ASCENT: A Benchmark for Evaluating and Advancing Stepwise Diagnostic Reasoning in Large Language Models on Common Clinical Scenarios

Yera Choi YERA.CHOI@NAVERCORP.COM
NAVER Applied AI Research, Republic of Korea

Yeong Hwa Kim* YHKIMMAY@EWHAIN.NET
Ewha Womans University Mokdong Hospital, Republic of Korea

JaeDeok Lee* JAEDEOK.D.LEE@GMAIL.COM
NAVER Applied AI Research, Republic of Korea
Healthcare AI Research Institute (HARI), Seoul National University Hospital, Republic of Korea

Taekang Kim* TAEKANG@YONSEI.AC.KR
Yonsei University Wonju College of Medicine, Republic of Korea

Sangdoo Yun SANGDOO.YUN@NAVERCORP.COM
NAVER AI Lab, Republic of Korea

Seong-Eun Moon* SEONGEUN.MOON89@GMAIL.COM
NAVER Applied AI Research, Republic of Korea

Abstract

Large language models (LLMs) excel at medical question answering yet are rarely evaluated on the *stepwise* diagnostic reasoning that defines real clinical workflows, where impressions are revised as information accumulates. We build **Annotated Stepwise Clinical rEasoning for NaturalisTic Diagnosis (ASCENT)**, a clinician-annotated benchmark and training resource of 3,078 stepwise problems derived from MedQA-USMLE that decomposes each vignette into EMR-aligned steps (Findings, Impression, supporting Rationale), enabling evaluation of intermediate reasoning under incomplete information.

Experiments and training with ASCENT revealed insights into how current LLMs handle stepwise diagnostic reasoning. Even strong reasoning models that perform well on MedQA-USMLE leave substantial headroom on ASCENT, and general-purpose frontier models trail further—exposing a persistent gap between fully informed and stepwise diagnosis. Fine-tuning Qwen2.5-7B and 32B on ASCENT yields measurable F1 gains over both pretrained and HuatuoGPT-o1 CoT-trained baselines, with gains driven primarily by precision. Complementary robustness analyses (counterfactual perturbation, format-vs-content control, judge agreement, and rollout) further show that ASCENT-fine-tuned models rely on the diagnostic content of prior impressions rather than imitating their output format, while error propagation under rollout remains a key challenge for clinical deployment.

Data and Code Availability This paper uses the MedQA-USMLE benchmark dataset, which is available on the HuggingFace Datasets repository (Jin et al., 2020). The proposed dataset and code accompanying the paper will be made publicly available[1].

Institutional Review Board (IRB) The current research does not require IRB approval, as it utilized only publicly available medical license examination materials.

1. Introduction

Recently, large language models (LLMs) have been progressively utilized in the medical domain and have

* Work done while at NAVER Applied AI Research (formerly NAVER Digital Healthcare Lab).

1. The data and code will be released at https://github.com/naver-ai/ascent.

demonstrated remarkable performance on tasks including medical question answering (QA) (Singhal et al., 2023) and differential diagnosis (Savage et al., 2023). With the advent of reasoning models that leverage test-time scaled Chain-of-Thought (CoT) prompting (Wei et al., 2022) and effective learning strategies incorporating reinforcement learning (RL), LLMs are even considered capable of generating complex clinical reasoning and solving complex diagnostic problems (Chen et al., 2024).

Nevertheless, the application of model-generated clinical reasoning outcomes to real-world clinical settings has been limited, which is in part due to the difficulty of assessing their clinical validity and effectiveness (Meng et al., 2024). Unlike mathematical or coding problems, which generally allow logical step-by-step formulation and automatic verification, medical problems are difficult to convert into clear step-by-step structures and require expert involvement for rigorous evaluation of solutions (Wu et al., 2025a).

Accordingly, although still in the early stages, existing works have attempted to ground the generation of clinical reasoning through LLMs, medical knowledge resources, or human experts, and increasingly emphasize the importance of stepwise verification to create valid clinical reasoning paths. HuatuoGPT-o1 and ReasonMed rely on LLM-based medical verifiers (Chen et al., 2024; Sun et al., 2025), while MedReason implements a medical knowledge graph to create reasoning chains (Wu et al., 2025a). MedCaseReasoning focuses on clinician–model alignment in case studies (Wu et al., 2025b).

A more recent study, SD Bench (Nori et al., 2025), highlights a common shortcoming of prior approaches–their reliance on complete patient vignettes that include all available information for diagnosis. Such approaches tend to overlook the importance of "sequential diagnosis," a crucial component of real-world clinical decision-making (Daniel et al., 2019). While this study draws important attention to the need for stepwise reasoning, it, along with other recent works, primarily focuses on complex and challenging case studies. Although interpreting such difficult cases is valuable, such a focus may lead to overlooking the need to assess models' capabilities in commonly encountered scenarios. Moreover, SD Bench does not explicitly evaluate intermediate reasoning outputs under uncertainty, leaving LLMs' stepwise reasoning capacity underexplored.

To support the development and evaluation of stepwise clinical reasoning, we propose **Annotated Stepwise Clinical rEasoning for Naturalis-Tic Diagnosis (ASCENT)**, a clinician-annotated dataset derived from the widely used MedQA-USMLE benchmark (Jin et al., 2020). ASCENT targets a gap that prior datasets leave open: although diagnostic reasoning is crucial for accurate clinical decision-making, its core cognitive processes are typically acquired implicitly rather than explicitly taught (Graber et al., 2005), so expert reasoning in common scenarios is rarely articulated in a stepwise manner—precisely the regime LLMs need to learn.

ASCENT structures each case incrementally along sections commonly found in Electronic Medical Records (EMRs)—such as Present Illness and Physical Examination—approximating the natural flow of information at the bedside. At every step, clinicians annotate the Findings, the Impression (suspected disorders given the information so far), and the Reason (the rationale supporting the Impression). This dual annotation of both structured diagnostic targets and rationales enables fine-grained, intermediate evaluation that goes beyond final-answer correctness, and—because ASCENT inherits from MedQA-USMLE—covers a broad range of commonly encountered, representative clinical scenarios that are often underexplored despite forming the basic foundation of medical reasoning.

We further demonstrate that training on ASCENT yields substantial empirical gains. Models ranging from 7B to 32B in size exhibit an F1 score improvement of 0.05–0.11 over their pre-trained baselines. Moreover, compared to models trained on complex CoT annotations from HuatuoGPT-o1, models trained on ASCENT outperform by 0.04–0.11 F1, suggesting that ASCENT effectively complements existing datasets by recovering basic but essential reasoning steps that are often overlooked.

The key contributions of the current study are as follows:

- We build and release **ASCENT**, a benchmark dataset of structured patient vignettes segmented into **intermediate diagnostic steps** with **stepwise evaluation targets**. Each step includes clinician-verified findings, suspected impressions, and supporting rationales, offering a transparent view of the stepwise clinical reasoning process.

- ASCENT reflects the structure of real-world EMR sections (Present Illness, Physical Exam-

ination, and so on), approximating the natural flow of information in clinical settings.

- Unlike previous datasets that focus primarily on complex or rare cases, ASCENT emphasizes **common and exemplar clinical scenarios**, which are often governed by implicit reasoning and underrepresented in prior work.
- By reconstructing these implicit decisions into explicit, interpretable reasoning chains, ASCENT offers a framework to enhance and evaluate stepwise diagnostic reasoning in LLMs.

2. Related Work

LLM-based clinical reasoning and differential diagnosis. LLMs have shown increasing potential for clinical reasoning, with early efforts exploring how specialized optimization can enhance diagnostic performance. **Articulate Medical Intelligence Explorer (AMIE)** (McDuff et al., 2025) represents one of the first attempts to tailor LLMs for diagnostic reasoning, demonstrating strong performance on 302 challenging case-study problems. Brodeur et al. (2024) report a complementary finding that a frontier general-purpose LLM matches or exceeds physician performance on standardized diagnostic-reasoning tasks, further motivating fine-grained evaluation of how this capability transfers to stepwise clinical workflows.

Subsequent research has advanced long-form medical reasoning in LLMs by improving reliability and stepwise explainability. **HuatuoGPT-o1** (Chen et al., 2024) built 40K verifiable exam-style problems and applied supervised fine-tuning (SFT) with verifier-based reinforcement learning with verifiable rewards (RLVR), while **m1** (Huang et al., 2025) showed that test-time scaling alone can boost sub-10B models to rival larger ones, though excessive "overthinking" can hurt accuracy. Complementary efforts such as **MedReason** (Wu et al., 2025a) and **ReasonMed** (Sun et al., 2025) focus on intermediate reasoning chains—via knowledge-graph paths or multi-agent verification—yielding improved faithfulness and enabling sub-10B models to approach or surpass much larger models. Together, these studies highlight the central role of curated step-by-step reasoning data in developing trustworthy medical LLMs.

Complementing training-focused efforts, recent benchmarks emphasize sequential diagnosis and clinician-aligned reasoning. **SD Bench** (Nori et al., 2025) transforms 304 cases from the New England Journal of Medicine Clinicopathological Conferences (NEJM-CPC) into interactive stepwise encounters; **MedCaseReasoning** (Wu et al., 2025b) offers 14K QA cases with clinician-authored reasoning to assess diagnostic accuracy and reasoning fidelity; **MedThink-Bench** (Zhou et al., 2025) provides 500 expert-annotated cases with an LLM-as-a-Judge framework for fine-grained evaluation; and **HealthBench** (Arora et al., 2025) introduces 5,000 multi-turn conversations evaluated against physician-authored rubrics for open-ended healthcare interactions, including settings where the model must elicit missing patient information through follow-up questions, although its scope spans a broader range of healthcare interactions. While these efforts substantially broaden evaluation coverage, no existing benchmark systematically evaluates intermediate reasoning steps under conditions of clinical uncertainty in common diagnostic scenarios. Table 1 compares datasets focused on medical reasoning within diagnostic scenarios.

Dataset	Expert Review	GT Rationale	Intermediate Target
HuatuoGPT-o1	✗	✓	✗
m1	✗	✓	✗
MedReason	△	✓	✗
ReasonMed	✗	✓	✗
MedCaseReasoning	✓	✓	✗
SD Bench	✓	✗	✗
MedThink-Bench	✓	✓	✗
`ASCENT` (ours)	✓	✓	✓

Table 1: Comparison between datasets. **Expert Review**: Whether the dataset's annotations were reviewed by medical experts. **GT Rationale**: Whether the dataset provides ground-truth explanations. **Intermediate Target**: Whether the dataset includes intermediate reasoning targets for sequential diagnostic evaluation. In the case of MedReason, △ under Expert Review indicates partial evaluation, where over 25 sampled questions across seven specialties were evaluated by directly comparing anonymized HuatuoGPT-o1 and MedReason Chain-of-Thought data.

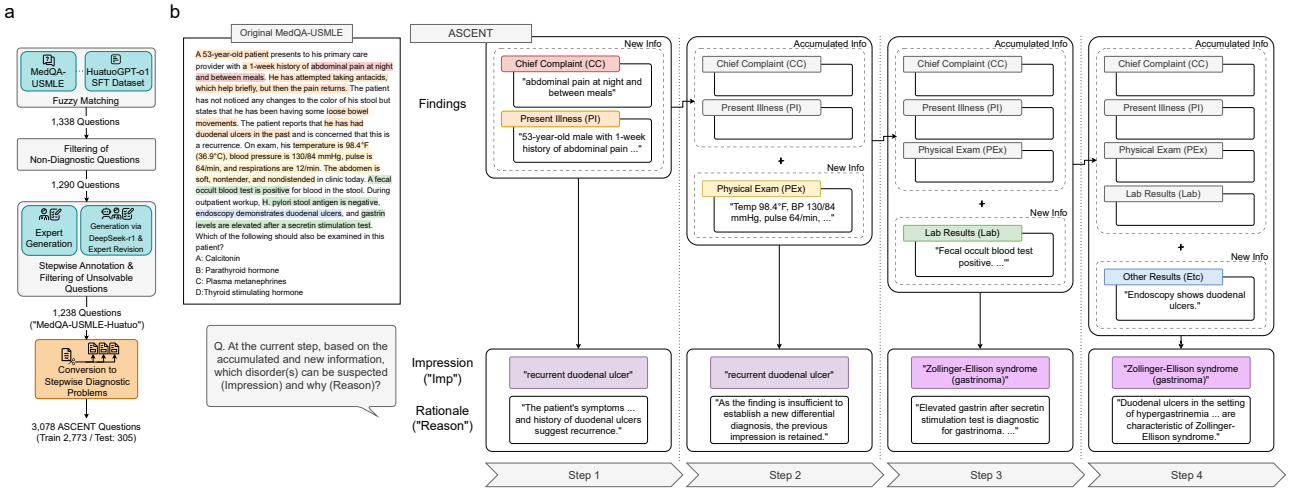

Figure 1: **a**. Data construction workflow of ASCENT. Candidate questions from MedQA-USMLE and the HuatuoGPT-o1 SFT dataset were filtered and stepwise-annotated by experts, either directly or via DeepSeek-R1 with expert revision. The finalized questions were converted into stepwise diagnostic problems across clinical stages. **b**. The structure of ASCENT data. At each step, clinical findings are paired with a list of suspected disorders (**Impression**) and the corresponding rationale (**Reason**). As steps progress and new clinical information accumulates, the Impression is refined or updated to reflect the evolving diagnostic reasoning. In evaluations with ASCENT, each step is treated as a separate question, and models are prompted to generate either only the Impression or both the Impression and the corresponding Reason.

3. ASCENT

3.1. Structured design for Stepwise Diagnosis

Figure 1-b illustrates the stepwise structure of the ASCENT dataset, including the conversion of MedQA-USMLE questions into ASCENT data samples. To emulate a more naturalistic flow of information in clinical settings, the content of each original MedQA-USMLE question was segmented and classified into one of the following EMR sections.

- **Chief Complaint (CC)** briefly states the primary symptom that serves as the reason for seeking medical care, as reported by the patient.
- **Present Illness (PI)** provides a concise narrative of the patient's chief complaint, describing their onset, course, and associated features, while incorporating relevant contextual information such as age and gender.
- **Physical Examination (PEx)** consists of objective findings obtained by the clinician, typically through direct inspection, palpation, percussion, or auscultation.
- **Laboratory Results (Lab)** include outcomes from laboratory tests performed on biological specimens (e.g., blood, urine tests, or culture tests).
- **Imaging Results (Img)** present interpretative findings derived from radiological studies (e.g., X-ray, computed tomography (CT), or magnetic resonance imaging (MRI)).
- **Other Results (Etc)** encompass additional diagnostic and procedural findings not covered in the previously mentioned sections (e.g., electrocardiography (ECG), endoscopic evaluations, and biopsy results), as well as the post-treatment status and planned management.

Formally, each clinical case $\mathcal{C}$ derived from a medical exam question is represented as an ordered sequence of steps:

$$\mathcal{C} = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n],$$

where each step $\mathbf{s}_t$ consists of:

- f_t: clinical findings newly acquired at step t corresponding to the information from one of the predefined EMR sections,
- i_t: impression (suspected diagnosis or system) at step t based on the collected clinical information through previous and current steps $f_0, \ldots, f_t$,
- r_t: rationale that supports the impression i_t.

The first step $\mathbf{s}_0$, corresponding to the chief complaint (CC), includes only clinical findings f_0; both impression i_0 and rationale r_0 are defined as `None`, as this section is composed of briefly stated symptoms. As the CC step includes limited information, we construct segments only for steps $t \geq 1$. For each such step t, a diagnostic problem is constructed as:

- **Stepwise Diagnostic Context** X_t: the accumulated findings, impressions, and rationales from steps 0 through $t-1$, along with the current step's findings:

$$X_t = \left[\bigcup_{i=0}^{t-1} (f_i,\ i_i,\ r_i,\ ),\ f_t \right].$$

- **Target** Y_t: depending on the task setup, the target can be one or more of the following: i_t, r_t.

This format ensures that the model predicts the impression (and/or the rationale) at each stage using only the information available up to that point, without access to the current step's impression or rationale.

3.2. Data Generation

Problem Filtering. For direct comparison with HuatuoGPT-o1 and its Supervised Fine-Tuning (SFT) dataset, we retained only the MedQA-USMLE (Jin et al., 2020) problems also present in the HuatuoGPT-o1 SFT dataset, identified via fuzzy string matching with Python's FuzzyWuzzy library at a similarity threshold of 0.85. To remove questions that do not require diagnostic reasoning, we filtered out those containing keywords related to research or experimental studies (e.g., "experiment," "epidemiologist," "cohort") and quantitative terms (e.g., "probability"), followed by manual inspection. In total, 48 of the 1,338 questions were excluded by the filtering. Among the filtered questions, simple knowledge-based questions that did not present clinical scenarios (e.g., items asking solely about pharmacology mechanisms or basic anatomy without a patient context) were excluded. In addition, questions requiring interpretation of the accompanying images were also omitted due to restricted accessibility to the images. The resulting dataset containing 1,238 questions, referred to as "MedQA-USMLE-Huatuo," was then subjected to stepwise annotation described in the following sections.

Stepwise Diagnostic Problems Generated by Experts. Four board-certified physicians were recruited to annotate the MedQA-USMLE-Huatuo dataset into the ASCENT structure, which segments each case into sequential clinical steps and pairs the findings at each step with corresponding impressions and rationales. Along with the dataset, the physicians were provided with the raw outputs from a Llama-3.1-8B-Instruct model fine-tuned on a separate in-house dataset in the same structure. They were guided to generate sequentially structured data in accordance with the dataset design. After the initial annotation, two separate reviewers with medical expertise (a board-certified physician, Y.H.K., and a medical student, T.K.), along with another reviewer (Y.C.) to address formatting and consistency, thoroughly reviewed and revised the outcome. Overall, more than half of the dataset (680 questions) were annotated during this process; the corresponding edit distribution is reported in Appendix C.6.

DeepSeek-R1-Based Generation. The remaining part of the MedQA-USMLE-Huatuo dataset was also used to generate the ASCENT dataset. Specifically, the DeepSeek-R1 model (DeepSeek-AI, 2025) was provided with a detailed instruction to classify the sections and generate the impression and rationales for 558 questions, based on the corresponding HuatuoGPT-o1 complex CoT. The prompt template for DeepSeek-R1-based generation is shown in Appendix B.

The generation outcomes were then meticulously reviewed and revised by two reviewers with medical expertise (Y.H.K., a board-certified physician, and T.K., a medical student), along with another reviewer (Y.C.) to examine and correct formatting issues. All DeepSeek-R1-generated cases received independent review by Y.H.K. and T.K., with Y.H.K. thoroughly reviewing all of T.K.'s revisions; no case bypassed clinical review.

The DeepSeek-R1 drafts required substantially less revision than the Llama-3.1-8B drafts used for the expert-annotated subset: 57.1% no edit and 17.5% major edits (similarity <0.85) for DeepSeek-R1, versus 97.7% major edits for the 680 Llama-3.1-8B drafts. Full provenance is reported in Appendix C.6.

Construction into Stepwise Diagnostic Problems. After annotation, each problem was segmented into multiple stepwise diagnostic problems, where each step consists of a Stepwise Diagnostic Context (X_t) and a corresponding Target (Y_t). If there existed a preceding Impression but the current Findings were not decisive enough to change it, the rationale was filled in with "As the finding is insufficient to establish a new differential diagnosis, the previous impression is retained." The resulting ASCENT dataset comprises 3,078 questions. Both the expert-generated and the DeepSeek-R1-generated, expert-reviewed subsets were randomly divided at the level of the original MedQA-USMLE questions in a 9:1 training-to-test ratio (seed=3407), yielding 2,773 training and 305 test questions. Figure 1-a provides an overview of the data generation workflow.

4. Experiment

To evaluate the clinical reasoning capabilities of LLMs in stepwise diagnostic scenarios, we investigate how different training regimes affect performance on the ASCENT benchmark. We also examine the role of model scale, motivated by the observation that larger LLMs tend to exhibit stronger few-shot and CoT capabilities.

4.1. Experimental Settings

Models. We employed a diverse set of LLMs spanning both open-source and proprietary models, as well as reasoning-augmented models for the comprehensive evaluation of stepwise reasoning abilities. Specifically, our experiments include Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct, Qwen3-8B, Qwen3-32B, HuatuoGPT-o1-7B, and DeepSeek-R1 (*open-source models*), alongside OpenAI GPT-4o, o3-mini, and GPT-5.2 (*closed-source models*). GPT-5.2 was included to assess potential saturation under a strong proprietary general-purpose model. Note that Qwen3-8B, Qwen3-32B, HuatuoGPT-o1-7B, DeepSeek-R1, OpenAI o3-mini, and GPT-5.2 were evaluated using their thinking modes, which are designed to enhance reasoning.

In addition, we fine-tuned Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct on two counterpart datasets derived from the same MedQA-USMLE questions: (1) the ASCENT training set and (2) the MedQA-USMLE-Huatuo dataset, which contains complex CoT data extracted from the HuatuoGPT-o1 SFT dataset by matching with MedQA-USMLE. All models were fine-tuned for three epochs on eight NVIDIA H100 (80 GB) GPUs with a batch size of 16, using the AdamW optimizer and cosine learning rate scheduler with the learning rate of 1e-5, BF16 precision, and weight decay of 0.01.

Inference Settings. The maximum generation length was set to 1,024 tokens for standard models and 4,096 tokens for reasoning models during the evaluation. This restriction of generated token length is based on the findings in Huang et al. (2025), which report that excessively long generation lengths (i.e., over 4K tokens) can induce overthinking and degraded reasoning performance for problems in the medical domain. We further validate this choice via a token-length ablation in Appendix C.3. Apart from the maximum generated token length, all experiments adopted the following inference settings for fair and reproducible comparison: temperature 0.6, top-P 0.95, top-K 20, presence penalty 1.0, frequency penalty 0.0, and minimum probability 0.0. All experiments were run on eight H100 GPUs, while proprietary models were evaluated via their commercial APIs.

Experimental Tasks for ASCENT. As the ASCENT dataset provides two types of stepwise outputs—impressions and their supporting rationales—we conducted corresponding experiments on two tasks: (1) to generate impressions only (***Imp***) and (2) to generate supporting rationales followed by impressions (***Imp + Reason***). The prompt templates for *Imp* and *Imp + Reason* tasks can be found in Appendix B.

4.2. Evaluation Metrics

Model-generated impressions at each step are compared against the ground-truth impressions in ASCENT for evaluation. As both the predicted and ground-truth impressions can include multiple items, we use the micro-averaged F1 score as our evaluation metric.

For better understanding of model behavior under insufficient clinical context, we categorized impres-

Targets	Models	Step 1	Step 2	Step 3+	Overall	95% CI
Imp	**Small Models (<10B)**					
	Qwen2.5-7B-Instruct	0.22	0.49	0.61	0.40	[0.35, 0.45]
	+ SFT w/ MedQA-USMLE-Huatuo	0.23	0.53	0.65	0.43	[0.38, 0.48]
	+ SFT w/ ASCENT (ours)	**0.28**	**0.63**	**0.69**	**0.50**	[0.44, 0.55]
	HuatuoGPT-o1-7B	**0.33**	0.59	0.69	0.50	[0.45, 0.55]
	Qwen3-8B	0.29	**0.62**	**0.73**	**0.51**	[0.45, 0.56]
	Large Models (≥10B)					
	OpenAI GPT-4o	0.43	0.43	0.66	0.49	[0.44, 0.54]
	OpenAI GPT-5.2	**0.47**	0.56	0.57	0.53	[0.48, 0.57]
	Qwen2.5-32B-Instruct	0.27	0.51	0.65	0.43	[0.38, 0.48]
	+ SFT w/ MedQA-USMLE-Huatuo	0.27	0.48	0.68	0.43	[0.38, 0.49]
	+ SFT w/ ASCENT (ours)	0.35	**0.66**	**0.70**	**0.54**	[0.49, 0.60]
	Qwen3-32B	0.35	0.59	0.62	0.49	[0.44, 0.54]
	DeepSeek-R1	0.56	0.70	0.75	0.65	[0.60, 0.70]
	OpenAI GPT-o3-mini	**0.66**	**0.74**	**0.76**	**0.72**	[0.67, 0.76]
Imp + Reason	**Small Models (<10B)**					
	Qwen2.5-7B-Instruct	0.31	0.58	**0.71**	0.48	[0.43, 0.53]
	+ SFT w/ MedQA-USMLE-Huatuo	**0.34**	0.56	0.70	0.50	[0.44, 0.55]
	+ SFT w/ ASCENT (ours)	0.32	**0.70**	0.69	**0.54**	[0.49, 0.59]
	HuatuoGPT-o1-7B	0.37	0.63	0.71	0.54	[0.48, 0.59]
	Qwen3-8B	**0.41**	**0.68**	**0.78**	**0.59**	[0.54, 0.65]
	Large Models (≥10B)					
	OpenAI GPT-4o	0.39	0.55	0.67	0.51	[0.46, 0.56]
	OpenAI GPT-5.2	**0.53**	0.62	0.58	0.57	[0.52, 0.62]
	Qwen2.5-32B-Instruct	0.42	0.65	0.76	0.58	[0.53, 0.63]
	+ SFT w/ MedQA-USMLE-Huatuo	0.39	0.63	0.76	0.56	[0.51, 0.61]
	+ SFT w/ ASCENT (ours)	0.44	**0.74**	**0.81**	**0.63**	[0.58, 0.68]
	Qwen3-32B	0.44	0.66	0.76	0.58	[0.53, 0.63]
	DeepSeek-R1	0.51	0.70	0.74	0.63	[0.58, 0.68]
	OpenAI GPT-o3-mini	**0.58**	**0.75**	**0.77**	**0.69**	[0.64, 0.74]

Table 2: Stepwise diagnosis performance of open-source and proprietary models on the ASCENT test set ($n = 305$), evaluated on two tasks: (1) generating impressions only (***Imp***) and (2) generating supporting rationales followed by impressions (***Imp + Reason***). Step columns report micro-F1 per stage ("Step3+" aggregates Step 3 onward); the "Overall" column reports micro-F1 across all steps. The "95% CI" column reports bootstrap 95% confidence intervals on Overall micro-F1, computed via 10,000 test-record resamples (percentile method); under *Imp*, the top reasoning cluster (o3-mini, DeepSeek-R1) is statistically separable from the lower cluster with non-overlapping CIs, whereas under *Imp + Reason* the second-tier models cluster more closely and several CIs overlap (notably Qwen2.5-32B-Instruct + SFT w/ ASCENT and DeepSeek-R1). Models are grouped by size and whether they are specialized in reasoning, with bold numbers indicating the best result within each model category.

sion outputs into four types: (1) **Correct Match** (e.g., ground truth: *intrauterine growth restriction*; prediction: *fetal growth restriction*), (2) **Overly Specific** (e.g., ground truth: *respiratory disorder*; prediction: *Chronic Obstructive Pulmonary Disease (COPD)*), (3) **Overly Broad** (e.g., ground truth: *lupus nephritis*; prediction: *glomerulonephritis*), and (4) **Others (Unrelated or Irrelevant)** (e.g., *essential tremor*; prediction: *Parkinson's disease*), where the latter three indicate varying degrees of deviation from the ground-truth impression. Model outputs were evaluated using OpenAI GPT-4.1 (see Appendix B for the full prompt; we additionally assess inter-judge agreement with GPT-5.2 in Appendix A), which classified each predicted impression into one of these categories. These labels were subsequently used to compute the F1 scores against the ground-truth entities and to analyze failure cases.

4.3. Results

Main Results Table 2 presents the micro-averaged F1 score for impressions. Notably, the well-known LLMs, such as OpenAI's models and DeepSeek-R1, still underperform relative to their impressive results on the original MedQA-USMLE dataset. Models fine-tuned with conventional clinical CoTs, i.e., Qwen2.5 + SFT w/ MedQA-USMLE-Huatuo, also offer only marginal gains.

In contrast, models trained with ASCENT consistently outperformed their counterparts trained on the MedQA-USMLE-Huatuo dataset, highlighting the value of structured, stepwise supervision for diagnostic reasoning. For example, Qwen2.5-7B-Instruct fine-tuned on ASCENT achieved an F1 score of 0.50, surpassing the MedQA-USMLE-Huatuo counterpart (0.43). When prompted to generate rationales first (***Imp + Reason***), the F1 score further improved to 0.54 versus 0.50 for the MedQA-USMLE-Huatuo version. These improvements reflect that the stepwise, incremental structure of ASCENT helps models draw candidate diagnoses from insufficient patient information, whereas models trained on the entire vignette show limited performance under stepwise evaluation.

Model size also contributed significantly to performance gains for Qwen2.5-Instruct models. Larger models consistently achieved higher F1 scores, especially with additional training on ASCENT. Qwen2.5-32B-Instruct fine-tuned on ASCENT reached an F1 score of 0.63 with rationale generation, clearly outperforming its 7B counterpart (0.54) with a large gap even compared to the other experiment cases. Across both 7B and 32B fine-tuning, F1 gains were driven primarily by precision: precision improved by 0.11–0.17 over pre-trained baselines while recall stayed essentially flat (Table 6, Appendix C), indicating that ASCENT fine-tuning encourages more focused, clinically appropriate differential diagnoses rather than broader candidate enumeration. The Qwen3 series, however, violates this scaling pattern: Qwen3-32B underperforms Qwen3-8B due to consistently more *Overly Broad* errors at every step (Appendix C.2). We hypothesize that the larger model's broader knowledge base produces more candidate disorders without a commensurate ability to narrow them under early-step uncertainty.

Reasoning-oriented "thinking models" outperformed their non-thinking counterparts even without task-specific fine-tuning, with the largest gains observed for OpenAI o3-mini and DeepSeek-R1; the Qwen3 series showed more modest improvements. OpenAI o3-mini achieved the highest overall F1 score (0.72). Interestingly, rationale generation (***Imp + Reason***) did not consistently benefit all thinking models; while OpenAI o3-mini and DeepSeek-R1 exhibited slight drops (from 0.72 to 0.69 and from 0.65 to 0.63, respectively), Qwen3-8B and Qwen3-32B showed improvement with rationale generation (from 0.51 to 0.59 and from 0.49 to 0.58, respectively). This suggests that extended reasoning can sometimes induce overthinking or misalignment between reasoning and final answers.

Notably, even GPT-5.2—a strong proprietary general-purpose model—achieves only 0.53 / 0.57 (*Imp* / *Imp + Reason*), above GPT-4o (0.49 / 0.51) in point estimates but well below o3-mini in both modes (with non-overlapping 95% CIs); the gap to DeepSeek-R1 is evident under *Imp* (non-overlapping CIs), although their CIs overlap under *Imp + Reason* (Table 2). Reasoning-specialized models consistently outperform stronger general-purpose ones, suggesting that stepwise diagnostic reasoning rewards structured reasoning supervision more than raw capability, and that ASCENT remains far from saturation.

Overall, these findings demonstrate that ASCENT enhances diagnostic reasoning in both small and large models, scales effectively with model size, and narrows the gap with inherently reasoning-oriented models, while also highlighting that the impact of rationale generation can vary across models.

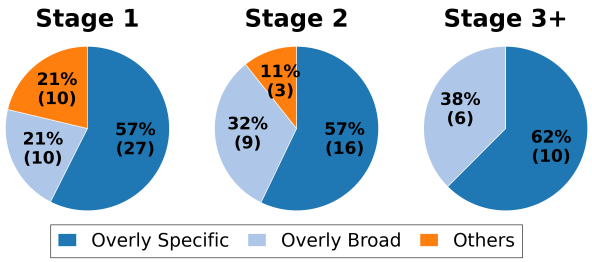

Figure 2: Error type analysis for OpenAI o3-mini on the *Imp* task. Chart labels show the percentage of each error type, with corresponding case counts in parentheses.

Stepwise Analysis The diagnostic performance was further investigated in a stepwise manner, with the outcomes presented in the Step columns of Table 2. We observe that, in contrast to human healthcare professionals, who can readily capture early clinical clues to infer candidate diagnoses, LLMs generally perform poorly in the early stages, where clinical evidence is sparse and uncertain, underscoring the need for sequential diagnostic training and intermediate-output evaluation. Since our setting is a mitigated scenario in which models receive ground-truth context, this gap is even more pronounced under realistic rollout conditions where errors propagate—an intuition confirmed by the rollout experiment in the Robustness Analysis section to follow.

Our results also indicate that training with structured, stepwise-annotated data can help address these shortcomings, particularly in the early stages. Models trained on ASCENT demonstrate improved accuracy at initial diagnostic steps, where less information is available, and *Imp + Reason* training further improves later-stage performance (up to F1 0.81 at Step 3+ for ASCENT-trained Qwen2.5-32B-Instruct).

Failure Case Analysis Figure 2 depicts the distribution of error types for the OpenAI o3-mini results on the ***Imp*** task. Here, OpenAI o3-mini was selected as the representative case because it achieved the best performance among the evaluated models, while other models exhibited similar error distribution trends. Overall, the absolute number of errors in each category tends to decrease as more clinical context accumulates across steps.

This decline is most pronounced for the Others error type, where the sharp drop across diagnostic steps indicates that highly irrelevant predictions—those that fail to meaningfully leverage the given clinical context—become substantially less frequent as more information is obtained. This trend further highlights the difficulty faced by LLMs in generating appropriate diagnoses when patient information is insufficient. This finding may also be partly explained by the structure of the clinical scenarios in MedQA-USMLE, where later-stage information inherently reflects preceding outcomes because subsequent evaluations and tests are often selectively performed based on the diagnostic hypotheses formed in earlier stages.

Qualitative Evaluation on Model-Generated Rationales Beyond the model-judged error categories above, we additionally conducted a small-scale qualitative review of rationale quality with a clinician rater. The case study covers 20 problems, evaluating their generated rationales on a 3-point Likert scale based on three criteria: (1) **Comprehensiveness and Relevance**; (2) **Clinical Validity**; (3) **Soundness of Differential Diagnosis** (see Appendix D for the detailed scoring rubric). A board-certified physician (Y.H.K.) developed the criteria and served as the rater. The rater, provided with the ground-truth rationales, remained blinded to the identities of the models. We treat this evaluation as indicative rather than conclusive given its limited scale.

Figure 3 illustrates the stepwise assessment results for representative cases, while the full results are available in Appendix E. All models achieved average scores of 2.0 or higher across all reasoning steps and evaluation criteria. However, distinct patterns of stepwise performance were observed. Most models demonstrate performance gains as diagnostic steps progress, most notably when moving from Step 1 to Step 2. This jump likely stems from Step 1 including only patient-provided information, whereas Step 2 typically incorporates more detailed observations from physical examinations.

Models trained on ASCENT follow a diagnosis-centered approach, emphasizing targeted supporting evidence for suspected diagnoses. This style can result in lower early-stage performance, particularly in Comprehensiveness and Relevance, as broad, nonspecific information in the early differential stages is often omitted. Similar trends were also observed in the model-based assessment of impressions.

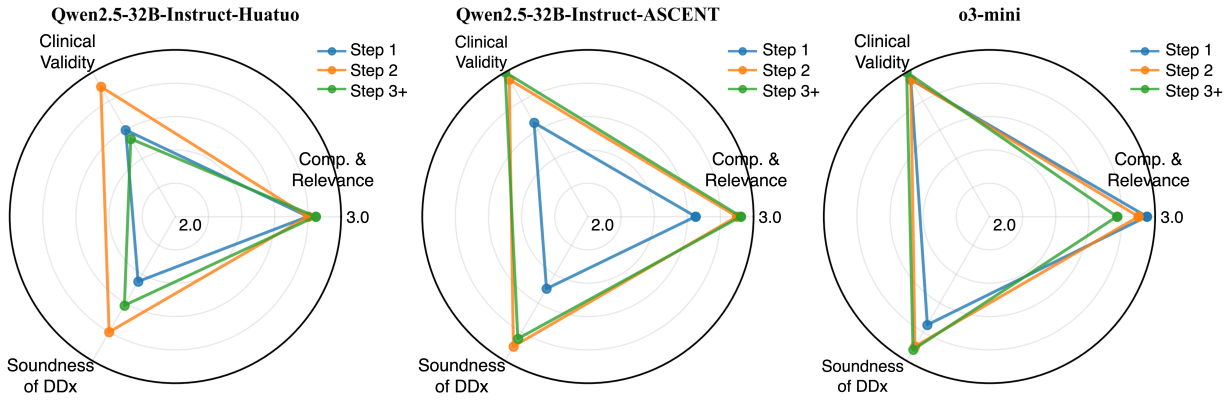

Figure 3: Expert assessment of stepwise rationales. The quality of model-generated rationales was assessed on three clinician-defined criteria: (1) **Comprehensiveness & Relevance** (*Comp. & Relevance*), (2) **Clinical Validity**, and (3) **Soundness of Differential Diagnosis** (*Soundness of DDx*). Solid lines indicate sequential diagnostic steps.

Intriguingly, models fine-tuned with MedQA-USMLE-Huatuo show generally lower and less stable performance, with a marked decline at Step 3+ in Clinical Validity and Soundness of Differential Diagnosis. This suggests that complex narrative CoTs from HuatuoGPT-o1 alone may be insufficient to enhance stepwise diagnostic reasoning. A similar Step 3+ decline is also observed from other models in Appendix E, although the contributing factors may differ between models.

Robustness Analysis. The evaluation above isolates per-step reasoning by providing gold context up to each step (thus *mitigated*). To characterize behaviour beyond this controlled setting—and to probe what ASCENT fine-tuning actually learns—we conducted four complementary analyses on o3-mini (strongest evaluated) and Qwen2.5-32B-Instruct fine-tuned on ASCENT (henceforth *ours-32B*). We first present three analyses that support ASCENT-trained models' learned behaviour, then turn to a key caveat: error propagation under realistic rollout.

(1) Counterfactual perturbation. On a stratified 75-case subset, we perturbed the most recent prior impression as *overly broad*, *overly specific*, *irrelevant* (unrelated diagnosis), or *order shift* (step swap). o3-mini was broadly resilient (largest drop: -0.05 F1 for *overly specific*). ours-32B exhibited a striking asymmetry: collapse under *overly broad* priors (F1 $= 0.15$) but robustness to *irrelevant* (0.63), *overly specific* (0.59), and *order shift* (0.61) (full results in Appendix C.5). We interpret this as the ASCENT-trained model treating prior impressions as an informative latent state summary: *overly broad* priors carry too little signal, while *irrelevant* ones are confidently incompatible and discarded.

(2) Format vs. content. The MedQA-USMLE-Huatuo SFT baseline (Qwen2.5 fine-tuned on HuatuoGPT-o1 complex CoT data) serves as a *format-matched control*: it produces step-level outputs in the same ASCENT-style format. Despite this identical output structure, it trails ASCENT fine-tuning by 0.04–0.11 F1 in Table 2, indicating that matching the surface format alone cannot explain ASCENT's gains. Combined with the counterfactual asymmetry above, this rules out a pure style-transfer interpretation: a stylistic adapter would degrade comparably to general-purpose baselines under content perturbation, whereas ASCENT fine-tuning instead amplifies sensitivity to prior impression *content*—consistent with the model learning to use accurate priors as diagnostic anchors.

(3) Judge-protocol robustness. To rule out artifacts specific to a single LLM judge, we re-judged the full set of 313 impression-level prediction–reference pairs with the stronger GPT-5.2 model and compared against the original GPT-4.1 judgments. Inter-judge agreement was substantial under the 4-class scheme (Cohen's quadratic-weighted $\kappa = 0.78$) and remained substantial when *Overly Specific* and *Overly Broad* were collapsed into a single granularity-mismatch class ($\kappa = 0.76$); only 0.64% of pairs reverse the dom-

inant correct-vs-other direction, confirming that the F1 rankings reported above are stable across judge choice (full confusion matrices in Appendix A).

(4) Rollout evaluation. Despite the positive evidence above, both evaluated models remain limited under realistic rollout, where each step receives the model's own prior prediction rather than gold context. For o3-mini, overall F1 collapses from 0.72 (mitigated) to 0.10 (rollout) for *Imp* and from 0.69 to 0.09 for *Imp+Reason*, with per-step degradation worsening monotonically from −0.52 at Step 1 to −0.78 at Step 3 (Appendix C.4); ours-32B collapses even further, from 0.54 to 0.01 (*Imp*) and 0.63 to 0.01 (*Imp+Reason*)—a larger *relative* drop (∼98% vs ∼86% for o3-mini) that reinforces the content-reliance interpretation above (a style-transfer adapter would be expected to degrade comparably to general-purpose baselines, not more). Strikingly, of 275 cases in which o3-mini produced a wrong impression, *zero recovered* at any subsequent step (Table 9)—indicating that error propagation is a fundamental obstacle for clinical deployment and motivating rollout-aware training as an important next step. While the rollout setting is closer to realistic deployment conditions, the mitigated setting remains informative on its own: it isolates per-step reasoning quality under controlled context and already exposes meaningful headroom on ASCENT, with rollout further compounding these gaps via error propagation.

Limitations and Future Work. Building a dataset like ASCENT requires substantial human expert effort, underscoring the need for automated, reliable methods for generating stepwise diagnostic data. Moreover, as highlighted in prior work (McDuff et al., 2025), even with expert involvement, determining whether an impression is sufficiently specific can be subjective, which may lead to occasional disagreements over diagnostic granularity.

Due to limited resources, several complementary validations remain future work: (i) a quantitative human baseline on the ASCENT test set; (ii) a comprehensive inter-rater agreement analysis over the full dataset annotations, beyond the dual-reviewer revision protocol described in Section 3.2; and (iii) a comprehensive human review of LLM-as-a-Judge outputs to complement the model-vs-model judge agreement reported in Section 4.3.

Scaling ASCENT to larger sample sizes is also an important direction, both to support reliable per-chapter rate estimates and to broaden specialty coverage. Relatedly, although our automated ontology mapping showed promise (Appendix C.7), finer-grained stratification of model errors along well-established medical ontologies is left for future work, and transfer to other sequential diagnostic datasets and to fully open-ended clinical settings is similarly not yet demonstrated. Finally, because ASCENT specifically targets stepwise diagnostic reasoning under EMR-aligned uncertainty, performance on ASCENT alone should not be interpreted as a comprehensive measure of clinical competence, which encompasses a variety of abilities such as patient communication, treatment planning, procedural skills, and longitudinal management.

5. Conclusion

We introduced ASCENT, a clinician-annotated benchmark that decomposes MedQA-USMLE vignettes into EMR-aligned diagnostic steps, making intermediate reasoning an explicit evaluation target—a regime that final-answer benchmarks cannot reach.

Our experiments yield the following salient findings. First, strong reasoning models substantially underperform their MedQA-USMLE accuracy on ASCENT, exposing a persistent gap between fully informed and stepwise diagnosis. Second, fine-tuning on ASCENT improves both small and large models over pre-trained and HuatuoGPT-o1 CoT-trained baselines, with gains driven primarily by precision—evidence that stepwise supervision teaches focused differential diagnosis rather than broader candidate enumeration. Third, robustness analyses reveal that ASCENT fine-tuning produces reliance on the content (not format) of prior impressions, while the strongest evaluated model (o3-mini) never recovers from a first wrong impression.

These findings call for rollout-aware training and for evaluation protocols that report both mitigated and rollout performance, since the two diverge sharply in our experiments. We release ASCENT as a public benchmark and substrate for stepwise clinical reasoning research.

Acknowledgments

We thank the physicians who annotated and the medical-expert reviewers who refined and qualitatively evaluated the ASCENT data. We are also grateful to the CHIL 2026 reviewers and area chair for feedback that shaped the robustness analyses.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *ArXiv*, abs/2505.08775, 2025. URL https://api.semanticscholar.org/CorpusID:278535396.

Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie E. Abdulnour, Adrian D. Haimovich, Jason Freed, Andrew P. J. Olson, Daniel J. Morgan, Jason Hom, Robert J. Gallo, Eric Horvitz, Jonathan H. Chen, Arjun K. Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician. *ArXiv*, abs/2412.10849, 2024. URL https://api.semanticscholar.org/CorpusID:274777011.

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

Michelle M Daniel, Joseph J. Rencic, Steven J. Durning, Eric S. Holmboe, Sally A. Santen, Valerie J Lang, Temple A. Ratcliffe, David Gordon, Brian S Heist, Stuart Lubarsky, Carlos A. Estrada, Tiffany Ballard, Anthony R. Artino, Ana Sergio Da Silva, Timothy J. Cleary, Jennifer N Stojan, and Larry D. Gruppen. Clinical reasoning assessment methods: A scoping review and practical guidance. *Academic Medicine*, 94(6):902–912, 2019. URL https://api.semanticscholar.org/CorpusID:73416893.

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

Mark L Graber, Nancy Franklin, and Ruthanna Gordon. Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13):1493–1499, 2005. URL https://api.semanticscholar.org/CorpusID:42753279.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak N. Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Jacob Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642:451 – 457, 2025. URL https://api.semanticscholar.org/CorpusID:265551974.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jiaming Ji, Zifeng Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, Zexuan Guo, Youlan Lei, Chunli Shao, Wen yao Wang, Haojun Fan, and Yifang Tang. The application of large language models in medicine: A scoping review. *iScience*, 27(5):109713:1–16, 2024. URL https://api.semanticscholar.org/CorpusID:269354751.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott M. Lundberg, Marco Túlio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P. Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025.

Thomas Savage, Ashwin Nayak, Roberta Gallo, Ekanath Srihari Rangan, and Jonathan H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability

in medicine. *npj Digital Medicine*, 7:20:1–7, 2023. URL https://api.semanticscholar.org/CorpusID:260887092.

K. Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather J. Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, S. Lachgar, P. A. Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee C Wong, Christopher Semturs, Seyedeh Sara Mahdavi, Joëlle K. Barral, Dale R. Webster, Greg S Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models, 2022.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs, 2025a.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports, 2025b.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. Automating expert-level medical reasoning evaluation of large language models, 2025.

Appendix A. Judge Reliability

To assess the reliability of our LLM-as-a-Judge protocol, we re-ran the impression-classification step using GPT-5.2 as a stronger reference judge on the full test set, yielding 313 impression-level pairs of judgments. Table 3 summarizes agreement statistics; Tables 4 and 5 show the 4-class and equiv-23 (collapsed) confusion matrices.

Setting	**Agree.**	**κ unwt.**	**κ lin**	**κ quad**
4-class standard	0.84	0.62	0.70	0.78
3-class equiv-23	0.85	0.63	0.69	0.76

Table 3: Inter-judge agreement between GPT-4.1 (Judge A) and GPT-5.2 (Judge B) on 313 impression-level pairs. The equiv-23 collapse merges *Overly Specific* and *Overly Broad* into a single granularity-mismatch class. Both settings show substantial agreement under quadratic-weighted Cohen's κ. The micro-F1 of GPT-5.2 was higher than GPT-4.1 by +0.08, but only 0.64% of pairs reverse direction (i.e., trend-reversing disagreements), indicating that residual disagreements concentrate at class boundaries rather than systematic inversions.

A \ B	B=1	B=2	B=3	B=4
A=1	220	2	1	2
A=2	21	25	1	4
A=3	7	2	6	9
A=4	0	0	0	13

Table 4: 4-class confusion matrix between GPT-4.1 (rows, Judge A) and GPT-5.2 (columns, Judge B). Class labels: 1 = Correct Match, 2 = Overly Specific, 3 = Overly Broad, 4 = Others. 95.9% of disagreements fall within two ordinal classes ($|A - B| \leq 2$); strictly adjacent disagreements ($|A - B| = 1$, e.g., 1 vs. 2 or 2 vs. 3) account for 71.4%.

A \ B	B=1	B=2/3	B=4
A=1	220	3	2
A=2/3	28	34	13
A=4	0	0	13

Table 5: Confusion matrix after collapsing *Overly Specific* (2) and *Overly Broad* (3) into a single granularity-mismatch class.

Appendix B. Prompts for Data Generation, Inference, and Evaluation

B.1. Data Generation Prompt Template.

Figures 4 and 5 show the prompt template used to generate ASCENT data with DeepSeek-R1. The resulting outputs were meticulously reviewed and revised by reviewers prior to inclusion in ASCENT.

B.2. Inference Prompt Template.

Figure 6 presents the prompt template for the impression (***Imp***) and for the impression and rationale (***Imp + Reason***) settings.

B.3. Evaluation Prompt Template and Post-Processing of Evaluation Results.

Figure 7 includes the prompt template used to evaluate and classify model-predicted impressions. The output was evaluated using OpenAI GPT-4.1 (temperature=0.0). To improve the consistency of the LLM-as-a-Judge approach, inference was repeated 5 times using the same prompt.

The resulting evaluation outputs were post-processed to manually correct LLM-induced formatting errors (e.g., missing \\boxed{{}}). We also handled cases where the number of evaluation outputs did not match the number of provided impressions by applying position-wise majority voting to determine the final values. We then retained only the first k positions, where k equals the number of provided impressions, discarding any surplus values to maintain consistency of evaluation results.

Prompt: DeepSeek-R1-Based Data Generation

You are an assistant that helps with EMR documentation.
You are provided with a complex_cot corresponding to the given question.
Classify the contents of complex_cot into EMR sections according to the **EMR Section Classification Rules** below.
Then, for each classified section, divide the content into Findings, Reason, and Imp according to the **Diagnostic Reasoning Rules**.

EMR Section Classification Rules
1. EMR Section Names
- PI: Information related to the symptoms or conditions the patient is currently experiencing. Age, gender, and circumstances are also included in PI.
- CC: Among the contents corresponding to PI, extract the patient's **chief complaint** and express it in medical terminology (e.g. dyspnea). If the patient visits without a chief complaint, enter as appropriate for the situation, such as health checkup, prenatal examination, or regular checkup.
- PEx: Physical examination findings that the physician can confirm in the consultation room. Height, weight, and vital signs are also included in PEx.
- Lab: Results of laboratory tests performed on specific specimens, such as blood tests or urine tests.
- Img: Results of imaging tests such as X-ray, CT, or ultrasound.
- Etc: Other tests and results not included in PEx, Lab, or Img, such as ECG, endoscopy, or biopsy. The patient's post-treatment status or future plans presented in the question are also included in Etc.
2. Caution
- The order of EMR section classification follows the order of content in complex_cot.
- Content classified as PI, PEx, Lab, Img, or Etc should not overlap.
- Only generate EMR sections that are mentioned in complex_cot.

Diagnostic Reasoning Rules
1. Diagnostic Reasoning Composition
- Findings: Content from each classified section provided in the question
- Imp: The diagnosis or suspected system name mentioned based on the content
- Reason: The basis for inferring the Imp of the section based on the Findings from each classified section
2. Caution
- Among the EMR sections, CC does not require diagnostic reasoning. Therefore, do not write Findings, Reason, or Imp related to CC.
- Write Findings based on the format of the content provided in the question.
- If there is no Imp or Reason that can be inferred from the Findings, enter nan for Imp and Reason.
- When entering units such as °C, mL, use the units as written in the question.

Final Answer Rules
1. Answer: Based on the diagnostic reasoning, understand the patient's situation and select one answer from the options to the question being asked.
2. Reason: Write the rationale for the correct answer in connection with the content of the diagnostic reasoning.

Figure 4: The prompt template for DeepSeek-R1-based generation.

Prompt: DeepSeek-R1-Based Data Generation (Continued)

<Example>
question
A 29-year-old African American female presents to your office with extreme fatigue and bilateral joint pain. Serologies demonstrate the presence of rheumatoid factor along with anti-Smith and anti-dsDNA antibodies. A VDRL syphilis test is positive. You order a coagulation profile, which reveals normal bleeding time, normal PT, and prolonged PTT as well as normal platelet count. Further evaluation is most likely to reveal which of the following?
A: Palmar rash
B: HLA-B27 positivity
C: Factor VIII deficiency
D: History of multiple spontaneous abortions

complex_cot
Alright, we've got a 29-year-old African American woman showing up with extreme fatigue and pain in her joints. That makes me think autoimmune issues might be involved. Let's look at the test results... (...)

Diagnostic Reasoning
- [CC] fatigue, joint pain
- [PI] Findings: A 29-year-old African American female presents with extreme fatigue and bilateral joint pain
- [PI] Imp: autoimmune diseases
- [PI] Reason: Autoimmune diseases are suspected in a young female patients with fatigue and joint pain.
- [Lab] Findings: Serologies show rheumatoid factor (+), anti-Smith (+), anti-dsDNA (+). VDRL (+). Coagulation profile reveals normal bleeding time, PT, platelet count; prolonged PTT.
- [Lab] Imp: antiphospholipid syndrome
- [Lab] Reason: Anti-Smith and anti-dsDNA antibodies are specific for SLE. Positive VDRL (a false positive due to antiphospholipid antibodies) and prolonged PTT (due to lupus anticoagulant) suggest antiphospholipid syndrome.
Final Answer
Answer: D. History of multiple spontaneous abortions
Reason: The presence of lupus anticoagulant (prolonged PTT) and antiphospholipid antibodies (false-positive VDRL) in a patient with SLE (anti-Smith and anti-dsDNA) suggests antiphospholipid syndrome, which is associated with thrombosis and pregnancy complications like recurrent miscarriages.
<End of Example>

question
{question}
complex_cot
{complex_cot}
Diagnostic Reasoning

Figure 5: The prompt template for DeepSeek-R1-based generation (continued).

Prompt: Inference (*Imp*)

Diagnosis is a multi-step and iterative process.
You are given patient information and the physician's reasoning up to the previous stage.
Based on the provided information and medical knowledge, write the name(s) of the disorder(s) or the related system(s) that should be suspected at the final Findings stage inside \\boxed{{}}.
The diagnosis should be appropriate for inference at the current stage and should be neither overly specific nor overly vague.
{diagnostic_context}

Prompt: Inference with Rationale (*Imp + Reason*)

Diagnosis is a multi-step and iterative process.
You are given patient information and the physician's reasoning up to the previous stage.
Based on the provided information and medical knowledge, consider the name(s) of the disorder(s) or the related system(s) that should be suspected at the final Findings stage.
Then, concisely state the reason for your suspicion, beginning with "Reason:".
Subsequently, write the name(s) of the suspected disorder(s) or system(s) inside \\boxed{{}}.
The diagnosis should be appropriate for inference at the current stage and should be neither overly specific nor overly vague.
{diagnostic_context}

Figure 6: The prompt template for inference for the impression (top: ***Imp***) and for the impression and rationale (bottom: ***Imp + Reason***).

Prompt: Evaluation

You are a diagnostic accuracy evaluator. Determine whether the Suggested Impression (i.e., the suspected disease or related system), proposed as an answer to the given Question, matches the Ground-Truth Impression. Even if the Suggested Impression does not exactly match the Ground-Truth, consider it correct if it refers to the medically equivalent disease or system.

If multiple impressions are listed in the Suggested Impression, evaluate each one individually against the Ground-Truth. If the Ground-Truth contains multiple entities, consider the Suggested entity correct if it matches any of the Ground-Truth entities.

For each entity, assign one of the following values and list the results in order, separated by commas, inside \\boxed{{}}:
1 — Correct match
2 — Overly specific compared to the Ground-Truth (e.g., Ground-Truth: respiratory disorder; Suggested: COPD)
3 — Overly broad compared to the Ground-Truth (e.g., Ground-Truth: Lupus nephritis; Suggested: glomerulonephritis)
4 — Incorrect and does not fall into the above categories

Question
{inference_prompt}
Suggested Impression
{model_generated_impression}
Ground-Truth Impression
{ground_truth_impression}

Figure 7: The prompt template for model output evaluation.

Targets	Models	Precision [95% CI]	Recall [95% CI]	F1 Score [95% CI]
Imp	**Small Models (<10B)**			
	Qwen2.5-7B-Instruct	0.33 [0.29, 0.38]	0.50 [0.44, 0.55]	0.40 [0.35, 0.45]
	+ SFT w/ MedQA-USMLE-Huatuo	0.38 [0.34, 0.43]	0.50 [0.44, 0.55]	0.43 [0.38, 0.48]
	+ SFT w/ ASCENT (ours)	**0.50** [0.44, 0.55]	**0.50** [0.45, 0.55]	**0.50** [0.44, 0.55]
	HuatuoGPT-o1-7B	0.46 [0.41, 0.51]	**0.54** [0.49, 0.60]	0.50 [0.45, 0.55]
	Qwen3-8B	**0.50** [0.45, 0.56]	0.51 [0.46, 0.56]	**0.51** [0.45, 0.56]
	Large Models (≥10B)			
	OpenAI GPT-4o	0.44 [0.40, 0.49]	0.54 [0.48, 0.59]	0.49 [0.44, 0.54]
	OpenAI GPT-5.2	0.44 [0.39, 0.49]	**0.65** [0.60, 0.70]	0.53 [0.48, 0.57]
	Qwen2.5-32B-Instruct	0.38 [0.33, 0.43]	0.51 [0.45, 0.56]	0.43 [0.38, 0.48]
	+ SFT w/ MedQA-USMLE-Huatuo	0.41 [0.35, 0.46]	0.47 [0.41, 0.52]	0.43 [0.38, 0.49]
	+ SFT w/ ASCENT (ours)	**0.55** [0.49, 0.60]	0.54 [0.48, 0.59]	**0.54** [0.49, 0.60]
	Qwen3-32B	0.43 [0.38, 0.48]	0.57 [0.51, 0.62]	0.49 [0.44, 0.54]
	DeepSeek-R1	0.59 [0.54, 0.65]	**0.73** [0.68, 0.78]	0.65 [0.60, 0.70]
	OpenAI GPT-o3-mini	**0.72** [0.67, 0.77]	0.71 [0.66, 0.76]	**0.72** [0.67, 0.76]
Imp + Reason	**Small Models (<10B)**			
	Qwen2.5-7B-Instruct	0.43 [0.38, 0.48]	**0.55** [0.50, 0.60]	0.48 [0.43, 0.53]
	+ SFT w/ MedQA-USMLE-Huatuo	0.48 [0.43, 0.54]	0.52 [0.46, 0.57]	0.50 [0.44, 0.55]
	+ SFT w/ ASCENT (ours)	**0.55** [0.50, 0.61]	0.53 [0.48, 0.58]	**0.54** [0.49, 0.59]
	HuatuoGPT-o1-7B	0.49 [0.43, 0.54]	**0.59** [0.54, 0.65]	0.54 [0.48, 0.59]
	Qwen3-8B	**0.61** [0.56, 0.67]	0.57 [0.52, 0.62]	**0.59** [0.54, 0.65]
	Large Models (≥10B)			
	OpenAI GPT-4o	0.48 [0.43, 0.53]	0.54 [0.49, 0.60]	0.51 [0.46, 0.56]
	OpenAI GPT-5.2	0.50 [0.46, 0.55]	**0.66** [0.61, 0.72]	0.57 [0.52, 0.62]
	Qwen2.5-32B-Instruct	0.54 [0.49, 0.60]	0.63 [0.57, 0.68]	0.58 [0.53, 0.63]
	+ SFT w/ MedQA-USMLE-Huatuo	0.54 [0.49, 0.60]	0.58 [0.52, 0.63]	0.56 [0.51, 0.61]
	+ SFT w/ ASCENT (ours)	**0.65** [0.59, 0.70]	0.62 [0.57, 0.67]	**0.63** [0.58, 0.68]
	Qwen3-32B	0.55 [0.50, 0.61]	0.62 [0.57, 0.67]	0.58 [0.53, 0.63]
	DeepSeek-R1	0.60 [0.55, 0.66]	0.67 [0.61, 0.72]	0.63 [0.58, 0.68]
	OpenAI GPT-o3-mini	**0.69** [0.64, 0.74]	**0.68** [0.63, 0.73]	**0.69** [0.64, 0.74]

Table 6: Overall diagnosis performance (micro precision, micro recall, and micro F1 score, with bootstrap 95% CIs in brackets) of open-source and proprietary models on the ASCENT test set ($n = 305$), evaluated on two tasks: (1) generating impressions only (***Imp***) and (2) generating supporting rationales followed by impressions (***Imp + Reason***). CIs use 10,000 test-record bootstrap resamples (percentile method); within each resample, P, R, and F1 are computed from the same aggregated TP/FP/FN counts so the three values remain self-consistent. Bold marks the best value within each model category.

Appendix C. Stepwise Diagnosis Performance

C.1. Overall diagnostic performance on the ASCENT test set.

Table 6 shows the overall diagnostic performance (micro precision, micro recall, and micro F1 score) of both open-source and proprietary models on the ASCENT test set.

C.2. A note on Qwen3 model scaling.

As noted in the main manuscript, model size contributed to performance gains for Qwen2.5-Instruct models. Larger models consistently achieved higher F1 scores, especially with additional training on ASCENT. However, Qwen3 models did not follow this trend, as the larger model did not outperform the smaller one. A closer inspection of stepwise errors, where a single question can produce multiple impressions and thus multiple errors, shows that Qwen3-32B generated more overly broad errors across all steps (Step 1: 60 errors, Step 2: 32 errors, Step 3: 16 errors) than Qwen3-8B (Step 1: 29 errors, Step 2: 25 errors, Step 3: 10 errors).

C.3. Token-length ablation.

We varied the maximum generation length for o3-mini from 512 to 4,096 tokens to assess sensitivity of stepwise diagnostic performance to the token budget. Table 7 reports overall micro-F1 across budgets for both *Imp* and *Imp+Reason*. Performance is essentially flat above 1,024 tokens: maximum variation is 0.03 for *Imp* and essentially flat for *Imp+Reason* (0.003), well within one bootstrap standard error. Only the 512-token setting showed non-trivial degradation, and we observe no evidence of overthinking-induced regression at higher budgets, supporting the choice of 4,096 tokens for reasoning models in our main experiments.

Token budget	Imp F1	Imp+Reason F1
512	0.69	0.65
1,024	0.69	0.67
2,048	0.69	0.67
4,096 (default)	0.72	0.67

Table 7: Token-length ablation for OpenAI o3-mini on the full ASCENT test set ($n = 305$). Performance is approximately constant beyond 1,024 tokens.

C.4. Rollout (error propagation) evaluation.

We evaluated o3-mini under a rollout setting in which each step receives the model's own predicted impression from the prior step as context, rather than the gold impression used in the mitigated setting. Both task modes (*Imp* and *Imp+Reason*) were evaluated on the full ASCENT test set ($n = 305$) using the standard GPT-4.1 judge with five repeats and majority vote. Table 8 reports per-step F1 under both protocols, and Table 9 reports the recovery rate after the first incorrect impression.

	Imp			Imp+Reason		
	Mit.	Roll.	Δ	Mit.	Roll.	Δ
Step 1	0.66	0.14	−0.52	0.58	0.12	−0.45
Step 2	0.74	0.10	−0.64	0.75	0.08	−0.67
Step 3	0.82	0.03	−0.78	0.83	0.06	−0.77
Step 4	0.62	0.00	−0.62	0.62	0.07	−0.55
Overall	0.72	0.10	−0.62	0.69	0.09	−0.60

Table 8: Per-step and overall F1 for OpenAI o3-mini under **mitigated** (gold prior) and **rollout** (model's own prior) protocols (Step 5 omitted, $n = 2$). Degradation worsens monotonically with step depth in both task modes. We focus the per-step analysis on o3-mini because it is the strongest model on ASCENT (Table 2) and therefore offers the most informative test of error propagation: if even the best evaluated model cannot recover under rollout, the collapse is unlikely to be an artifact of weak baseline capability. The corresponding overall rollout collapse for ours-32B is reported in the main text.

Model	Total	First-wrong	Recovered
o3-mini (*Imp*)	305	275	**0 (0.0%)**

Table 9: Recovery rate after first incorrect impression under rollout. Of the 275 o3-mini cases (90.2% of the test set) that produced a wrong impression at any step, none recovered at any subsequent step.

C.5. Counterfactual perturbation.

On a stratified 75-case subset, we perturbed the most recent prior impression with one of four variants: *overly broad*, *overly specific*, *irrelevant* (an unrelated diagnosis), and *order shift* (swapping two consecutive steps' prior impressions). Both o3-mini (the strongest evaluated model) and ours-32B (Qwen2.5-32B-Instruct fine-tuned on ASCENT) were evaluated. Table 10 reports the resulting F1 and resilience rate (fraction of cases producing at least one correct impression) for each variant.

C.6. Draft revision analysis.

To confirm that expert reviewers devoted comparable effort to revising and curating annotations from both draft sources—Llama-3.1-8B and DeepSeek-R1—we computed textual similarity between each draft and its final expert-revised version. A "major edit" corresponds to similarity < 0.85. Table 11 reports the resulting edit distribution. The substantially lower edit rate for the DeepSeek-R1-drafted subset likely reflects DeepSeek-R1's stronger generation capacity relative to Llama-3.1-8B rather than reduced reviewer effort; all DeepSeek-R1 drafts received the same independent two-reviewer protocol described in Section 3.1.

C.7. Ontology coverage.

As supporting evidence for the clinical groundedness of ASCENT's ground-truth impressions, we programmatically mapped all 1,152 entities to UMLS (2025AA), identifying 1,035 (89.8%) matches. Of the matched entities, 836 (80.8%) carry at least one SNOMED CT code, and 401 are linked through UMLS to ICD-10 codes spanning all 20 disease chapters (Table 12). This aggregate coverage suggests that ASCENT's impression vocabulary aligns well with standard medical ontologies and indicates potential utility for downstream tasks that benefit from standardized terminology; we leave full annotation of the entity-level mapping, which would warrant additional expert review, to future work.

Appendix D. Expert Evaluation Criteria and Scoring Rubric

To further examine stepwise clinical reasoning, we conducted a qualitative case study on 20 problems. Model-generated rationales were evaluated on a 3-point Likert scale based on the following criteria: (1) **Comprehensiveness and Relevance**; (2) **Clinical Validity**; (3) **Soundness of Differential Diagnosis**. The detailed scoring rubric for the criteria is provided below.

- **Comprehensiveness and Relevance** assesses whether all key relevant information is included and unnecessary details are excluded.
  3 — All key information is included and directly contributes to the differential diagnosis.
  2 — Some highly relevant information is missing, or some unnecessary details are present.
  1 — Important information is missing, or the majority of the content consists of irrelevant details.

- **Clinical Validity** evaluates whether the clinical content and medical knowledge are accurately and appropriately described.
  3 — All clinical information is accurate and aligns well with established medical knowledge.
  2 — Most information is accurate, but there are instances of incompleteness or ambiguity.
  1 — Contains inaccurate or misinterpreted clinical information.

- **Soundness of Differential Diagnosis** judges whether the resulting diagnosis is clinically appropriate and well-supported by the provided information.
  3 — The diagnosis is clear, well-supported, and clinically valid.
  2 — The diagnosis is somewhat plausible but either too broad or too specific.
  1 — The diagnosis is not clinically convincing or does not align with the information provided.

Model	Variant	F1	Resilience
o3-mini	overly broad	0.66	0.69
	overly specific	0.62	0.65
	irrelevant	0.65	0.68
	order shift	0.68	0.72
ours-32B	overly broad	**0.15**	**0.16**
	overly specific	0.59	0.63
	irrelevant	0.63	0.67
	order shift	0.61	0.64

Table 10: Counterfactual perturbation of the most recent prior impression on a 75-case subset. **Resilience** is an auxiliary metric we introduce for this analysis: the fraction of cases in which the model still produces at least one correct impression somewhere in the diagnostic trajectory under the perturbed prior, complementing the per-step F1 column. o3-mini (gold-prior baseline F1 on this subset = 0.67) is broadly resilient, with all four variants staying within roughly 5 F1 points of baseline. ours-32B exhibits a striking asymmetry: collapse under *overly broad* priors (F1 = 0.15, far below its overall test-set F1 of 0.54 in Table 6) but robustness to *irrelevant*, *overly specific*, and *order shift* variants, consistent with content-based (rather than format-based) reliance on the prior impression.

Draft source	No edit	Minor	Major
DeepSeek-R1 ($n = 558$)	57.1%	25.5%	17.5%
Llama-3.1-8B ($n = 680$)	—	2.3%	97.7%

Table 11: Edit distribution of physician revisions over the two draft sources used to construct ASCENT. "Major" is defined as similarity < 0.85 between draft and final. The reviewer protocol is described in Section 3.1.

Appendix E. Expert Evaluation on Model-Generated Rationales

Figure 8 displays the full results of expert assessment on stepwise rationales.

Resource	Coverage
UMLS (2025AA)	1,035 / 1,152 (89.8%)
SNOMED CT (via UMLS)	836 / 1,035 (80.8%)
ICD-10 (via UMLS)	401 entities, 20 chapters

Table 12: Ontology coverage of ASCENT ground-truth impression entities. The 117 unmatched UMLS entities are free-text terms. **"via UMLS"** indicates that mappings to SNOMED CT and ICD-10 were obtained indirectly, using each UMLS Concept Unique Identifier (CUI) as a bridge to the corresponding SNOMED CT and ICD-10 codes in the UMLS Metathesaurus (2025AA); we did not query SNOMED CT or ICD-10 directly. Consequently, SNOMED CT coverage is reported as a fraction of UMLS-matched entities, and the 401 ICD-10-linked entities span the full disease taxonomy (all 20 chapters).

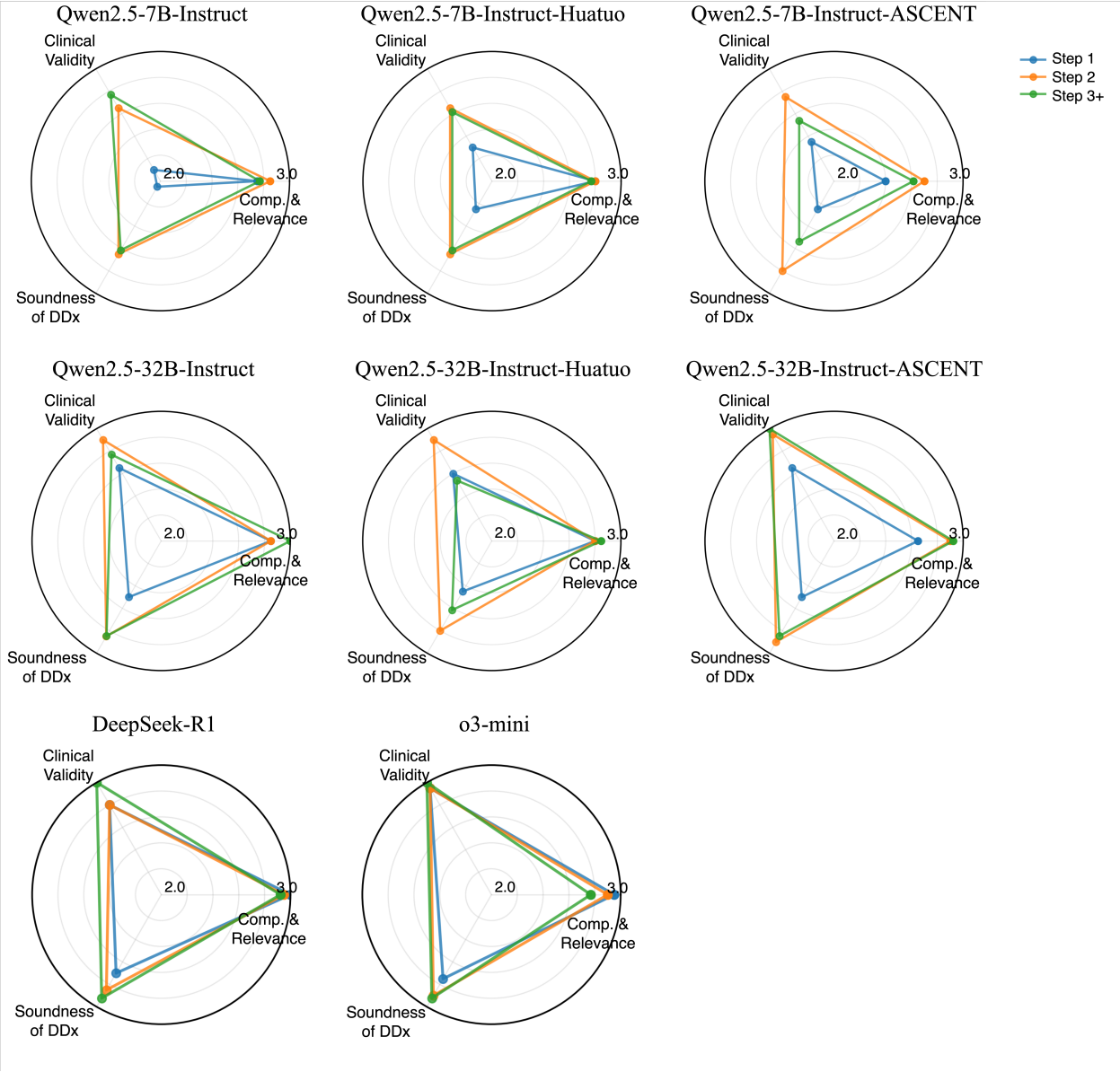

Figure 8: Expert assessment of stepwise rationales (full results). The quality of model-generated rationales was assessed on three clinician-defined criteria: (1) **Comprehensiveness & Relevance** (*Comp. & Relevance*), (2) **Clinical Validity**, and (3) **Soundness of Differential Diagnosis** (*Soundness of DDx*). Solid lines indicate sequential diagnostic steps.