DiagnosisArena: Benchmarking Diagnostic Reasoning for Large Language Models

Yakun Zhu[1,2,3*] **Zhongzhen Huang**[1,3*] **Linjie Mu**[1,3*] **Yutong Huang**[1]
Wei Nie[5] **Jiaji Liu**[6] **Shaoting Zhang**[1†] **Pengfei Liu**[1,2,4†] **Xiaofan Zhang**[1,2,3†]
[1]Shanghai Jiao Tong University , [2]SII, [3]SPIRAL Lab, [4]Generative AI Research Lab (GAIR)
[5]Shanghai Chest Hospital, [6]Beijing Anzhen Hospital, Capital Medical University

Abstract

The emergence of groundbreaking large language models capable of performing complex reasoning tasks holds significant promise for addressing various scientific challenges, including those arising in complex clinical scenarios. To enable their safe and effective deployment in real-world healthcare settings, it is urgently necessary to benchmark the diagnostic capabilities of current models systematically. Given the limitations of existing medical benchmarks in evaluating advanced diagnostic reasoning, we present *DiagnosisArena*, a comprehensive and challenging benchmark designed to rigorously assess professional-level diagnostic competence. *DiagnosisArena* consists of 1,113 pairs of segmented patient cases and corresponding diagnoses, spanning 28 medical specialties, deriving from clinical case reports published in 10 top-tier medical journals. The benchmark is developed through a meticulous construction pipeline, involving multiple rounds of screening and review by both AI systems and human experts, with thorough checks conducted to prevent data leakage. Our study reveals that even the most advanced reasoning models, o3-mini, o1, and DeepSeek-R1, achieve only 45.82%, 31.09%, and 17.79% accuracy, respectively. This finding highlights a significant generalization bottleneck in current large language models when faced with clinical diagnostic reasoning challenges. Through *DiagnosisArena*, we aim to drive further advancements in AI's diagnostic reasoning capabilities, enabling more effective solutions for real-world clinical diagnostic challenges. We openly share the benchmark and evaluation tools for further research and development[1].

1 Introduction

Recent advances in the reasoning capabilities of large language models (LLMs) have transformed the landscape for addressing complex scientific problems, such as mathematics and programming (OpenAI, 2024b; DeepSeek-AI, 2025). Using step-by-step problem-solving and iterative refinement processes, LLMs have demonstrated performance that exceeds humans in diverse tasks (Qin et al., 2024; Huang et al., 2024; Guan et al., 2025; Chen et al., 2025). Emerging studies have increasingly highlighted the promising potential of integrating these reasoning capabilities into clinical scenarios (Nori et al., 2024; Huang et al., 2025; Chen et al., 2024b; Yu et al., 2025; Jiang et al., 2025). However, there remains a considerable gap in assessing the readiness of such models for real-world clinical deployment. Consequently, how to benchmark the clinical diagnostic capabilities of reasoning models has become a focal point for further application.

Existing benchmarks primarily use medical examination-style questions to evaluate the capabilities of LLMs in the medical domain (Nori et al., 2024). Such tasks are mainly reliant on specific knowledge and have become relatively straightforward for contemporary LLMs. In fact, state-of-the-art models now achieve precision exceeding 90% on established benchmarks, including MMLU (Hendrycks et al., 2020) and MedQA (Jin et al., 2021). The saturation of these existing benchmarks impedes an accurate evaluation of LLM capabilities in realistic diagnostic scenarios. Although recent efforts have attempted to construct benchmarks based on real-world clinical cases (Chen et al., 2024a; Wang et al., 2023b; zhao zy15, 2024), these evaluations often suffer from limitations such as restrictive multiple-choice formats or insufficient clinical relevance. Fundamentally, diagnostic reasoning requires clinicians to identify a specific disease or condition from a set of possible alternatives by analyzing a patient's symptoms, medical history, physical examination, and diagnostic test results. However, multiple-choice

[*] Co-first authors
[†] Corresponding author
[1] https://github.com/SPIRAL-MED/DiagnosisArena

Findings of the Association for Computational Linguistics: ACL 2026, pages 3074–3098

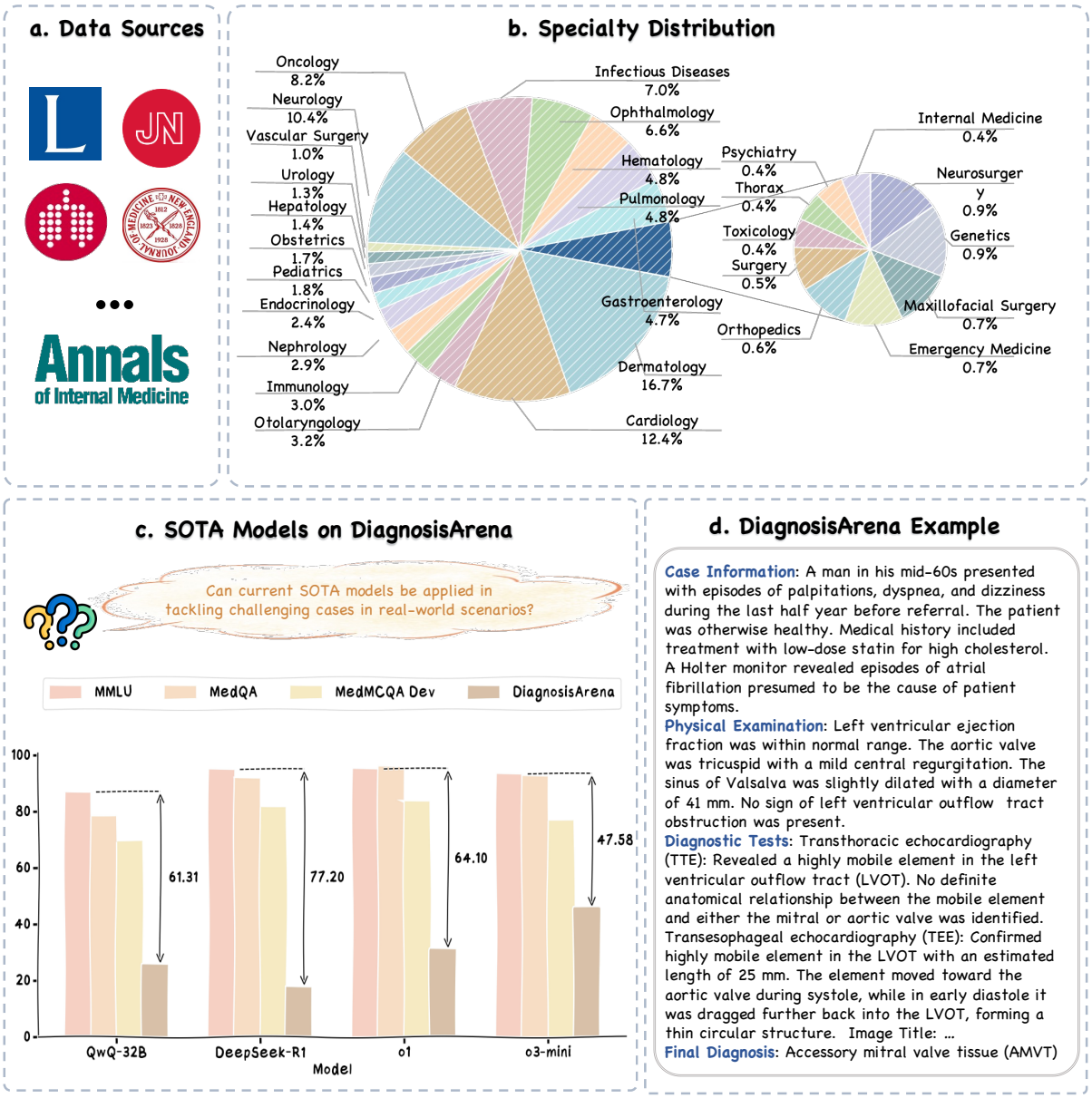

Figure 1: **Overview of the DiagnosisArena Benchmark. (a)** DiagnosisArena is sourced from 10 top-tier medical journals. **(b)** DiagnosisArena is highly diverse, covering 28 medical specialties. **(c)** DiagnosisArena, which emphasizes real-world clinical scenarios, yields significantly lower performance on current leading models, in sharp contrast to results observed on conventional medical benchmarks. **(d)** DiagnosisArena boasts clearly defined segments and offers information-dense clinical cases, which align more closely with clinical practice and present greater reasoning complexity.

formats inherently constrain the scope of differential diagnostics, artificially simplifying the reasoning process. Moreover, effective clinical decision-making demands that clinicians determine individual patient nuances from extensive and complex information. Existing studies typically perform evaluations using overly simplified case information, thus diverging considerably from real-world clinical complexity. These limitations highlight the urgent need to develop benchmarks capable of rigorously evaluating LLMs in a manner that not only closely mirrors real-world clinical complexities but also emphasizes highly challenging diagnostic cases.

In this paper, we introduce *DiagnosisArena*, a comprehensive and highly challenging benchmark comprising professional diagnostic problems. Given that, in real-world scenarios, treatment recommendations from AI models require extensive validation through prolonged clinical trials and experiments—currently difficult to achieve—we focus primarily on the diagnostic aspect. *Diagnosis-*

Arena is developed through a meticulous construction pipeline, involving data collection, data segmenting, iterative filtering, and expert-AI collaborative verification. Initially, we collected an extensive set of real-world case reports from all of the professional top-tier medical journals—including Lancet, NEJM, JAMA, and so on—to ensure authenticity and diversity. Subsequently, we perform segmented data transformation, converting raw case reports into standardized segmented formats. Each segmented case encompasses detailed *case information*, *physical examination findings*, and *diagnostic test results*, with the final diagnosis serving as the ground truth. To ensure the complexity and quality of the benchmark, iterative filtering was conducted based on analyses from AI experts, coupled with AI-based reviews to verify that each included case contains sufficient and unambiguous information necessary for arriving at the final diagnosis. Moreover, to enhance robustness and minimize potential errors, we employ an expert-AI collaborative verification process: cases are excluded if frontier LLMs failed to arrive at the correct diagnosis within 8 attempts, or if they fail approval by board-certified physicians. The inclusion of highly challenging cases and a multi-stage curation process culminates in a professional-grade benchmark comprising 1,113 structured clinical cases across 28 medical specialties, designed specifically for evaluating the diagnostic reasoning capabilities of LLMs in complex clinical scenarios.

To quantitatively evaluate the performance of diagnostic outputs, we adopt GPT-4o (OpenAI, 2024a), one of the most powerful models for knowledge-intensive tasks, as a judge to categorize the relationship between a model's diagnostic results and the ground truth diagnosis as either "identical", "relevant", or "irrelevant". For each case, we generate five candidate diagnostic outputs and calculate both top-1 and top-5 accuracy scores. We conduct extensive experiments involving proprietary models (OpenAI, 2024b,a; Anthropic, 2024; Team, 2024; OpenAI, 2025; Deepmind, 2025) and open-source models (DeepSeek-AI, 2025; Bingning Wang et al., 2025; Team, 2025b; DeepSeek-AI, 2024; Team, 2025a), including advanced reasoning models (e.g. o1 and DeepSeek-R1). Results on *DiagnosisArena* reveal that: (1) Current leading models exhibit notably low performance in diagnosing professional-level clinical cases. Even the most advanced reasoning models, o3-mini, o1, and DeepSeek-R1, achieve only 45.82%, 31.09%, and 17.79%, respectively, while other models struggle to surpass 20%. These findings underscore a substantial gap between existing LLM capabilities and the performance required for professional-level medical diagnostics. (2) We examine potential data leakage effects by examining the publication dates of the original case reports. Our analysis indicates that instances of data leakage within our benchmark are negligible, as indicated by the absence of significant divergence in model performance for cases published before versus after the training data cut-off date. This observation highlights the necessity of developing more sophisticated paradigms to enhance AI capabilities for addressing challenging clinical problems. (3) By converting the diagnostic task in *DiagnosisArena* into the multiple-choice format based on model-generated diagnoses as *DiagnosisArena-MCQ*, we observe a marked increase in model performance, with o1 reaching 61.90%, further suggesting that multiple-choice formats inherently reduce task difficulty and thus fail to accurately reflect the models' true abilities in addressing complex clinical problems.

In summary, the contributions of our study are threefold:

- We introduce *DiagnosisArena*, a comprehensive and challenging benchmark of 1,113 pairs of segmented patient cases designed to rigorously assess professional-level diagnostic competence. Through meticulous filtering and data augmentation, *DiagnosisArena* ensures high difficulty and robustness.
- We conduct extensive evaluations on *DiagnosisArena* using a range of open-source and proprietary large language models, including 11 frontier models. Moreover, the analysis of data leakage confirms that data leakage is negligible within the benchmark.
- The experiments reveal the limitations of current models: Although models exhibit promising capabilities in clinical diagnostic reasoning, a significant gap remains between their performance and the requirements of real-world clinical applications.

2 Related Work

Reasoning LLMs. Endowing LLMs with reasoning abilities has been regarded as a challenging task. As the overall capabilities of LLMs continue to improve, enhancing their reasoning abilities has

increasingly become a focal point of research. Initially, the focus shifted from the few-shot approach to making LLMs mimic the reasoning process to fixed paradigms such as Chain-of-Thought (Wei et al., 2022) and ReAct (Yao et al., 2023b) to stimulate LLMs' reasoning abilities. Later, heuristic search and process-level reward models were introduced. Heuristic search drew inspiration from traditional search algorithms, such as Monte Carlo Tree Search (Silver et al., 2017), and expanded the LLM's single chain of thought into a more complex tree structure of thought, as seen in (Yao et al., 2023a; Shinn et al., 2023). On the other hand, the process-level reward model models the reasoning process as a Markov Decision Process, gradually assigning rewards to intermediate steps, guiding the model to generate more accurate chains of thought (Lightman et al., 2023; Jiao et al., 2024; Wang et al., 2023a). Then, the advent of OpenAI's o1 (OpenAI, 2024b) sparked a wave of research on reasoning. Journey Learning (Qin et al., 2024) explores multiple strategies to replicate the o1-like slow-thinking reasoning. The o1-coder (Zhang et al., 2024) integrates Monte Carlo Tree Search in code-related domains. STILL-2 (Min et al., 2024) distills long-form reasoning data, expands potential solution paths, and iteratively optimizes. These efforts culminated in breakthroughs exemplified by DeepSeek's achievements (DeepSeek-AI, 2025), which underscore the critical role of reinforcement learning in augmenting reasoning performance. The development of reasoning models continues to evolve.

Benchmark for Medical LLMs. As LLMs continue to advance in medicine, the benchmark for medical applications is also evolving. Before the emergence of LLMs, most commonly medical benchmarks like MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) were derived from medical licensing examinations, such as USMLE, CHMLE, or others. They primarily focus on assessing LLMs' mastery of standardized medical knowledge. As the application scenarios of LLMs expand, benchmarks based on subdomain applications have emerged, such as medical calculators (Khandekar et al., 2024; Zhu et al., 2024), medical visual X-ray (Zhou et al., 2024), and medical code (Lee and Lindsey, 2024). However, benchmarks focused on knowledge assessment do not align with expectations for LLMs to perform diagnosis and treatment in clinical settings. Recent studies have increasingly focused on clinical scenarios, with these benchmarks moving away from reliance on medical licensing examinations and instead drawing data from broader sources to better align with real-world clinical scenarios. Examples of such work include Medbullets (Chen et al., 2024a), CMB-Clin (Wang et al., 2023b) and RareArena (zhao zy15, 2024). With the development of reasoning models in the medical field, research indicates that traditional benchmarks have gradually lost their challenge (Nori et al., 2024). As a result, researchers have started to introduce more difficult and multi-level filtration benchmarks (Zuo et al., 2025; Qiu et al., 2025). We believe that the fundamental reason for this phenomenon lies in the insufficient assessment of model reasoning abilities in clinical scenarios by traditional benchmarks. We focus on medical diagnosis, a scenario that demands complex reasoning over multidimensional patient records. Building on this, we introduce *DiagnosisArena*, a benchmark designed to comprehensively evaluate diagnostic reasoning.

3 DiagnosisArena

3.1 Overview

We introduce *DiagnosisArena*, a comprehensive and challenging benchmark for medical diagnosis, designed to evaluate the capabilities of LLMs in diagnosing challenging cases in real-world scenarios. In Section 3.2, we introduce the construction pipeline, which comprises data collection, data segmenting, iterative filtering, and expert-AI collaborative verification. In Section 3.3, we compare *DiagnosisArena* with other existing benchmarks. With comprehensive case information and alignment with realistic clinical diagnostic scenarios, *DiagnosisArena* serves as an effective benchmark for evaluating the performance of LLMs in addressing complex diagnostic tasks.

3.2 Construction

In real-world clinical scenarios, diagnostic decision-making requires physicians to meticulously analyze extensive patient records—including symptoms, medical history, physical examination, and diagnostic test results—to piece together the full picture of the patient's condition. This process inherently demands sophisticated reasoning abilities. Although current reasoning models have demonstrated significant proficiency in medical licensing exams, their performance in real-world diagnostic

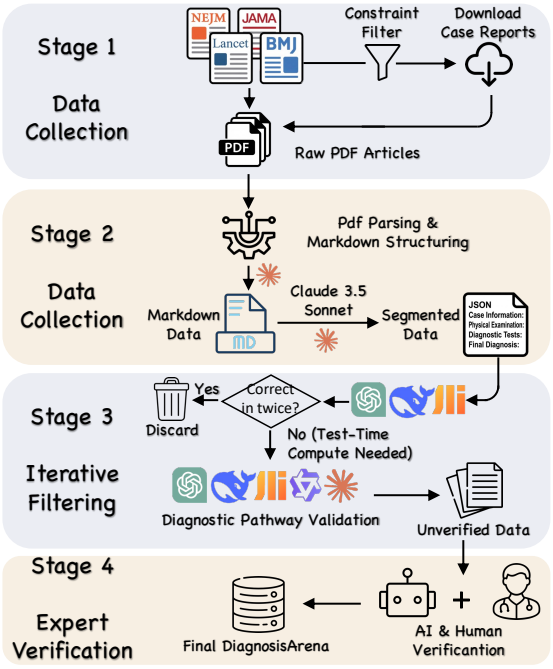

Figure 2: The pipeline for constructing the DiagnosisArena dataset consists of four stages: data collection from the journals, data segmenting, iterative filtering of non-reasoning examples, and expert-AI collaborative verification.

contexts remains uncertain. To address this gap, we constructed a benchmark designed to closely replicate authentic clinical situations and to pose substantial reasoning challenges, thereby effectively evaluating the upper limits of existing reasoning models. Our meticulous data development pipeline encompasses data collection, data segmenting, iterative filtering, and expert-AI collaborative verification, as shown in Figure 2.

Data Collection. Performing medical diagnosis is an information-dense task, wherein overlooking subtle yet critical observations can lead to entirely different diagnostic conclusions. Therefore, it is essential to ensure sufficient and detailed case information during data collection. Moreover, since LLMs can perform effectively on trivial diagnostic tasks (Nori et al., 2024), it is imperative to incorporate more challenging scenarios to better evaluate the upper limits of their capabilities. Based on these considerations, we focused specifically on case reports published in top-tier, high-impact medical journals, as these reports typically present challenging cases of substantial research value while providing comprehensive diagnostic information. Furthermore, we endeavored to include a broad spectrum of cases spanning various medical specialties to enable our benchmark to thoroughly evaluate the diagnostic proficiency of LLMs across multiple clinical disciplines. Consequently, we conducted an extensive review of numerous medical journals and ultimately selected 10 target journals as our data sources, from which we collected a total of 4,175 case reports.

Data Segmenting. Since the raw data collected may include treatment details or follow-up information that could imply the actual diagnostic outcomes, it is crucial to distinguish between prognostic and diagnostic information, incorporating only diagnostic-relevant content into our benchmark. To achieve this, we apply a combination of rule-based filtering and model-based segmenting to convert the unsegmented raw data into a standardized Markdown format. Specifically, explicit treatment plans and prognostic information were initially filtered out manually based on chapter headings. Subsequently, we employed `Claude-3.5-sonnet` to systematically extract and structure the diagnostic-related content. Consequently, each case report was reorganized into four sections: case information, physical examination, diagnostic tests, and final diagnosis. The first three encapsulate data about the patient's clinical presentation, and the last serves as the ground truth.

Iterative Filtering. Clinical diagnoses must be grounded in both patient-specific information and medical knowledge, guided by inductive reasoning. While certain typical cases can be addressed through straightforward knowledge recall, such scenarios are inadequate for evaluating the true capabilities of LLMs when confronted with complex clinical situations. We perform iterative filtering to ensure the complexity and quality of the benchmark. First, we employ `Baichuan-M1`, `DeepSeek-V3`, and `GPT-4o` to eliminate overly simple cases. Each model conducts two sampling attempts per case. If either model produces a correct answer in any of the two attempts, the case is considered too simple and is excluded. Second, we utilize AI experts to assess whether each remaining case contains sufficient contextual clues to support a logically sound diagnostic pathway. Only cases unanimously judged as reasonable by all AI expert reviewers are retained. Following this procedure, we retain 1,783 cases for the next stage.

Expert-AI Collaborative Verification. To ensure the accuracy of the final diagnostic results and minimize potential errors, we employed an Expert-AI Collaborative Verification mechanism. First, we used the advanced model `DeepSeek-R1` to perform

Benchmark	# Sample Size	# Average Length	Problem Type	Exams & Tests	Clinical Scenarios	Data Source
PubMedQA (Jin et al., 2019)	1, 000	328.41	*Close-Ended*	✘ / ✘	✘	PubMed
MMLU (Medical) (Hendrycks et al., 2020)	1, 089	100.07	*MCQ*	✘ / ✘	✘	Licensing Exams
MedQA-USMLE (Jin et al., 2021)	1, 273	215.46	*MCQ*	✔ /✘	✔	Licensing Exams
MedMCQA-Dev (Pal et al., 2022)	4, 183	53.84	*MCQ*	✘ /✘	✘	Licensing Exams
CMExam (Liu et al., 2023)	6, 811	150.67	*MCQ*	✘ /✘	✘	Licensing Exams
C-Eval (Medicine) (Huang et al., 2023)	375	32.89	*MCQ*	✘ /✘	✘	Licensing Exams
CMB-Clin (Wang et al., 2023b)	74	792.55	*Open-Ended*	✔ / ✔	✔	Hospital
MMLU-Pro (Medical) (Wang et al., 2024)	586	166.63	*MCQ*	✘ / ✘	✘	Licensing Exams
Medbullets (Chen et al., 2024a)	124	209.95	*MCQ*	✔ / ✔	✔	Question Bank
RareArena (zhao zy15, 2024)	72, 661	310.36	*Open-Ended*	✔ / ✔	✔	PubMed
MedXpertQA Text (Zuo et al., 2025)	2, 450	257.43	*MCQ*	✔ / ✔	✔	Exams & Boards
MedR-Bench (Qiu et al., 2025)	1, 453	335.37	*Open-Ended*	✔ / ✔	✔	PubMed
DiagnosisArena	1,113	545.02	*Open-Ended & MCQ*	✔ / ✔	✔	Top-tier Journals

Table 1: **Comparisons with existing medical benchmarks.** We categorize existing benchmarks into three types based on chronological milestones: those developed before the emergence of LLMs, after the introduction of LLMs and following the advent of reasoning-based models. In terms of **Exams & Tests**, **Exams** refer to physical examinations, whereas **Tests** denote diagnostic tests. And *MCQ* refers to multiple-choice questions. Unlike prior benchmarks that are primarily derived from licensing examinations or focus on simplified cases with limited patient information, *DiagnosisArena* presents significantly greater challenges for state-of-the-art LLMs. This increased difficulty stems from its inclusion of rich patients' records and complex clinical scenarios that cannot be addressed through pretrained knowledge alone.

multiple rounds of sampling. Specifically, cases are excluded if the model failed to arrive at the correct diagnosis within 8 attempts. Next, we enlisted board-certified physicians to conduct reviews. If the experts identified missing information or ambiguity in the diagnosis, the corresponding cases were also excluded. In total, 1,113 cases were retained through this process.

3.3 Comparison

With the above pipeline, we constructed *DiagnosisArena*, a benchmark comprising 1,113 cases across 28 medical specialties. Table 1 compares *DiagnosisArena* with other existing benchmarks. We divide existing benchmarks into three categories. The first one consists of traditional and widely used medical benchmarks, with data primarily derived from questions of the Medical Licensing Examination. The second one emerged alongside the development of LLMs, encompassing a wide range of task types—from question banks to clinical case analyses. The third one represents recent benchmarks that appeared following the release of `OpenAI-o1`. While these benchmarks aim to address more complex medical problems, they often remain limited to multiple-choice formats or lack real-world clinical context.

In contrast, *DiagnosisArena* is sourced from clinical case reports published in 10 top-tier medical journals. These case reports are comprehensive and inherently challenging, making them well-suited for evaluating the diagnostic reasoning capabilities of LLMs. In clinical practice, comprehensive physical examinations and diagnostic tests (e.g., blood tests, CT scans) are essential for assessing a patient's condition and identifying health issues. By incorporating such clinical information, *DiagnosisArena* closely mirrors real-world medical scenarios and includes detailed patient data such as age, sex, presenting symptoms, examination findings, etc. Furthermore, our rigorous filtering pipeline excludes cases that can be resolved solely using memorized specific knowledge, thereby emphasizing the evaluation of reasoning abilities rather than simple retrieval.

4 Experiments

4.1 Implementation Details

Setup. To fairly evaluate the diagnostic capabilities of current LLMs on our *DiagnosisArena*, we use a unified prompt to instruct models to generate five possible diagnostic outcomes in descending order of confidence. To minimize potential biases, our prompt did not impose strict constraints on the output format.

Models. Our experiments include both proprietary and open-source models, as well as domain-specific medical models and general models, such as Baichuan-M1 (Bingning Wang et al., 2025), DeepSeek-V3 (DeepSeek-AI, 2024), GPT-4o (OpenAI, 2024a), Claude-3.5-Sonnet (Anthropic, 2024), Qwen2.5-Max (Team, 2024). Notably, we

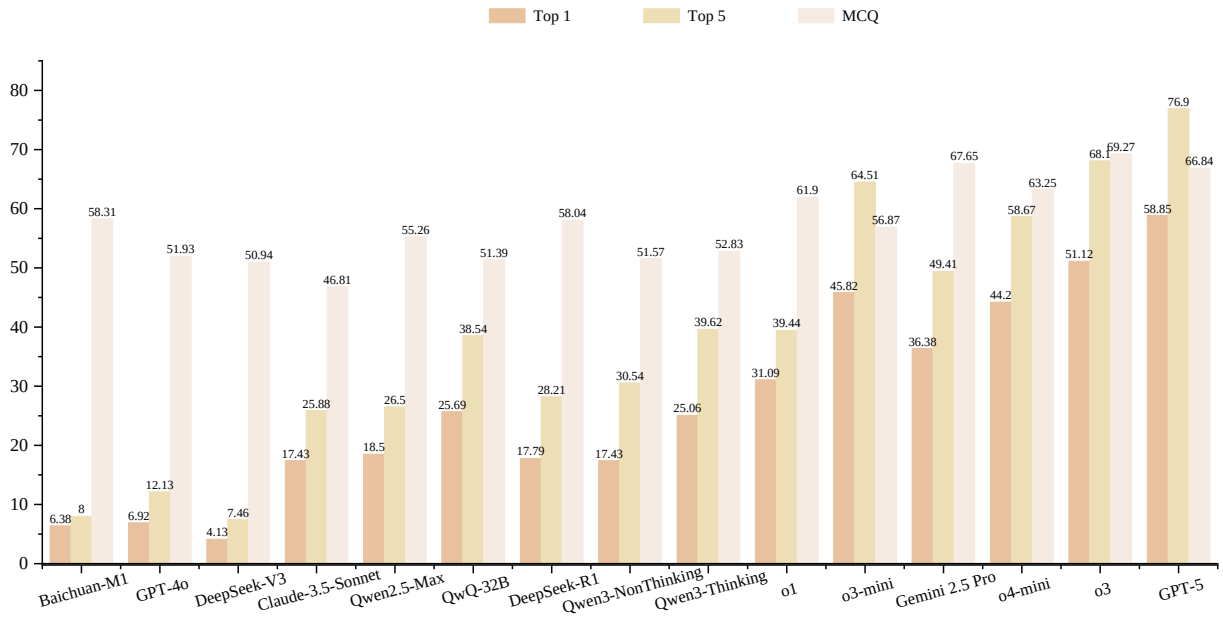

Figure 3: **Performance of Different Models on *DiagnosisArena*.** (a) The Top-k metric represents the proportion of cases where the correct answer is included among the Top-k predictions generated by the model, ranked in descending order of confidence. The results reveal that while the o3-mini outperforms others, *DiagnosisArena* remains a significant challenge for all existing models. (b) The MCQ presents the multiple-choice version of *DiagnosisArena*. A marked increase in model performance can be observed, with o1 reaching 61.90%.

also focused on assessing models that leverage inference-time scaling, such as open-source QwQ-32B (Team, 2025b), OpenAI's o1 (OpenAI, 2024b), DeepSeek-R1 (DeepSeek-AI, 2025), Qwen3-235B-A22B (Team, 2025a), Gemini 2.5 Pro (Deepmind, 2025), and o3-mini (OpenAI, 2025). Additionally, we tested both the thinking and non-thinking configurations of Qwen3-235B-A22B, denoted as Qwen3-Thinking (with enable_thinking set to true) and Qwen3-NonThinking (with enable_thinking set to false).

4.2 Evaluation

Open-Ended Question Evaluation. For open-ended questions, the ground truth is a clear diagnostic conclusion. In clinical scenarios, medical diagnoses are generally classified into three categories: "identical", "relevant", and "irrelevant" (zhao zy15, 2024). Therefore, we use `GPT-4o` as the judge to evaluate the results into these three categories (McDuff et al., 2025). Among them, only when the model's result is judged to be "identical" is it considered correct. Additionally, we instruct the LLM to generate the k possible diagnostic outcomes in descending order of confidence, and then calculate the Top k accuracy, which is the hit rate of the correct answer within the top k predicted results.

Multi-Choice Question Evaluation. After open-ended question evaluations, we select partially correct diagnostic results from the answers generated by LLMs—primarily from the evaluation results of o1 and DeepSeek-R1—as distractor options and construct multiple-choice questions with four options. For multiple-choice questions, the ground truth is a fixed option. We instruct the LLMs to answer and perform rule-based extraction and comparison to calculate the accuracy.

4.3 Main Results

Figure 3 presents the main evaluation results of LLMs on the *DiagnosisArena*. Based on our analysis of the experimental data, we draw the following conclusions: (1) Even the most advanced reasoning LLMs struggle with this task. The best-performing model, o3-mini, achieved an accuracy of only 45.82% on *DiagnosisArena*, while o1 and DeepSeek-R1 performed even worse, with an accuracy as low as 31.09% and 17.79%. This substantial performance gap highlights the significant difficulty of our benchmark and the limitations of current models. (2) Models endowed with explicit reasoning capabilities demonstrate a clear advantage in clinical diagnostic tasks. Notably, even powerful models such as Claude-3.5-Sonnet and Qwen2.5-Max achieve suboptimal performance—below 20% accuracy. In contrast, reasoning-enhanced models, including those with relatively smaller parameter sizes (e.g., QwQ-32B), attain significantly higher

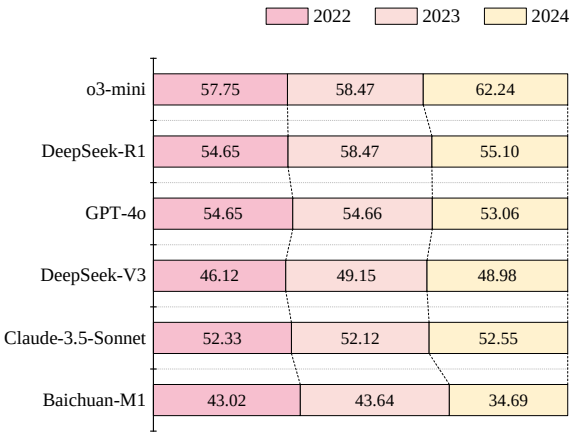

Figure 4: **Leakage Detection on *DiagnosisArena*.** For all models, the experimental results maintained a generally consistent trend across different years, with only minor fluctuations.

accuracy, reaching 25.69%. A particularly illustrative case is DeepSeek-R1, a reasoning-oriented model trained by DeepSeek-V3, which achieves a 13.66% improvement in accuracy over its base model. These findings underscore the critical role of reasoning in clinical diagnosis. (3) Solely prompting LLMs to select the best answer in a multiple-choice format does not accurately reflect their true capabilities in handling clinical tasks. A significant performance improvement is often observed in the multiple-choice setting—for example, o1 achieves 61.90%. Even Baichuan-M1-14B, which scores below 10% in open-ended settings, attains 58.31% in the multiple-choice format. This discrepancy can be attributed to the inherently simplified nature of multiple-choice questions, where predefined options narrow the problem space. In such cases, LLMs can leverage superficial cues or rely on partial knowledge to eliminate incorrect choices and select the most plausible answer. However, this does not constitute a complete deductive reasoning process as required in real-world diagnostic scenarios.

4.4 Data Leakage

The scope of pretraining corpora is increasingly broad, with academic journal papers often included in the pretraining data of LLMs. This may cause LLMs to memorize specific cases and reproduce knowledge to answer questions. To ensure the robustness of *DiagnosisArena*, we conducted experiments to detect data leakage.

To verify the presence of data leakage from sources, we conducted the following analysis. We collected 690 journal entries from 2022 to 2024, including publications from JAMA and NEJM. Data from 2025 were excluded due to insufficient volume for a meaningful comparison. The datasets were balanced across years, with approximately 200 entries per year. We performed preliminary processing and evaluation on these datasets, and the results are shown in Figure 4.

We can observe that for mainstream models remained largely stable across the three years, exhibiting only minor fluctuations. This suggests that either no data leakage occurred or any leakage that did occur had a negligible impact on model evaluation. However, for models in specific, the Baichuan-M1, the accuracy exhibited a slight decline in 2024, indicating that data from prior to 2024 may have leaked into its training corpus, thereby affecting its evaluation performance.

Based on the pilot studies, to directly perform data leakage detection, we conducted another statistical analysis on the constructed *DiagnosisArena*, with the results shown in Figure 11. As observed, compared to the results of the Pre-experiment small sample Leakage Detection, DiagnosisArena exhibited more irregular fluctuations. However, from the overall trend over the decade, neither mainstream general models nor domain-specific medical models showed significant accuracy fluctuations over time. This result indicates that the evaluation in *DiagnosisArena* did not experience performance anomalies due to data leakage.

5 Conclusion

We introduce *DiagnosisArena*, a challenging medical benchmark for evaluating LLMs' diagnostic reasoning capabilities in clinical settings. It comprises 1,113 structured clinical cases spanning 28 specialties, reflecting real-world diagnostic complexity. Our experiments reveal that even SOTA reasoning models struggle significantly on *DiagnosisArena*, despite performing well on its multiple-choice variant. This gap underscores that multiple-choice formats provide shortcuts that mask models' true reasoning limitations. Case studies further show that models tend to favor common diseases over evidence-based inference, indicating insufficient adaptation to clinical reasoning demands. We hope that *DiagnosisArena* will contribute to advancing reasoning abilities in the field of medicine.

Limitations

Despite its rigorous design and comprehensive coverage, DiagnosisArena faces several inherent limitations that warrant consideration. First, the dataset's scale remains constrained. Although we systematically collected data from all of the top-tier journals to ensure high-quality cases, the final sample comprises only 1,113 cases. While this sample size is adequate for benchmarking purposes, it may be inadequate for more effective development and training applications. Second, the validation of automated evaluation methods requires further establishment. Although we drew upon existing methodologies for LLM-based diagnostic evaluation and demonstrated strong concordance between human assessment and GPT-4o evaluation through experimental validation, such automated evaluation frameworks have not yet achieved widespread acceptance in the clinical community. Broader experimentation and consensus-building efforts will be necessary to establish their validity and reliability as standard evaluation tools.

References

Anthropic. 2024. claude-3-5-sonnet.

Huozhi Zhou Liang Song Mingyu Xu Wei Cheng Xiangrong Zeng Yupeng Zhang Yuqi Huo Zecheng Wang Zhengyun Zhao Bingning Wang, Haizhou Zhao and 1 others. 2025. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*.

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*.

Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, and 1 others. 2025. An empirical study on eliciting and improving r1-like reasoning models. *arXiv preprint arXiv:2503.04548*.

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028.

Google Deepmind. 2025. Gemini 2.5 pro.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. 1978. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

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. *arXiv preprint arXiv:2501.06458*.

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

Fangkai Jiao, Chengwei Qin, Zhengyuan Liu, Nancy F Chen, and Shafiq Joty. 2024. Learning planning-based reasoning by trajectories collection and process reward synthesizing. *arXiv preprint arXiv:2402.00658*.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad Safranek, Abid Anwar, Andrew

Zhang, and 1 others. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745.

Simon A Lee and Timothy Lindsey. 2024. Can large language models abstract medical coded language? *arXiv preprint arXiv:2403.10822*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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

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

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, and 1 others. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*.

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. *arXiv preprint arXiv:2411.03590*.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Learning to reason with llms.

OpenAI. 2025. Openai o3-mini.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and 1 others. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weike Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, and 1 others. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025a. Qwen3.

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

Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2023a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and 1 others. 2023b. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. 2025. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv preprint arXiv:2501.09213*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

zhao zy15. 2024. Rarearena. Website. https://github.com/zhao-zy15/RareArena.

Yang Zhou, Tan Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. 2024. Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. *Advances in Neural Information Processing Systems*, 37:6625–6647.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2024. Menti: Bridging medical calculator and llm agent with nested tool calling. *arXiv preprint arXiv:2410.13610*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

A Further Experiments

A.1 Human-LLM Agreement Validation

Our evaluation methodology follows the approaches established in (zhao zy15, 2024; McDuff et al., 2025). To further validate the reliability of GPT-4o as an evaluator, we conducted a comparative study between human evaluators and GPT-4o. We randomly selected 150 diagnosis results for independent evaluation, comprising 50 cases from each of three representative models: DeepSeek-R1-0528, gpt-oss-120b-high, and o3. Both human evaluators and GPT-4o assessed the consistency of these diagnosis results using identical scoring criteria.

The comparative analysis revealed strong inter-rater reliability, with a 93.3% complete agreement rate and **a Cohen's kappa coefficient of 0.840**, indicating substantial agreement between the human evaluation and GPT-4o. These findings validate the reliability of the GPT-4o-based evaluation for assessing diagnosis results.

A.2 Hypothetico-Deductive Reasoning Analysis

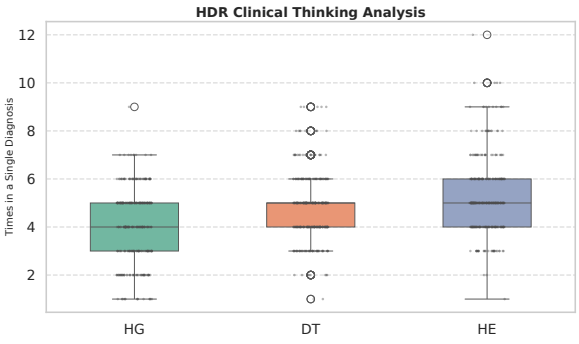

Figure 5: Distribution of Hypothetico-Deductive Reasoning Steps in Clinical Diagnosis by Qwen3-235B-A22B-2507

To investigate the specific reasoning approaches employed by existing large language models in clinical diagnostic reasoning, we selected Qwen3-235B-A22B-2507 as our research subject due to its outstanding performance among open-source models. As a reasoning model, it provides complete inference processes suitable for analysis. We utilized Deepseek-V3.1 to classify reasoning patterns and extract hypothetico-deductive steps from the thinking processes. Results are presented in Figure 5. Statistical analysis revealed that among 373 correctly diagnosed cases, Qwen3-235B-A22B-2507

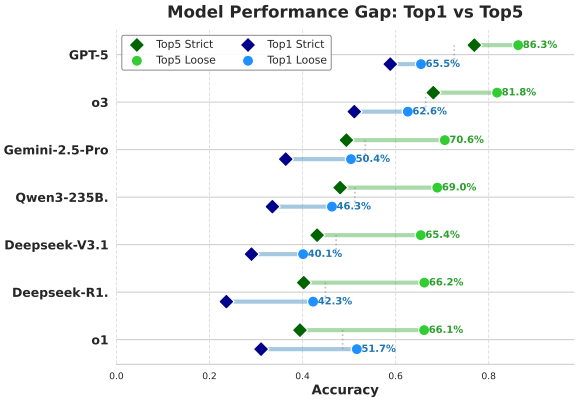

Figure 6: Performance Comparison under Strict vs. Loose Matching Criteria.

engaged in extensive hypothetico-deductive reasoning in 98.4% of cases. The figure illustrates the breakdown of hypothetico-deductive steps per case on average, including Hypothesis Generation (HG), Deductive Testing (DT), and Hypothesis Evaluation (HE), which constitute the specific components of hypothetico-deductive reasoning (Elstein et al., 1978).

The distribution of Hypothesis Generation exhibits notable convergence characteristics, with a median frequency stabilizing at 4 (IQR: 3-5). This indicates that when confronted with clinical presentations, Qwen3-235B-A22B-2507 adopts a focused differential diagnosis strategy. Rather than engaging in divergent speculation, the model leverages prior medical knowledge to establish a bounded problem space. For Deductive Testing, the median of 5 slightly exceeds that of Hypothesis Generation, with the overall distribution shifting upward. This "more testing than hypotheses" pattern demonstrates that the model conducts an average of more than one active verification per diagnostic hypothesis, seeking confirmatory or disconfirmatory evidence through laboratory tests, imaging features, and other clinical data, rather than engaging in arbitrary speculation. This provides evidence that Qwen3-235B-A22B-2507 does not rely on superficial pattern matching (System 1), but rather employs a deliberative System 2 thinking mechanism analogous to human physicians (Croskerry, 2009).

Hypothesis Evaluation demonstrates the greatest distributional variance and significant outliers extending to 12 instances. This reflects the model's adaptive evaluation capabilities. For typical cases, the model achieves rapid diagnosis through standard 3-4 evaluations; however, for complex or am-

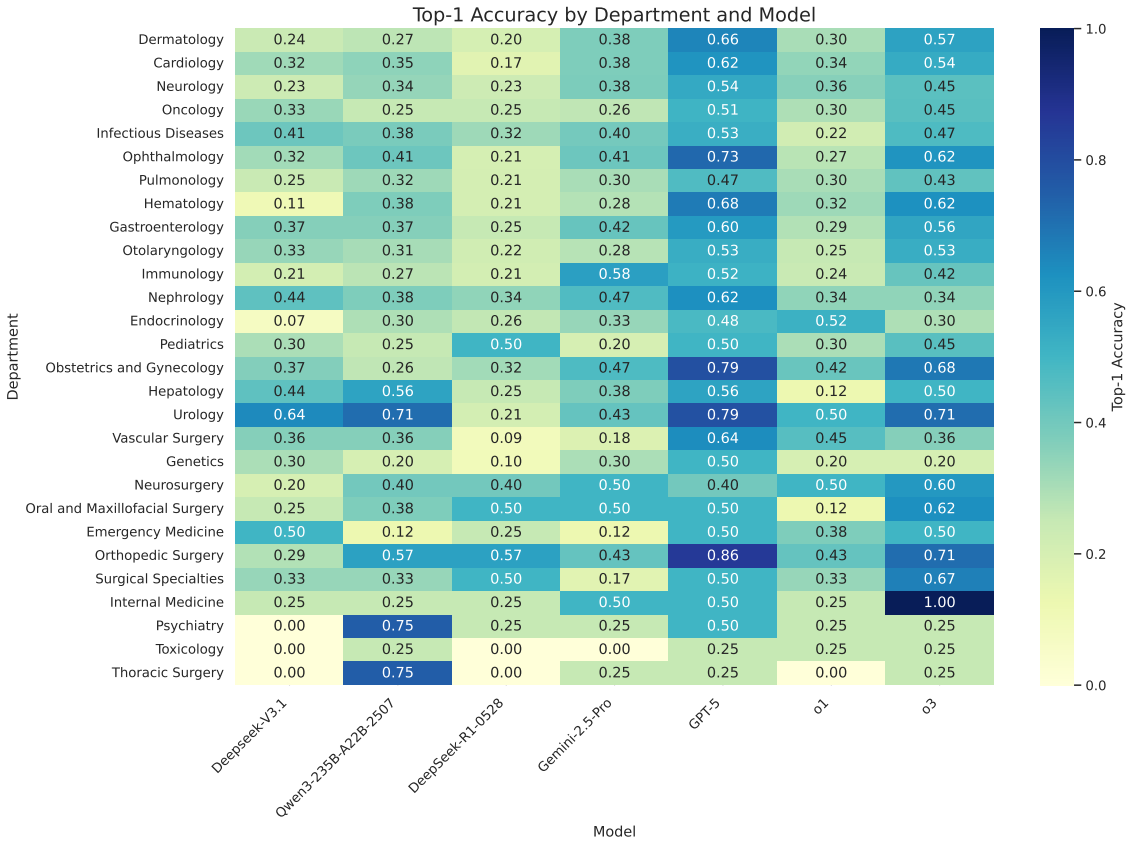

Figure 7: Heatmap of Top-1 Diagnostic Accuracy Across 28 Medical Departments.

biguous cases, the model exhibits remarkable persistence, employing iterative evaluation-revision-reevaluation cycles to process uncertain information. In summary, Qwen3-235B-A22B-2507's clinical reasoning represents not merely mechanical text generation, but rather a structured hypothetico-deductive process.

A.3 Impact of Matching Strictness

To comprehensively evaluate models' diagnostic capabilities using more refined criteria, we compared these scoring mechanisms, as illustrated in the figure. Specifically, we categorized the relationship between diagnostic results and ground truth into three types: "identical", "relevant" and "irrelevant". Under the strict criterion, only "identical" matches are considered correct and awarded 1 point. Under the loose criterion, "relevant" matches are also recognized as partially correct and awarded 0.5 points. The figure 6 visually demonstrates the score distributions under both loose and strict scoring standards.

Experimental results reveal that all evaluated models exhibit significant performance degradation when transitioning from loose to strict criteria. Taking the best-performing GPT-5 as an example, while it achieves an impressive 86.3% accuracy under the Top-5 Loose setting, its score drops by 9.4% under the strict standard. This phenomenon is even more pronounced in the Top-1 metric, indicating that generating perfectly accurate answers remains challenging even for state-of-the-art models.

However, despite the differences in absolute values between the two scoring standards, the relative rankings of models remain highly consistent. As shown in the figure, GPT-5 and o3 consistently occupy the top tier under both strict and loose criteria, significantly outperforming other models. Gemini-2.5-Pro, Qwen3-235B-A22B-2507, and other models follow closely behind, with clear tier stratification across the board. The performance curves corresponding to squares and circles exhibit remarkable similarity. This ranking consistency demonstrates that stricter scoring criteria neither introduce noise nor fundamentally alter our assessment of model capabilities. While the two standards provide different numerical perspectives, they converge in capturing the essential trends of model performance.

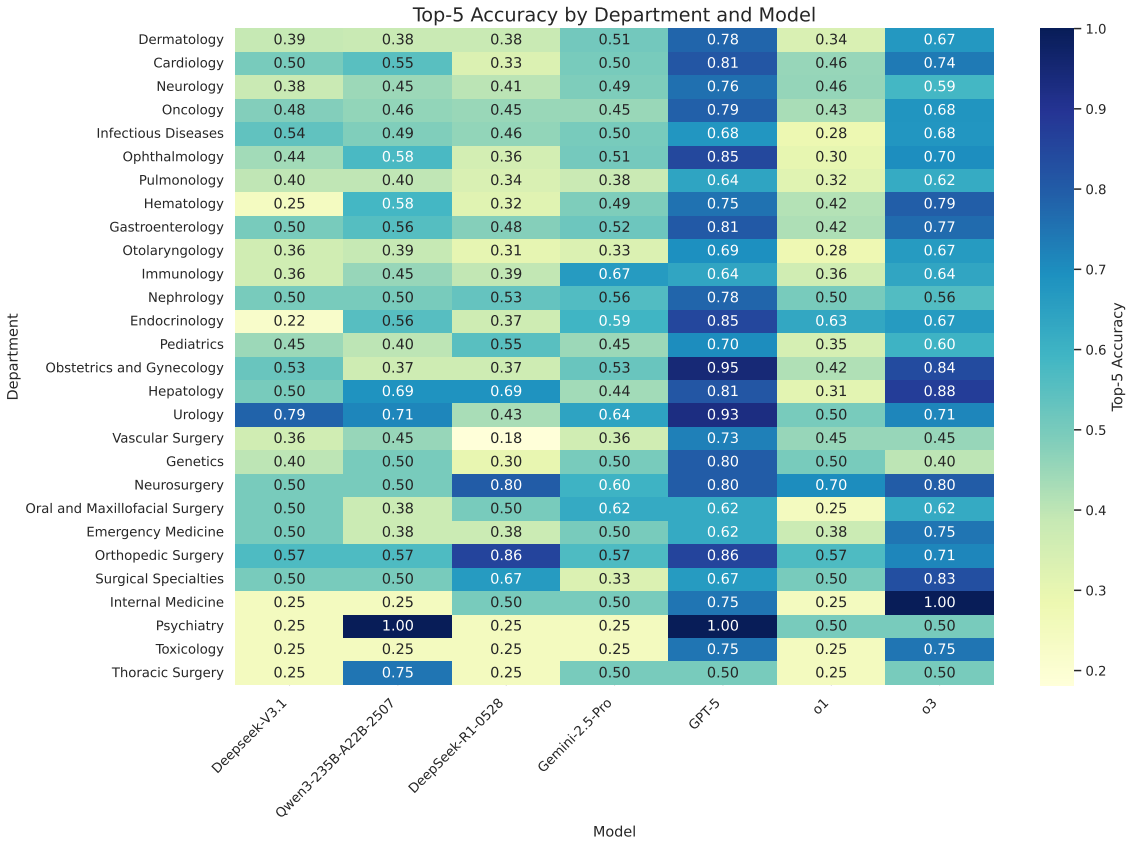

Figure 8: Heatmap of Top-5 Diagnostic Accuracy Across 28 Medical Departments.

A.4 Department Breakdown Analysis

To gain a more comprehensive understanding of the model's performance across different medical specialties, we conducted a stratified analysis of diagnostic results across 28 medical departments. The results are presented in the Figure 7, 8 and 9. In the figure, departments are arranged in descending order based on the number of cases.

Top-1 Accuracy Breakdown. The heatmap distribution reveals distinct performance tiers among models in medical diagnostic tasks. Leading models demonstrate significant advantages. State-of-the-art models such as GPT-5 and o3 maintain consistently high Top-1 accuracy across the vast majority of specialties. For instance, in Obstetrics and Gynecology, they achieve accuracy rates of 0.79 and 0.68 respectively, substantially outperforming other models in their cohort, which average between 0.26-0.47. Gemini-2.5-Pro and o1 constitute the second tier. In contrast, open-source models in the DeepSeek series exhibit performance bottlenecks across multiple specialties. For example, in Hematology, the lowest Top-1 accuracy drops to merely 0.11. This limitation likely stems from the challenge of capturing complex hematological pathological features without domain-specific fine-tuning.

Performance patterns across medical specialties show notable consistency. First, certain specialties present exceptional challenges. Toxicology and Thoracic Surgery not only demonstrate low overall accuracy but also exhibit multiple instances of models scoring 0. This may be attributed to toxicological diagnoses' heavy reliance on rare chemical substance characteristics and the atypical nature of case descriptions. Second, some specialties yield high accuracy rates: Urology and Ophthalmology display universally strong performance. In Urology, multiple models exceed 0.70 accuracy, suggesting that clinical features in this specialty (such as symptom descriptions and biochemical indicators) possess greater distinguishability in semantic space. Third, specialties with extreme variance exist. Internal Medicine exhibits dramatic performance disparities across models. While some models achieve perfect 1.00 accuracy, most hover around 0.25. This suggests that the dataset for this specialty may contain a limited set of typical reasoning patterns that leading models

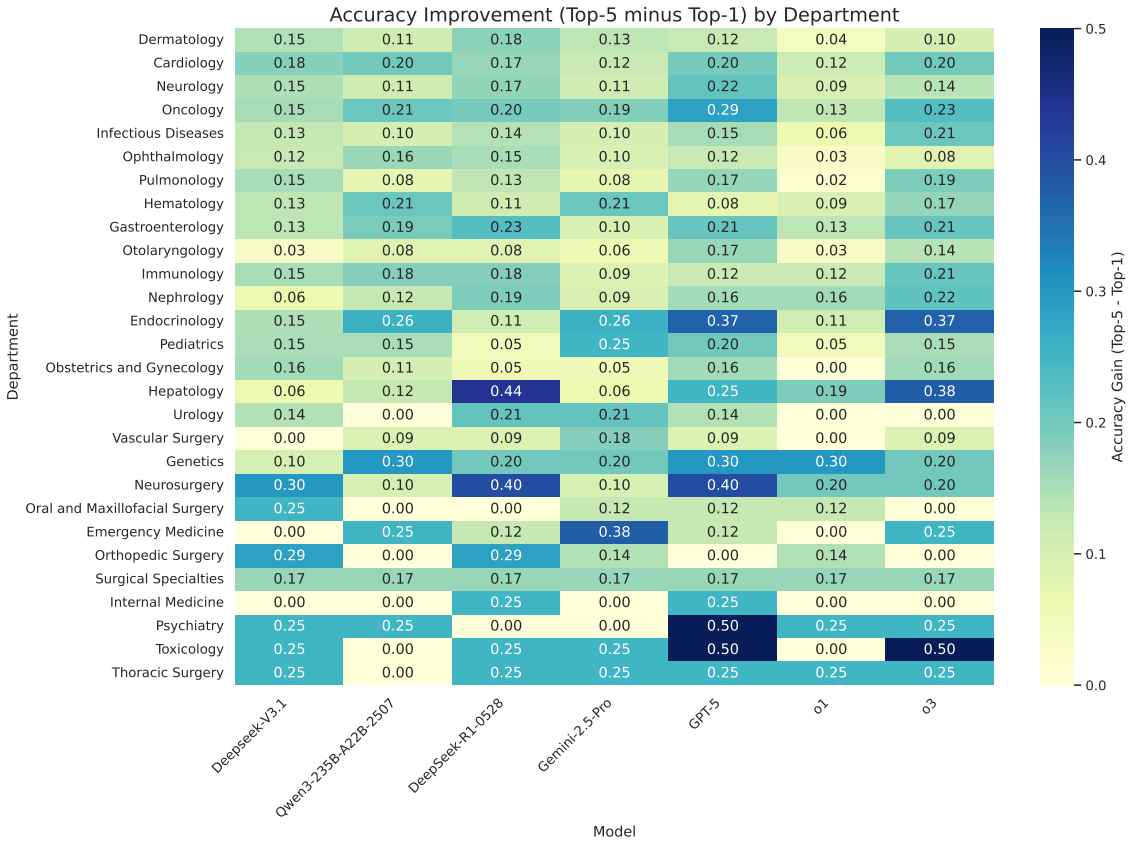

Figure 9: Heatmap of Accuracy Gain (Top-5 minus Top-1) Across Medical Departments.

can master, while weaker models completely lack such capabilities.

Top-5 Accuracy Breakdown. In medical diagnostic scenarios, Top-5 accuracy holds significant clinical importance as it reflects the model's capability to construct a differential diagnosis list. Comparing Top-1 and Top-5 data, we observe substantial performance improvements. Leading models demonstrate dominant stability (above 0.9) in certain specialties, indicating that current SOTA models have already acquired the capability to function as primary care physicians in differential diagnosis screening. However, distinct performance tiers remain evident across different departments.

Urology and Orthopedic Surgery emerge as the top-performing specialties, with average Top-5 accuracy consistently ranging from 0.6-0.8, and Urology achieving a remarkable peak accuracy of 0.93. Obstetrics and Gynecology also demonstrates excellent performance, with the highest model accuracy reaching 0.95. Nevertheless, certain specialties continue to show suboptimal results. Vascular Surgery ranks among the poorest-performing departments, with an average accuracy of merely 0.42 and a minimum score of 0.18. This may be attributed to the complexity of vascular surgery cases, which typically involve intricate anatomical descriptions and dynamic hemodynamic parameters that text-only models struggle to accurately capture. Otolaryngology and Pulmonology also exhibit below-average performance, suggesting that diagnostic features in these domains may present higher levels of ambiguity.

Accuracy Comparative. By calculating the difference between Top-5 and Top-1 accuracy ($\Delta = \text{Top5} - \text{Top1}$), we quantified the variations in reasoning and ranking capabilities across different medical specialties in diagnostic reasoning.

In certain high-difficulty specialties, while models struggle to make precise diagnoses directly (low Top-1), their accuracy shows explosive growth when the prediction scope is relaxed. This indicates that models possess relevant reasoning capabilities but lack precise discriminative ability. Toxicology demonstrates the most significant benefit from this approach. The average Top-1 accuracy is merely 14.3%, yet in Top-5, the average accuracy surges to 39.3% ($\Delta \approx +25\%$). This suggests that for poisoning cases, models can often delineate the range of possible toxins but struggle to pinpoint

the unique causative agent among several similar compounds. Neurosurgery and Endocrinology similarly exhibit substantial improvements (average increase of approximately 23-24%). Endocrine disorders typically involve complex hormonal feedback loops with highly non-specific symptoms (e.g., fatigue, weight changes), causing models to easily confuse superordinate concepts or similar etiologies, though they can effectively recall the correct answer within the Top-5 list.

In another category of specialties, Top-5 provides very limited improvement ($\Delta < 10\%$). One reason is that these specialties already perform exceptionally well, such as Dermatology and Internal Medicine, where some leading models achieve Top-1 accuracy approaching or reaching 1.0. Since predictions are already quite precise, Top-5 cannot offer additional benefits. For instance, Internal Medicine shows an average improvement of only 7.1%, the lowest among all specialties. Another scenario involves poorly performing specialties, such as Vascular Surgery. Despite its mediocre Top-1 accuracy (35%), Top-5 only improves to 42.5%. This suggests that models face fundamental deficiencies in reasoning capabilities or domain knowledge in this field, rather than simple ranking errors.

A.5 Model Performance Disparities

To investigate the interpretability of LLM decision-making, we analyzed a challenging case involving a rare vascular lesion in the cerebellopontine angle and internal auditory canal of a 4-week-old male infant. The diagnostic crux of this case lies in excluding common clinical phenotypes while recognizing the distinctive pathological features. The patient's age and lesion location strongly suggested a clinical diagnosis of infantile hemangioma, which represents a highly prevalent condition in this population. However, the definitive diagnosis relied entirely on interpreting specific histopathological descriptions—namely, "glomeruloid structures" and "spindle cells"—which constitute the gold standard diagnostic criteria for Kaposiform hemangioendothelioma (KHE). This case thus presents an exemplary scenario for examining whether LLMs can overcome the "System 1" intuitive priors regarding disease probability and engage in "System 2" reasoning to correct biases through evidence-based clinical decision-making and accurately identify microscopic pathological evidence (Croskerry, 2009).

Through experimental analysis (presented in Table 2), we observed distinct reasoning patterns across models. Among the top-performing models, GPT-5, o3, and Gemini 2.5 Pro successfully captured the pathological differential points and, through deliberate reasoning, ranked them as their primary diagnosis. Notably, GPT-5 not only achieved accurate Top-1 diagnosis but also predicted Tufted Angioma—which belongs to the same disease spectrum as KHE—as its Top-2 choice, explicitly articulating their spectral relationship. This demonstrates that GPT-5 engaged in more than simple keyword matching; it successfully established deep pathophysiological associations and comprehension of disease spectra.

While DeepSeek-R1-0528 also successfully identified the pathological differential points and its robust reasoning capabilities led to the correct answer, its final decision-making remained conservatively dependent on prior probabilities. It selected the statistically common infantile hemangioma as its primary choice, possibly to mitigate risk, while placing the correctly reasoned KHE in second position. Despite possessing reasoning capabilities, DeepSeek-R1, when faced with conflict, did not trust its reasoning results as confidently as top-tier models.

In contrast, the failures of o1, Qwen3-235B-A22B-2507, and Deepseek-V3.1 exposed cognitive limitations in complex medical reasoning. The o1 model exhibited pronounced anchoring bias, with its reasoning process locked onto two strong clinical features—"4-week-old" and "CPA location"—causing it to overlook decisive pathological descriptions such as glomeruloid structures. Consequently, it produced a standard list of common diseases while completely omitting the target rare disease. This mirrors human cognitive errors where initial impressions impede the updating of differential diagnoses. Qwen3-235B-A22B-2507's error stemmed from medical knowledge confusion; its proposed "congenital hemangioendothelioma" represents an erroneous concatenation of terminology, revealing conceptual ambiguity when processing rare diseases. While Deepseek-V3.1 approached the correct category with "hemangioendothelioma" in its Top-2 prediction, it failed to further specify the "Kaposiform" subtype. This coarse diagnostic granularity suggests the model defaulted to safer, higher-level concepts to mask its insufficient understanding of specific pathological entities.

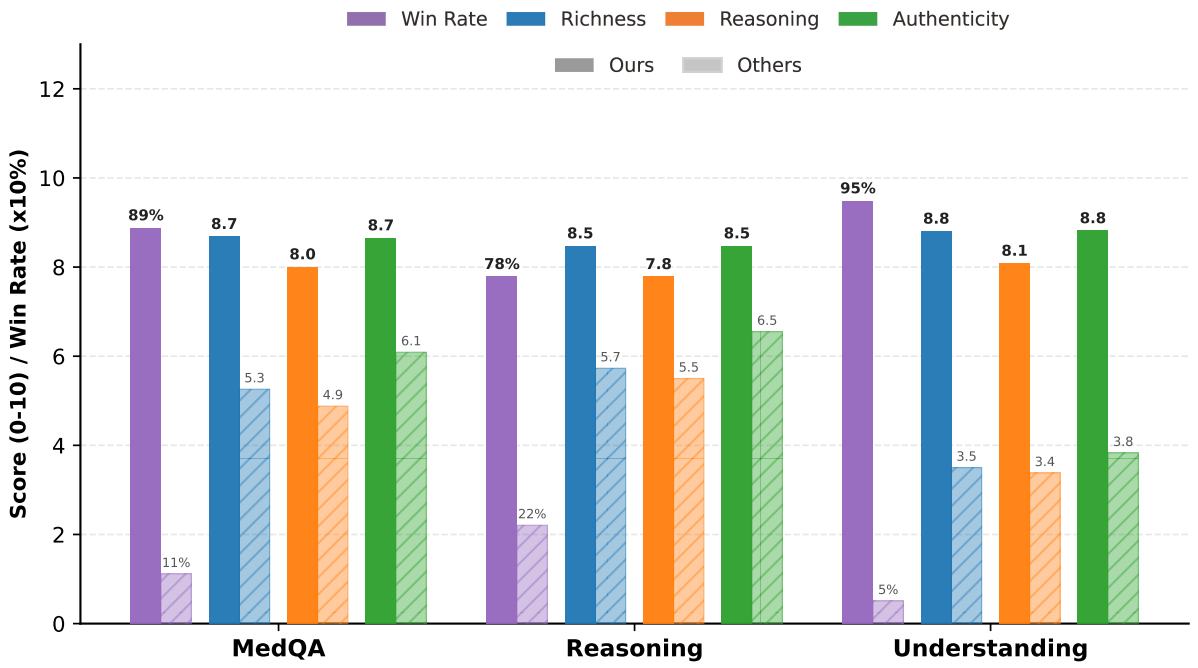

Figure 10: Pairwise quality comparison between DiagnosisArena and existing benchmarks (MedQA and MedXpertQA) evaluated by Deepseek-V3.1 across three dimensions: information richness, reasoning depth, and clinical authenticity.

A.6 Comparative Quality Assessment

To objectively evaluate the quality of DiagnosisArena relative to existing benchmarks, we conducted pairwise comparisons with two leading benchmarks: MedQA (Jin et al., 2021) (a standard benchmark for medical knowledge) and MedXpertQA (Zuo et al., 2025) (an advanced benchmark for complex clinical reasoning, divided into Reasoning and Understanding subsets). Using Deepseek-V3.1 as an impartial judge, we evaluated randomly sampled cases based on information richness, reasoning depth, and clinical authenticity. These metrics respectively represent the level of detail in the data, the complexity of reasoning required for diagnosis, and the consistency with real medical records. We sampled 915 instances from DiagnosisArena and matched them with equal-sized samples from the corresponding datasets through random pairing. To mitigate potential position bias in LLMs, we implemented random position swapping of the data. The results are presented in the Figure 10.

The quantitative results demonstrate that DiagnosisArena consistently outperforms both baseline datasets across all dimensions. DiagnosisArena achieves an 89% win rate against MedQA, showing significant superiority in information richness (8.7 vs. 5.3). This advantage likely stems from MedQA's standard exam-style questions, whereas DiagnosisArena provides heterogeneous data (vital signs, laboratory results) that better simulates the noise and complexity of real-world medical data.

When compared to the reasoning benchmark MedXpertQA, our performance remains impressive. Even against the Reasoning subset, DiagnosisArena achieves a 78% win rate. Overall, while MedXpertQA Reasoning scores slightly higher than MedQA, both fall short of our dataset. This is because both originate from medical clinical exam questions in standardized formats, whereas our data derives from actual clinical cases, resulting in superior comprehensive quality. Similarly, our reasoning depth score (7.8) significantly exceeds the baselines (5.5). This indicates that while MedXpertQA tests logic, DiagnosisArena requires a more complex, multi-step diagnostic process that integrates diverse clinical data. Against MedXpertQA Understanding (which contains only 589 instances, hence we randomly selected 589 DiagnosisArena instances for comparison), our advantage is most pronounced with a 95% win rate.

In conclusion, DiagnosisArena not only increases the difficulty of diagnostic reasoning but also introduces patterns and characteristics that more closely reflect real clinical cases. This contributes to the superior comprehensive quality of

our dataset.

A.7 Data Leakage Results

More results are shown in Figure 11.

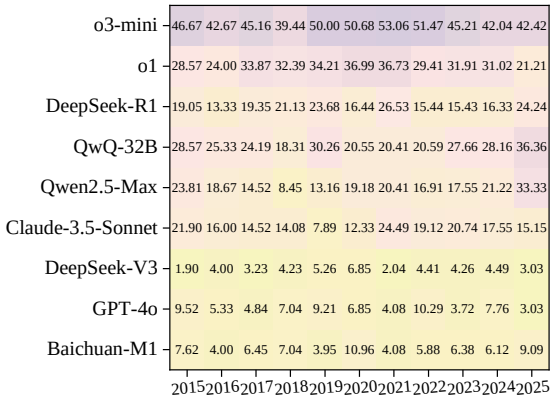

Figure 11: **Leakage Detection on *DiagnosisArena*.** Leakage Detection on the Constructed *DiagnosisArena*. Over the past decade, all models have demonstrated relatively stable accuracy, with no significant fluctuations over time.

B Detailed Statistics

B.1 Distribution of Problems

We collected 4,175 original data entries from the following ten top-tier journals (shown in Table 3). After rigorous screening and validation, we retained 1,113 high-quality data points. These data span 28 medical specialties, with the specific distribution shown in the Figure 1.

B.2 Anonymization

All data utilized in this study were systematically checked and fully anonymized to comply with the standards set forth by the HIPAA Safe Harbor Provision.

B.3 Recruitment

This study was conducted under the supervision of medical experts throughout the entire process, with physicians recruited to provide professional proofreading and participate in experimental procedures. The consultation fee for participating physicians was $40 per hour.

C Case Study

We analyze a case in *DiagnosisArena* and explore the reasons why current reasoning models struggle to solve complex diagnostic problems.

In this case, as shown in Figure 12, the final diagnosis is Accessory Mitral Valve Tissue (AMVT). The primary basis for this diagnosis was the discovery of a highly active abnormal structure in the left ventricular outflow tract (LVOT), which exhibited significant motion patterns during the cardiac cycle. The activity and position of this structure, as well as its unclear anatomical relationship with the mitral valve or aortic valve, supported the conclusion that it was an accessory tissue. Various imaging studies and the incidental finding of mitral annulus separation further confirmed the existence of this structure.

Observing the responses of various models, it is evident that, except for `o3-mini`, which successfully provided the correct answer in the top 1, the other models were far from the correct answer. Analyzing `DeepSeek-R1`'s response below, we can identify the following: (1) The diagnostic approach did not consider the possibility of AMVT. `DeepSeek-R1`'s diagnostic reasoning focused on tumors, emboli, fibrolamellar tumors, and other lesions related to the left ventricular outflow tract (LVOT), while overlooking potential mitral valve-related structural issues. (2) Misjudgment of imaging features. `DeepSeek-R1` identified an active structure through imaging and speculated that it might be a papillary fibrolamellar tumor or Lambl's excrescence. However, it did not consider that AMVT might present as a similar filamentous or active structure. (3) Inadequate consideration of the correlation with clinical symptoms. `DeepSeek-R1` linked the patient's symptoms of palpitations, dizziness, and shortness of breath to various diseases such as cardiac tumors, thrombi, and fibrolamellar tumors, but did not sufficiently consider that AMVT might cause mild to moderate left ventricular outflow tract obstruction or intermittent arrhythmias.

We believe that the root cause of this phenomenon lies in the fact that current SOTA reasoning models have not yet fully adapted to the complex reasoning requirements in medical scenarios. Medical reasoning requires attention to every subtle clue, such as the patient's symptoms and the detailed differences in imaging examinations, and gradually piecing together this information to reconstruct the complete truth of the event. However, in our case, despite the existence of numerous indirect pieces of evidence supporting the diagnosis of AMVT, `DeepSeek-R1` selectively ignored these clues and instead overly relied on the reasoning

Case Information

A man in his mid-60s presented with episodes of palpitations, dyspnea, and dizziness during the last half year before referral. The patient was otherwise healthy. Medical history included treatment with low-dose statin for high cholesterol. A Holter monitor revealed episodes of atrial fibrillation presumed to be the cause of patient symptoms.

Physical Examination

Left ventricular ejection fraction was within normal range. The aortic valve was tricuspid with a mild central regurgitation. The sinus of Valsalva was slightly dilated with a diameter of 41 mm. No sign of left ventricular outflow tract obstruction was present.

Diagnostic Tests

Transthoracic echocardiography (TTE): Revealed a highly mobile element in the left ventricular outflow tract (LVOT). No definite anatomical relationship between the mobile element and either the mitral or aortic valve was identified.
Transesophageal echocardiography (TEE): Confirmed highly mobile element in the LVOT with an estimated length of 25 mm. The element moved toward the aortic valve during systole, while in early diastole it was dragged further back into the LVOT, forming a thin circular structure.
Image Title: Figure 1 Two-dimensional TEE, mid-esophageal long-axis view
Image Description: Demonstrates the highly mobile element in the LVOT Cardiac; CT scan: Revealed a filamentous element located approximately 7 mm below the left aortic coronary cusp; Cardiac MRI: Unable to visualize the LVOT element. Incidentally found mitral annulus disjunction.
Image Title: Figure 3 Dynamic cardiac CT
Image Description: Shows element movement toward open aortic valve during systole (20% of R-R interval) and during diastole (70% of R-R interval)

Final Diagnosis

Accessory mitral valve tissue (AMVT)

Model Response

o3-mini:	(1). Anomalous (accessory) mitral valve chordae traversing the LVOT (2). Papillary fibroelastoma of the LVOT endocardium (3). Lambl's excrescence arising near the aortic valve (4). Vegetation from infective endocarditis (5). Cardiac myxoma
o1:	(1). Papillary fibroelastoma (2). Lambl's excrescence (3). Infective endocarditis (4). Subaortic membrane (5). Fenestrated aortic cusp
DeepSeek-R1:	(1). Papillary fibroelastoma (2). Lambl's excrescence (3). Cardiac myxoma (4). Thrombus (5). Infective endocarditis vegetation
QwQ-32B:	(1). Ruptured Aortic Sinus of Valsalva Aneurysm into LVOT (2). Aortic Root Aneurysm with Intraluminal Flap (3). Left Ventricular Outflow Tract Myxoma (4). Aortic Valve Leaflet Prolapse (5). Left Ventricular Outflow Tract Thrombus
Qwen2.5-Max:	(1). Lambl's Excrescence (2). Cardiac Papillary Fibroelastoma (3). Thrombus in the Left Ventricular Outflow Tract (4). Endocarditis with Vegetation (5). Myxomatous Degeneration of the Mitral Valve
Claude-3.5-Sonnet:	(1). Papillary fibroelastoma (2). Lambl's excrescence (3). Cardiac myxoma (4). Thrombus (5). Infective endocarditis vegetation
DeepSeek-V3:	(1). Lambl's excrescence (2). Fibroelastoma (3). Thrombus (4). Vegetation (5). Papillary fibroma
GPT-4o:	(1). Papillary fibroelastoma (2). Atrial fibrillation-related thrombus (3). Infective endocarditis with valvular vegetation (4). Cardiac myxoma (5). Lambl's excrescence

Figure 12: **A Case Study of *DiagnosisArena*.** Except for `o3-mini`, which successfully provided the correct answer in the top 1, the other models were far from the correct answer. Analyzing `DeepSeek-R1`'s response, we found that, despite numerous indirect pieces of evidence supporting the diagnosis of AMVT, `DeepSeek-R1` selectively ignored these clues and instead overly relied on the reasoning paths of common diseases. `DeepSeek-R1`'s response is shown in Appendix C.

paths of common diseases. This behavior essentially reflects that the model is still dependent on knowledge reproduction to solve problems, rather than deeply grasping and reasoning through the key details.

R1 Response of the Figure 12 Case

⟨think⟩
Okay, let's tackle this case. So the patient is a man in his mid-60s with episodes of palpitations, dyspnea, and dizziness over six months. He's on a statin for high cholesterol but otherwise healthy. The Holter monitor showed atrial fibrillation, which they thought was causing the symptoms.
Looking at the physical exam: normal LVEF, tricuspid aortic valve with mild regurgitation, dilated sinus of Valsalva (41 mm), no LVOT obstruction. The diagnostic tests are key here. TTE and TEE show a highly mobile element in the LVOT, 25mm long. It moves toward the aortic valve in systole and forms a thin circular structure in diastole. The cardiac CT found a filamentous element 7mm below the left coronary cusp. MRI didn't see the LVOT element but found mitral annulus disjunction. The dynamic CT shows movement with the aortic valve opening.
So, possible diagnoses. The mobile structure in LVOT is central. First thought is a cardiac tumor, like a myxoma. But myxomas are usually in the atria. Fibroelastoma? Those are small, mobile, and often on valves. But the TEE didn't link it to the valves. Maybe a papillary fibroelastoma? They can be on the LVOT structures.
Another possibility is a Lambl's excrescence. Those are filamentous, usually on valves, but maybe this one is in the LVOT. They can be mobile and cause embolic events. The filamentous description on CT fits Lambl's.
Also, considering mitral annular disjunction, but that's an incidental finding here. The main issue is the LVOT mass. Could it be a thrombus? But in the LVOT without severe LV dysfunction? Less likely since ejection fraction is normal.
Another thought: a vegetation from endocarditis. But the patient's history doesn't mention fever or other signs of infection. Also, the element is described as filamentous and mobile, which might not fit typical vegetation.
Then, maybe a flail mitral valve leaflet or chordae, but TEE didn't link it to the mitral valve. The movement described—toward the aortic valve in systole and back in diastole—suggests something attached in the LVOT.
Cardiac CT showing filamentous structure below the coronary cusp. Maybe a congenital band or a false tendon? But those are usually benign. However, this one is causing symptoms like palpitations and dizziness, possibly due to intermittent obstruction or triggering arrhythmias (like the atrial fib detected).
Putting it all together: The top differentials would be:
1. Papillary fibroelastoma: common on valves, but can be in LVOT. Highly mobile, filamentous, causing embolic symptoms or arrhythmias. The size (25mm) is a bit large but possible.
2. Lambl's excrescence: smaller, filamentous, typically on valves. Could be in LVOT, leading to similar symptoms.
3. Myxoma: though typically atrial, but rare in LVOT. Mobile and can cause obstruction.
4. Thrombus: less likely given normal LV function, but possible if there's another source.
5. Vegetation: if considering infective endocarditis despite lack of symptoms.
Wait, but the movement pattern and CT findings might make fibroelastoma more likely than Lambl's. Also, the length is 25mm, which is larger than typical Lambl's. Fibroelastomas can be larger. So maybe fibroelastoma is first, then Lambl's.
Another possibility is a ruptured chordae tendineae, but TEE didn't show mitral valve involvement. Or a flail aortic valve structure, but the aortic valve is tricuspid with mild regurgitation.
Mitral annular disjunction is incidental here, so not the main issue.
So top 5:
1. Papillary fibroelastoma
2. Lambl's excrescence
3. Myxoma
4. Thrombus
5. Vegetation
But also, could there be a false tendon or another structure? But those usually don't cause symptoms. The key is the mobile mass in LVOT causing possible embolic events (leading to AFib) or obstruction. Fibroelastoma is the most likely, followed by Lambl's. Myxoma less likely in LVOT. Then thrombus and

vegetation.
⟨/think⟩
1. Papillary fibroelastoma;
2. Lambl's excrescence;
3. Cardiac myxoma;
4. Thrombus;
5. Infective endocarditis vegetation

D Additional Implementation Details

D.1 Prompts

DiagnosisArena Segmenting Prompt 1

Your task is to convert raw text data into clear and well-formatted Markdown files. Please follow the guidelines below:

1. Formatting Adjustments
- Retain the original paragraph content and apply clear Markdown formatting while maintaining the original layout.
- Use appropriate Markdown syntax for headings based on the original heading hierarchy (e.g., '#', '##', '###').

2. Removal of Irrelevant Information
- Remove all references, citations, footnotes, bibliographies, and any content unrelated to the main topic of the article.
- Completely delete citations and footnotes, including their in-text markers.

3. Handling of Images and Tables
- In the main text: Keep all image and table label (e.g., "Figure 1", "Table 1") but do not alter their placement or content.
- Figure Information Section: At the end of the document, add a new section titled '## Figure Information', listing all image and table titles and descriptions in Markdown list format. Ensure their numbering matches the original document.

4. Maintaining Content Integrity
- Ensure that the actual content remains unchanged, preserving the accuracy and completeness of the original text.
- Only perform formatting and cleaning adjustments without modifying the original content.

Make sure the final output adheres to Markdown syntax standards, with clear content, neat formatting, and easy readability.

DiagnosisArena Segmenting Prompt 2

You will receive a medical paper of the type "Case Report." Your task is to **accurately extract** the following four sections and organize them into a JSON-formatted data structure for use in medical exam question design.

The definitions of each section are as follows:

1. **Case Information**: Includes the patient's basic details (e.g., gender, age, occupation), medical history (past medical history, family history, current medical history), and disease progression.
- **Important Requirement**: If the case report explicitly mentions a disease name or diagnostic speculation (e.g., "considering XX disease" or "highly suspected XX"), this information must be removed to avoid directly revealing the final diagnosis.

2. **Physical Examination**: Includes the results of the patient's physical examination during the initial consultation or hospital admission, presenting all clinical signs found in the examination.
- **Prohibited Content**: Any direct diagnostic statements involving disease names.

3. **Diagnostic Tests**: Includes laboratory tests, imaging examinations, pathological examinations, and other auxiliary test results, categorized by test type.
- **Handling of Images and Tables**: - If the case report contains images or tables of imaging studies, lab reports, etc., extract their **titles** and **descriptions** and classify them under the corresponding test category.
- Ensure that imaging studies (e.g., X-ray, CT, MRI), laboratory tests (e.g., blood, urine analysis), and pathological tests are **separately categorized** without mixing.
- **Prohibited Content**: Any direct

diagnostic statements involving disease names.

4. **Final Diagnosis**: The doctor's final diagnosis for the patient, which should concisely and accurately summarize the disease name or diagnostic conclusion.

Output Format Requirements:
- Your response should be a JSON dictionary containing four keys: "Case Information", "Physical Examination", "Diagnostic Tests", and "Final Diagnosis".
- Each key's value should **retain the original wording as much as possible**. If the original text is overly long or unclear, **it may be appropriately condensed for clarity**, but no subjective speculation should be added.
- **Exclude any non-exam-related information**, such as treatment plans, surgical procedures, prognosis, follow-up, etc.

Example Output:
```json
{
"Case Information": "A 45-year-old male patient was admitted due to recurrent chest pain for 3 months. Past medical history includes hypertension for 5 years, and a family history of coronary artery disease.",
"Physical Examination": "Examination: Blood pressure 150/90 mmHg, heart rate 80 bpm, systolic murmur heard at the apex.",
"Diagnostic Tests": "- Laboratory tests: Complete blood count showed no significant abnormalities. Serum biochemistry indicated elevated troponin levels. - Imaging studies: Coronary angiography revealed 70% stenosis in the left anterior descending artery. Image Title: Coronary Angiography Results. Image Description: The angiography shows luminal narrowing in the left anterior descending artery, with an estimated stenosis of 70%.",
"Final Diagnosis": "Coronary atherosclerotic heart disease."
}
```

Medical Paper:

DiagnosisArena Inference Prompt

As a medical expert, please make a diagnosis for the patient's disease based on the case information, physical examination, and diagnostic tests. Please enumerate the top 5 most likely diagnoses for the following patient in order, with the most likely disease listed first.

Case Information:
%s

Physical Examination:
%s

Diagnostic tests:
%s

Output the diagnosis in numeric order, one per line. For example:
1. Disease A;
2. Disease B;
...

DiagnosisArena-MCQ Inference Prompt

You are an expert in the field of rare diseases. You will receive a medical case file, including Case Information, Physical Examination and Diagnostic Tests. Please conduct a thorough analysis based on the provided information and select the most appropriate diagnosis from the following four options. The final answer must be formatted as \boxed{Correct Option Letter}.

Here is the medical case file:
Case Information:
%s

Physical Examination:
%s

Diagnostic Tests:
%s

Here are the four options:
%s

Output Format: Output the final answer in the

following format:
Final answer: \boxed{Correct Option Letter}

GPT-4o Evaluation Prompt

You are an expert in diagnosing challenging cases. You will receive a student's answer containing 5 differential diagnoses, as well as the reference diagnosis. You need to score each diagnosis from the student's answer according to the following rules:

2 = The student's diagnosis exactly matches the reference diagnosis;
1 = The student's diagnosis is a broad category that includes the reference diagnosis;
0 = The student's diagnosis does not meet the criteria for a score of 1 or 2.

Here is the student's answer:
%s

Here is the reference diagnosis:
%s

Output Format: Output the scores in the following format.
1. Disease 1 name: \\boxed{The Score of Disease 1};
2. Disease 2 name: \\boxed{The Score of Disease 2};
...

Model	Top1	Top5	Raw Scores	Response Preview
GPT-5	✔	✔	2 1, 0, 0, 0	1. Kaposiform hemangioendothelioma involving the left cerebellopontine angle/internal auditory canal; 2. Tufted angioma (angioblastoma), on the kaposiform hemangioendothelioma–tufted angioma spectrum; 3. Kaposiform lymphangiomatosis localized to the cerebellopontine angle/internal auditory canal; 4. Infantile capillary hemangioma of the internal auditory canal/cerebellopontine angle; 5. Lymphatic malformation (lymphangioma) of the cerebellopontine angle/internal auditory canal;
o3	✔	✔	2 0, 1, 0, 0	1. Kaposiform hemangioendothelioma (vascular tumor of infancy involving the cerebellopontine angle/internal auditory canal); 2. Tufted angioma (glomeroid/"cannon-ball" capillary angioma with similar biphasic, glomeruloid vascular nodules); 3. Infantile (capillary/cavernous) hemangioma of the cerebellopontine angle; 4. Hemangioblastoma presenting congenitally in the posterior fossa; 5. Epithelioid hemangioendothelioma (low-grade malignant vascular neoplasm).
Gemini-2.5-Pro	✔	✔	2 0, 0, 0, 0	1. Kaposiform Hemangioendothelioma 2. Infantile Hemangioma 3. Infantile Hemangiopericytoma / Solitary Fibrous Tumor 4. Atypical Teratoid/Rhabdoid Tumor 5. Complex Vascular Malformation / Angiomatous Lesion
o1	✘	✘	0 0, 0, 1, 0	1. Infantile Hemangioma; 2. Congenital Hemangioma; 3. Hemangiopericytoma; 4. Epithelioid Hemangioendothelioma; 5. Lymphangioma.
Qwen3-235B -A22B-2507	✘	✘	0 0, 0, 0, 0	1. Congenital hemangioendothelioma of the cerebellopontine angle; 2. Infantile myofibromatosis involving cranial nerves VII and VIII; 3. Vascular schwannoma with endothelial differentiation; 4. Lymphangiomatosis with atypical vascular proliferation; 5. Capillary hemangioma of the internal auditory canal.
Deepseek-V3.1	✘	✘	0 1, 0, 0, 0	1. Infantile hemangioma; 2. Hemangioendothelioma; 3. Lymphangioma; 4. Schwannoma; 5. Neurofibroma;
Deepseek-R1-0528	✘	✔	0 2, 1, 0, 0	1. Infantile hemangioma; 2. Kaposiform hemangioendothelioma; 3. Tufted angioma; 4. Lymphangiomatosis; 5. Venous malformation

Table 2: Experimental analysis of model performance disparities: A comparative example of outputs across different models.

Journal Name	Number
JAMA	488
Lancet	26
Annals of Internal Medicine	192
Cell	254
European Respiratory Journal	18
European Urology	4
Gastroenterology	14
Journal of Hepatology	6
Journal of Thoracic Oncology	24
NEJM	87

Table 3: The final distribution of Journal Names and their respective Numbers in *DiagnosisArena*.