributed to the carefully designed SFT. However, overall pass rates remain modest, primarily due to the inclusion of redundant examinations aimed at verifying elements of the differential diagnosis.

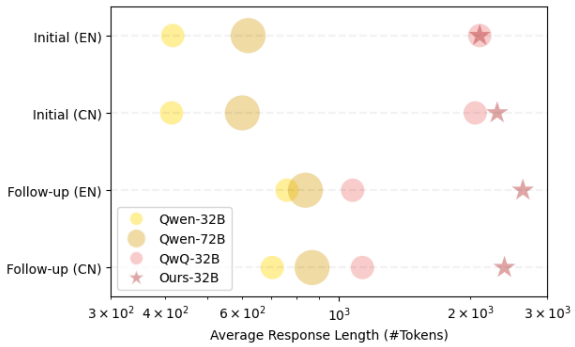

Figure 8: The average response lengths of 4 models in different stages.

5.4 Response Length

Following prior work (Huang et al., 2025), we examine the response lengths of different models across the two consultation stages (Fig. 8). In the first consultation, reasoning-enhanced models (QwQ-32B and ours) output around 2,000 tokens, while the larger Qwen-72B generates about 600 tokens and baseline Qwen-32B about 400 tokens. In the second consultation, base models (Qwen-32B and Qwen-72B) become more verbose, while QwQ-32B shortens, and our model maintains a similar length. This reflects two opposing but reasonable factors. Integrating new findings may increase length, while ruling out differentials is simpler than proposing them. Notably, the original models consistently generate responses of about 1,000 tokens, while ours maintains around 2,000. This may suggest a tendency toward over-reasoning induced by SFT.

5.5 Ablation Study

The ablation study focuses on two key aspects: (1) validating the effectiveness of incorporating two-turn diagnostic data during supervised fine-tuning (SFT data ablation), and (2) evaluating the necessity of the RAGES components (RAGES ablation).

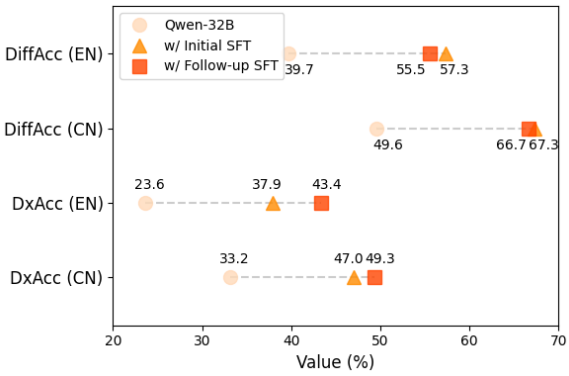

Figure 9: Ablation study on the usage of SFT data. *w/ Initial SFT* suggests only incorporating the data about initial consultation, and *w/ Follow-up SFT* further includes data about follow-up consultation.

SFT data ablation. We compare Qwen-32B with two SFT variants: one trained on initial consultation data only, and another on both consultation stages. The results are shown in Fig. 9. Even limited to the first-turn data, SFT significantly boosts performance in both stages. This improvement may stem from the model generating a more accurate differential diagnosis in the initial turn, which inherently facilitates more precise final diagnoses. Adding follow-up data slightly reduces differential accuracy but notably improves final diagnosis accuracy, highlighting the value of learning to reason with appended evidence.

Table 3: Ablation study on different phases of RAGES. *w/ GT* suggests reusing the original results, and *w/ KB* retrieves from the knowledge base.

RAGES		Correctness (%)		
w/ GT	w/ KB	EN	CN	Overall
		82.7	79.2	80.2
✓		**86.4**	80.3	82.0
	✓	84.5	80.7	81.8
✓	✓	**86.4**	**84.1**	**84.8**

RAGES ablation. As in Section 5.3, we also employ GPT-4o to assess the correctness of simulated examination results under four RAGES settings: vanilla generation, with reused text, with retrieved knowledge, and the full combination. The results are presented in Table 3. Including either reused or retrieved information improves output quality, while combining both yields the highest correctness, confirming their complementary value for factual and plausible result synthesis.

6 Conclusion

We present a two-turn reasoning workflow to simulate and evaluate the full hypothetico-deductive diagnostic process in pathology using large language models. We enable fine-grained supervision and evaluation by formalizing diagnosis as a two-turn task and introducing the RAGES method for follow-up examination simulation. Our experiments confirm the importance of evidence acquisition in LLM-assisted diagnosis and highlight the benefits of reasoning. Our model achieves superior or comparable performance to larger or reasoning-enhanced models, while also generating more plausible diagnostic reasoning and examination suggestions. This work lays a foundation for future works, including (1) incorporating more turns of evidence gathering and differential refinement, (2) combining comprehension of pathological images into the workflow, thus formulating a multimodal framework, and (3) including human-in-the-loop assessments for a more reliable model development and evaluation.

Limitations

This work represents an initial exploration of integrating two-turn reasoning into pathological diagnosis using LLMs. As a foundational step, our approach focuses exclusively on text-based reasoning, without incorporating the multi-modal interaction between pathology images and textual findings. Additionally, for clarity and evaluation feasibility, we model diagnosis as a two-turn process, which simplifies the inherently multi-turn and iterative nature of real-world clinical reasoning. For example, in real-world diagnosis, pathologists typically first order IHC tests, followed by molecular tests, and finally whole-genome sequencing, while in this work, the three kinds of tests are combined as a one-turn request in the follow-up consultation.

Regarding methodology, the limitations of this study fall into two main aspects. First, our approach relies solely on supervised fine-tuning (SFT) to activate the model's reasoning capabilities. While SFT is efficient and practical, reinforcement learning (RL) offers an alternative avenue for encouraging native reasoning. However, RL introduces significant challenges. Designing meaningful reward functions for diagnostic tasks and implementing robust training frameworks for flexible multi-turn dialogue remains non-trivial, especially within existing infrastructures such as OpenRLHF or VeRL. Second, our evaluation framework depends on LLM-based judgments rather than human experts. Although we mitigate this limitation by incorporating multiple strong LLMs (e.g., GPT-4o, DeepSeek-R1, Qwen-Max) and averaging results across them, automated evaluation still lacks the clinical authority and nuanced judgment that expert pathologists provide. Future work will include human-in-the-loop evaluation to ensure real-world applicability and safety.

Ethics Statement

This work leverages pathological case studies from multiple sources, raising two primary ethical considerations, i.e., patient privacy and data distribution.

Regarding patient privacy, all case reports collected from public websites and journals were already anonymized at the source. We carefully removed any personally identifiable information from the in-house dataset, preserving only essential information such as age and gender for clinical reasoning.

In terms of data distribution, we will not release the in-house dataset publicly due to institutional data protection policies. For externally sourced cases, we strictly adhered to the usage guidelines specified by each website or journal. To avoid unauthorized redistribution, we will release only the URLs linking to the original case sources, allowing other researchers to access the materials while respecting the original data ownership and licensing terms.

The risk of this work may lie in the improper responses (including repetitive patterns, false information, malignant output, etc.) since we do not specifically strengthen the safety of the model.

References

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

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

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: solving elaborate problems with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

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

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *Preprint*, arXiv:2310.15896.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025a. Meds^3: Towards medical small language models with self-evolved slow thinking. *Preprint*, arXiv:2501.12051.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025b. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *Preprint*, arXiv:2501.14654.

Taeyoon Kwon, Kai Tzu iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Yongsik Sim, Beomseok Sohn, Dongha Lee, and Jinyoung Yeo. 2024. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. *Preprint*, arXiv:2312.07399.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025a. Interactive evaluation for medical LLMs via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Zijie Liu, Xinyu Zhao, Jie Peng, Zhuangdi Zhu, Qingyu Chen, Xia Hu, and Tianlong Chen. 2025c. Dialogue is better than monologue: Instructing medical llms via strategical conversations. *Preprint*, arXiv:2501.17860.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Preprint*, arXiv:2311.16452.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. 2024. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *Preprint*, arXiv:2411.03590.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

QwenTeam. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. 2025. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2023. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Preprint*, arXiv:2308.06834.

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.

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.

Zhoujian Sun, Cheng Luo, Ziyi Liu, and Zhengxing Huang. 2024. Conversational disease diagnosis via external planner-controlled large language models. *arXiv preprint arXiv:2404.04292*.

Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *Preprint*, arXiv:2305.12031.

Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, and 1 others. 2025. Comparative benchmarking of the deepseek large language model on medical tasks and clinical reasoning. *Nature Medicine*, pages 1–1.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *Preprint*, arXiv:2502.07143.

A Appendix

A.1 Full Prompts

We append the full prompts used in SFT data, two-turn consultation, and automated evaluation for reproducibility.

Prompts used in SFT data. The whole prompt includes a system prompt to explicitly separate thinking and answering, a detailed instruction in the initial consultation, and a follow-up instruction, as illustrated in Fig. 10.

Prompts used in two-turn workflow. We use a simplified prompt for the original models during the two-turn workflow, as illustrated in Fig. 11. The difference is two-part. We drop the system prompt to avoid disturbing its original one, and we use a prompt without step-by-step instructions in the initial consultation since we find some interesting outcomes introduced later in Appendix A.2.

Prompts used in automated evaluation. We use LLMs to evaluate the differential and diagnostic accuracy, as well as the examinations and results. The prompts used in evaluating are shown in Fig. 12.

A.2 Ablation on Prompting

As stated before, we compare the performance of the original Qwen models with detailed instructions, as used in our model, to investigate the influence of prompting (Prompt ablation).

Prompt ablation. Figure 13 presents the differences. For a relatively weak model (Qwen-32B), a detailed step-by-step instruction can bring dramatic improvement, though not comparable with stronger models like Qwen-72B (larger), QwQ-32B (reasoning improved), and ours. For strong models, detailed instruction seems to limit their ability to think and might cause a decline in performance.

SFT Data Prompts

System

You are Qwen, created by Alibaba Cloud. You are a helpful assistant. A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>.

User

I need you to act as a professional pathologist. After carefully considering the given information, infer the possible differential diagnoses. Then, based on these differential diagnoses, suggest additional information that needs to be provided to rule out certain possibilities. Specifically:
1. First, you need to carefully analyze the given information, which mainly includes case background information, previous examination items, morphological descriptions of pathological sections, etc. Summarize the evidence points related to the diagnosis from this information.
2. Based on the given information, analyze what the possible differential diagnoses are and determine whether they are consistent with the given information. Note: These differential diagnoses should be as broad and accurate as possible (broad means considering less common diagnostic possibilities, and accurate means the listed differential diagnoses should not conflict with most of the background information).
3. According to the listed differential diagnoses, propose the further examination items. You need to specify the exact antigen - antibody, staining type, or molecular type. If the existing information is sufficient to confirm a specific disease, only output that disease and leave the additional examination items blank.
4. Finally, summarize the possible differential diagnoses and the required additional examination items in a given format. When summarizing the differential diagnoses, you need to rank the more likely diagnoses higher.
5. Output four sections in the specified format: "Case Analysis - Differential Diagnosis Analysis - Additional Examination Items - Summary".

Format requirements:
Case Analysis
...
Summary
Differential Diagnoses: \DiffList{Differential Diagnosis 1, ...}
Further Examinations: \ExamList{Examination Item 1, ...}

Here is the case:
<Case Information>

Assi.

<think> Okay, now I need to deal with this case. I'm going to conduct a differential diagnosis analysis based on the provided information, and recommend further examination items ... </think>
<answer> ...
Differential Diagnosis: \DiffList{...}
Further Examinations: \ExamList{...} </answer>

User

Now the results of the further examinations have come out. I need you to:
1. First, check the "Case Information" and the "First-round Diagnosis" to sort out the previous diagnostic chain of thought and related conclusions.
2. Then, check the "Results of Further Examinations". The additional test results may not fully match the items requested in the initial diagnosis. Based on the available test results, you need to conduct further differential analysis, and give the final diagnosis. Note: You are completely entitled to overturn the initial diagnostic approach and provide a diagnosis based on the current information after obtaining more data.
3. The final diagnosis must be output in the specified format, i.e., \boxed{Diagnosis Name}
Here is the information:
Results of Further Examinations: <Exam Results>

Assi.

<think> Okay, now I need to rethink this case ... </think>
<answer> ...
Final Diagnosis: \boxed{...} </answer>

Figure 10: A full illustration of SFT data.

Workflow Prompts

User

I need you to act as a professional pathologist. After carefully considering the given information, infer the possible differential diagnoses. Then, based on these differential diagnoses, suggest additional information that needs to be provided to rule out certain possibilities. Note: These differential diagnoses should be as broad and accurate as possible (broad means considering less common diagnostic possibilities, and accurate means the listed differential diagnoses should not conflict with most of the background information). You should summarize the possible differential diagnoses and the required additional examination items in a given format. When summarizing the differential diagnoses, you need to rank the more likely diagnoses higher.
Format requirements:
Case Analysis
...
Summary
Differential Diagnoses: \DiffList{Differential Diagnosis 1, ...}
Further Examinations: \ExamList{Examination Item 1, ...}

Here is the case:
<Case Information>

Assi.

...

Differential Diagnosis: \DiffList{...}
Further Examinations: \ExamList{...}

User

Now the results of the further examinations have come out. I need you to:
1. First, check the "Case Information" and the "First-round Diagnosis" to sort out the previous diagnostic chain of thought and related conclusions.
2. Then, check the "Results of Further Examinations". The additional test results may not fully match the items requested in the initial diagnosis. Based on the available test results, you need to conduct further differential analysis, and give the final diagnosis. Note: You are completely entitled to overturn the initial diagnostic approach and provide a diagnosis based on the current information after obtaining more data.
3. The final diagnosis must be output in the specified format, i.e., \boxed{Diagnosis Name}
Here is the information:
Results of Further Examinations: <Exam Results>

Assi.

<think> Okay, now I need to rethink this case ... </think>
<answer> ...
Final Diagnosis: \boxed{...} </answer>

Figure 11: A full illustration of prompts in the workflow.

Evaluation Prompts

User

I need you to act as a professional pathologist. After careful consideration based on the given disease candidates and the true diagnosis, determine whether the true diagnosis (or a close approximation) is among the candidates and, if present, its position in the list. If it is within the candidates, output \boxed{True} + "Hit candidate content" + "Position of the hit content" at the end; otherwise, output \boxed{False} + No hit + 0.

Assi.

...

\boxed{True | False} + ... + <digital>

User

I need you to assist me in determining whether some pathological content is reasonable. I will provide you with a list of differential diagnosis diseases, a set of further examination results, and the ground truth diagnosis. You need to determine:
1. Based on the list of differential diagnosis diseases, judge whether the additional examination items are reasonable and record the unreasonable items;
2. Based on the ground truth diagnosis, judge whether the further examination results are reasonable and record the incorrect results.

Notes:
1. When the additional examination items are "no need," both items can be directly considered reasonable.
2. When judging the plausibility of examination results, do not consider whether some results are omitted; only judge the reasonableness of the existing examination results.

The information you need to use is as follows:
- Differential diagnosis: ...
- Further examinations and results: ...
- Ground truth diagnosis: ...

After careful consideration, you need to summarize at the end of the output in the following format:
1. Exam: \boxed{True|False}, \List{Wrong Item 1, ...}
2. Result: \boxed{True|False}, \List{Wrong Item and Result 1, ...}

Assi.

...

1. Exam: \boxed{True | False}, \List {...}
2. Result: \boxed{True | False}, \List {...}

Figure 12: A full illustration of prompts for evaluation.

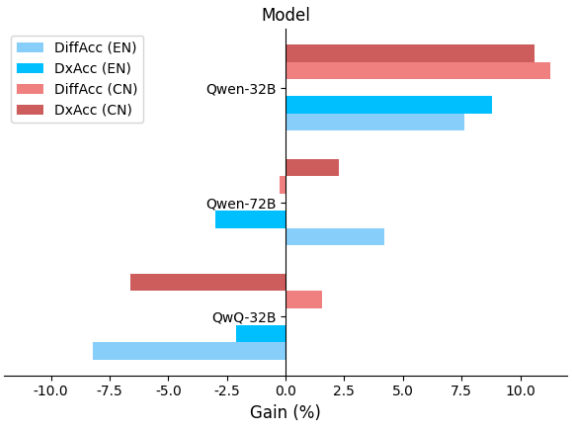

Figure 13: Performance gains with a detailed instructing prompt on original Qwen models.