`MedCaseReasoning`: Evaluating and learning diagnostic reasoning from clinical case reports

Kevin Wu[1,*] Eric Wu[1,*] Rahul Thapa[1]
Kevin Wei[2] Angela Zhang[3] Arvind Suresh[3] Jacqueline J. Tao[3]
Min Woo Sun[1] Alejandro Lozano[1] James Zou[1]

[1]*Stanford University*
[2]*University of Southern California*
[3]*University of California, San Francisco*
[*]*Equal contribution. Correspondence: kewu93@gmail.com*

Doctors and patients alike increasingly use Large Language Models (LLMs) to diagnose clinical cases. However, unlike domains such as math or coding, where correctness can be objectively defined by the final answer, medical diagnosis requires both the outcome and the reasoning process to be accurate. Currently, widely used medical benchmarks like MedQA and MMLU assess only accuracy in the final answer, overlooking the quality and faithfulness of the clinical reasoning process. To address this limitation, we introduce `MedCaseReasoning`, the first open-access dataset for evaluating LLMs on their ability to align with diagnostic reasoning documented in clinician-written case reports. The dataset includes 13,989 diagnostic question-and-answer cases, each paired with detailed reasoning statements extracted by an LLM from open-access medical case reports. We evaluate state-of-the-art reasoning LLMs on `MedCaseReasoning` and find significant shortcomings in their diagnoses and reasoning: for instance, the top-performing open-source model, DeepSeek-R1, achieves only 48% pass@10 diagnostic accuracy and mentions only 64% of the reference reasoning statements (recall). However, we demonstrate that fine-tuning LLMs on the reasoning traces derived from `MedCaseReasoning` significantly improves diagnostic accuracy and clinical reasoning recall by an average relative gain of 41% and 29% (13.5 and 10.4 absolute percentage points), respectively. Our results showcase the importance of aligning LLMs to reason like human experts.

Keywords: Clinical reasoning; large language models; diagnostic accuracy; benchmark dataset.

1. Introduction

Large language models (LLMs) are increasingly deployed in medicine,[1] where they show promise in performing complex tasks like clinical reasoning[2] and disease diagnosis.[3] While there are a number of benchmark datasets designed to assess the diagnostic capabilities of medical LLMs (e.g. MedQA,[4] MMLU,[5] MultiMedQA,[6] and MedXpertQA[7]), these datasets all share the same limitation in that they only assess the correctness of the model's final answer. Unlike domains like mathematics[8] or coding,[9] where the reasoning process is secondary to the correct answer, medical diagnoses require both the final diagnosis and the thought process to be defensible. In clinical practice, physicians need to articulate coherent rationales, which is important for ensuring informed consent, meeting the standard of care, facilitating documentation and billing, and enabling clinical

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

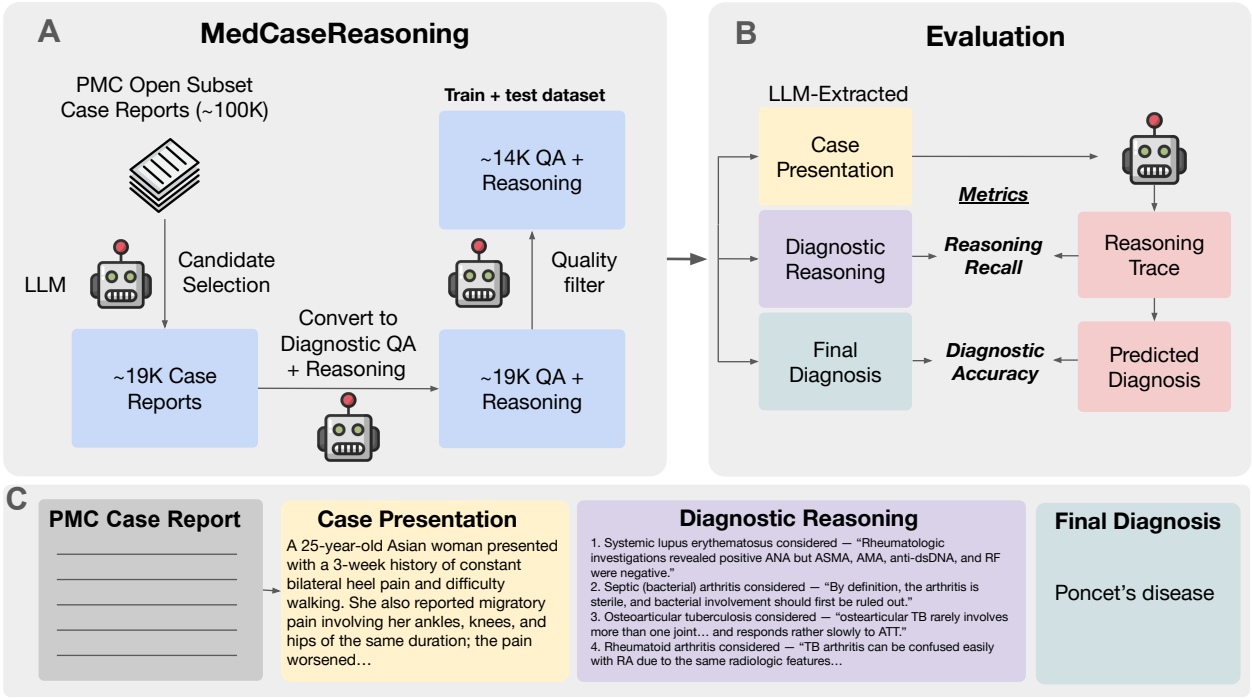


Fig. 1. A schematic of the `MedCaseReasoning` case processing pipeline. **A**: From the initial set of 98,994 PMC Open Subset Case Reports, we select 28,313 appropriate candidate cases, which are converted to QA format and then filtered for quality. The resulting 13,989 cases form the `MedCaseReasoning` dataset, with 897 test cases and 13,092 training cases. A full example is available in the Supplementary Material. **B**: For each test case, the case presentation is posed to an LLM to reason and then answer with a diagnosis. The reference reasoning extracted from the case report is compared with the LLM-generated reasoning to produce the reasoning recall score, and the case diagnosis is compared with the predicted diagnosis to produce diagnostic accuracy. **C**: The original case report text is converted into three sections: the case presentation that contains the relevant and sufficient patient information for making a diagnosis; the diagnostic reasoning that includes the diagnostic decision-making by the case author (in enumerated statements); and the final diagnosis.

collaboration.[10] Mistakes or deficiencies in reasoning, even if the correct diagnosis is given, can risk case mismanagement. Highlighting this gap, a recent study found that even frontier models like GPT-4 could produce the correct diagnosis for incorrect reasons in up to a third of clinical scenarios.[11] Thus, evaluating and improving model capabilities to reason and answer correctly is crucial for LLMs to gain clinical viability and trustworthiness.

In this work, we propose **`MedCaseReasoning`**, a diagnostic reasoning dataset with two complementary goals: (1) an evaluation benchmark to assess how well LLM reasoning aligns with the diagnostic reasoning documented in clinician-written case reports, and (2) a training dataset for improving diagnostic reasoning in LLMs from reasoning traces. `MedCaseReasoning` is an open-access dataset of 13,989 diagnostic cases drawn from 800+ medical journals and 30+ specialties, created from publicly available case reports published in PubMedCentral.[12] To produce it, we develop a scalable, multi-step, clinician-validated pipeline that filters an initial set of 98,994 case reports and converts them into QA-format diagnostic cases, yielding a high-quality

test subset of 897 cases. First, we observe that `MedCaseReasoning` is a meaningfully difficult and unsaturated benchmark, with the top commercial model, *OpenAI o3*, achieving only 65% on pass@10 diagnostic accuracy. Second, we find that even top-performing reasoning models are substantially limited in their ability to reason like clinicians, where the best open-source model, *DeepSeek-R1*, only recalls 64% of the reasons provided in the case report. Third, we find that training on `MedCaseReasoning` can significantly improve the diagnostic and reasoning capabilities of LLMs: both diagnostic accuracy and reasoning recall increased after fine-tuning three top medical open-source LLMs, with an average relative gain of 41% and 29% (13.5 and 10.4 absolute percentage points), respectively. Importantly, after training on `MedCaseReasoning`, their performance on NEJM CPC,[3,13] a held-out test dataset, also improved, showcasing the value of `MedCaseReasoning` in improving diagnostic generalizability.

Our work is the first medical benchmark dataset that explicitly evaluates the reasoning accompanying a diagnosis by comparing it to reasoning extracted from real-world case reports written by clinicians. These cases contain relevant differential diagnoses along with case presentations that present sufficient detail for an accurate diagnosis. Our findings underscore the significant challenges of aligning the reasoning of current top-performing LLMs with established clinical practice and provide a dataset for improving this alignment.

Throughout, *reference reasoning* denotes the reasoning statements an LLM extracts from a clinician-written case report; it is grounded in, but not identical to, the clinicians' own reasoning.

Related Works

Exam-style and literature-derived benchmarks. While commonly used medical benchmark datasets like MedQA,[4] MMLU,[5] and MedXpertQA[7] include diagnostic reasoning questions, the clinical cases presented are typically shorter, textbook-style cases intended to evaluate students and test medical knowledge, not real-world clinical cases that physicians may encounter. These benchmarks are also multiple-choice and score only the selected answer. Several datasets construct evaluation items from real-world sources, such as PubMedQA,[14] which generated QA examples based on PubMed articles, and MedAlign,[15] which generated clinical vignettes from electronic health records. However, PubMedQA asks yes/no/maybe questions about research abstracts rather than posing open-ended diagnostic problems, and MedAlign targets clinician instruction-following over EHRs, which requires credentialed data access; neither provides reference diagnostic reasoning against which a model's reasoning can be scored. MedHELM[16] broadens evaluation to a clinician-validated taxonomy of 121 medical tasks across 35 benchmarks, but it, too, scores model outputs rather than the reasoning that produced a diagnosis. By comparison, `MedCaseReasoning` is derived directly from clinician-authored case reports, which are valuable tools for advancing clinical practice,[17] and pairs each open-ended diagnostic question with reference reasoning extracted from the report.

Real-world diagnostic cases. The New England Journal of Medicine Clinicopathologic Conferences (NEJM CPC) is a collection of case reports sourced from Massachusetts General Hospital in Boston, MA.[3] Prior studies[3,13,18] have leveraged NEJM CPC to evaluate model capabilities, but primarily focus on generating differential diagnosis (DDx) lists, not full reasoning traces. `MedCaseReasoning` improves upon this dataset in several ways. First, the collection of case reports from PubMedCentral represents a globally diverse set of clinician and patient backgrounds, whereas NEJM CPC represents the medical practices of physicians from a single

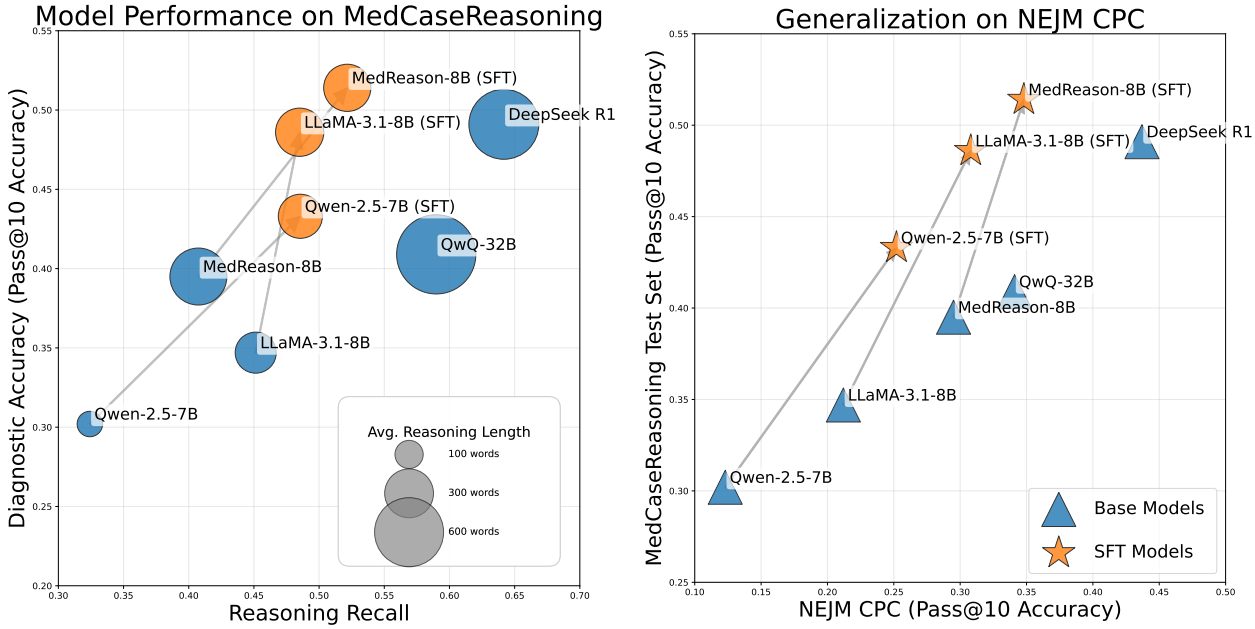


Fig. 2. (Left) Evaluation of LLMs on the `MedCaseReasoning` test dataset (N = 897), along with diagnostic accuracy and reasoning recall percentage. (Right) Model performance on the `MedCaseReasoning` test set and NEJM case studies (N = 302), showing a strong correlation between the two benchmarks (Pearson $r = 0.92$, $p < 0.001$, Pass@10 across the ten base and fine-tuned models). In both plots, accuracy is computed with pass@10 accuracy; circle size encodes the average length of the model-generated reasoning trace (trace length is correlated with reasoning recall; $r = 0.790$, $p = 0.0196$). LLaMA-3.1-8B and Qwen-2.5-7B are both Instruct variants (omitted for brevity).

hospital system on a specific population. Second, the NEJM CPC dataset consists of only 302 test cases, while `MedCaseReasoning` is comprised of 13,989 cases, allowing for more in-depth analysis as well as model fine-tuning. Third, `MedCaseReasoning` is derived from open-access articles on PubMedCentral, whereas NEJM CPC is available only under license.

Evaluating diagnostic reasoning. Several studies[19,20] have employed grading rubrics applied to curated clinical vignettes, but are restricted to a handful of cases, not openly accessible, and focus on human-AI comparison rather than establishing benchmarks for standalone LLM reasoning performance. One study[21] conducted manual clinician evaluations of reasoning traces from GPT-3.5 and GPT-4 to identify logical inconsistencies, but such methods are labor-intensive and inherently limited in scale, making them infeasible for large-scale benchmarking. `MedCaseReasoning` instead automates reasoning evaluation against reference reasoning for every case, with an LLM judge validated by a physician.

Training on medical reasoning traces. Recent models like HuaTuoGPT-o1[22] and MedReason[23] involve supervised fine-tuning (SFT) on diagnostic reasoning traces, but the traces are distilled from larger proprietary models (e.g., GPT-4) rather than being grounded in reasoning authored by clinicians based on real patient cases. `MedCaseReasoning` provides training traces derived from clinician-written case reports, and we show that they improve models already trained on synthetic traces (Section 3).

Table 1. Comparison of clinical diagnosis benchmark datasets. `MedCaseReasoning` is the only and largest open-access dataset with clinician reasoning. Existing datasets do not contain expert reasoning.

Dataset	N	Open Access	Reasoning	Question Type	Question Source
MedQA[4]	12,723	✓	✗	Clinical vignettes	USMLE
MedXpertQA[7]	1,050	✓	✗	Clinical vignettes	Licensing exams
NEJM CPC[3]	302	✗	✗	Real diagnostic cases	NEJM CPC
`MedCaseReasoning`	13,989	✓	✓	Real diagnostic cases	PMC Case Reports

2. Methods

2.1. *Data curation*

Case reports are typically written to highlight a rare or complex disease. A core tradeoff exists between the *novelty* and *usefulness* of a case report – models should learn from difficult cases, but should be able to practically deduce the diagnosis from the presented information. To this end, we produced a LLM-based pipeline with the goal of maximizing case novelty and usefulness. As expert verification on each step is prohibitively expensive, we perform our clinician validation only on the outputs of the last step.

(1) **Data Provenance:** We started with the PMC Open Subset pulled from Jan 1st, 2005-April 27th, 2025. The query was filtered for free-access articles with article type "case report", which yielded 98,994 case reports. The article's text was extracted from the main body in the XML file, along with the PMCID, publication date, title, journal name, and article link.

(2) **Candidate Selection:** Next, we excluded case reports without the word *differential* (ie. a discussion on differential diagnoses), which reduced the set to 28,313 case reports. These initial candidate case reports were converted into diagnostic QA with *o4-mini*, a strong reasoning model that can be run at scale. When the dataset was built, *o4-mini* offered strong reasoning at a cost that made processing tens of thousands of articles feasible; the pipeline is model-agnostic and can be rerun with newer models. The full prompt is available in the Supplementary Material and instructs models to identify the core conflict within the case report and provide an information cutoff that would ensure the presented information does not give away the final diagnosis. The reasoning and diagnosis are extracted from the full report, so reference reasoning can occasionally cite findings after the cutoff (e.g., histopathology; see Section 5). In parallel, each candidate case report was scored according to the following criteria: 1) thoroughness of case presentation, 2) presence of explicit differential diagnosis, 3) dependence on integrative clinical reasoning, 4) transparency of diagnostic reasoning process, and 5) presence of stated final diagnosis. These criteria were developed by clinicians, and *o4-mini* was used to produce the actual scores. Criteria 1, 3, and 4 are given out of 1-5 points, whereas criteria 2 and 5 are yes/no responses. We filtered out cases where 1) thoroughness of case presentation was either 1 or 2 (seriously deficient

Table 2. An example of comparing reasoning traces against the case report reasoning. The reasoning from the case report is extracted using an LLM as an enumerated list. Then, each element of the list is referenced against the model's reasoning trace, which gives us the "reasoning recall" for a given trace.

Reasoning	**LLaMA-3.1-8B-Instruct (Base)**	**LLaMA-3.1-8B-Instruct (SFT on `MedCaseReasoning`)**	**Groundtruth (from `MedCaseReasoning`)**
Trace	The patient has a history of Wilson disease, a genetic disorder that leads to copper accumulation in the body. However, the brain MRI shows no evidence of pathological copper deposition, and laboratory tests show normal copper metabolism – suggesting the current condition is not related to Wilson disease. The symptoms … are consistent with a psychotic disorder. The history of Wilson disease and orthotopic liver transplantation may have led to post-liver-transplant psychosis … Given the patient's symptoms and history, post-liver-transplant psychosis is the most likely diagnosis.	We first considered acute graft rejection due to non-adherence with tacrolimus and mycophenolate, but this was ruled out because tacrolimus trough levels were low. Next, tacrolimus neurotoxicity was excluded for the same reason. We also considered copper overload from Wilson disease, but MRI showed no pathological copper deposition. Delirium was excluded; finally, we evaluated substance-induced psychosis but found "no other medical condition or substance use disorder." With alternatives excluded, schizophrenia was suspected.	1. Wilson-disease recurrence — "no abnormalities, including pathological copper accumulation … no evidence of abnormal copper metabolism … psychosis due to recurrence of Wilson disease was ruled out." 2. Tacrolimus-induced psychosis — "Tacrolimus-induced psychosis was ruled out because the trough tacrolimus concentration had not increased before onset of psychosis, and psychosis persisted after the trough value decreased." 3. Substance-related or other medical causes — "No evidence of other medical conditions or a substance use disorder that could cause psychotic symptoms was found."
Answer	Post-liver-transplant psychosis	Schizophrenia	Schizophrenia
Coverage	**1/3 reasons recalled**	**3/3 reasons recalled**	**Groundtruth**

or major gaps), and also removed cases that either did not discuss 2+ plausible alternatives or did not state a final diagnosis. The full prompt is available in the Supplementary Material. After filtering, there remained 19,428 total case reports.

(3) **Quality Filter:** Next, to avoid model blind spots, we checked each of the generated case reports using a separate LLM (*gemini-2.5-pro*), which evaluated the generated case report's faithfulness to the source article and the plausibility of each case report. The prompt is also available in the Supplementary Material. We removed cases that had any flags, leaving a final total of 13,989 cases. We created an initial testing subset of 897 cases where the transparency and integrative diagnostic reasoning scores were at least 4 or 5.

The case reports used in `MedCaseReasoning` span over 800 different medical journals, with

the top 10 shown in the Supplementary Material. Additionally, the diagnostic case prompts from `MedCaseReasoning` are significantly longer and more detailed than the ones from MedQA (see Supplementary Material). Furthermore, case report publication dates are heavily enriched towards years after 2020, with over 16% after Jan 1 2024 (see Supplementary Material). Additionally, our pipeline can be regularly updated, incorporating new case reports with low marginal cost.

To validate the case prompts, diagnostic reasoning, and final diagnosis extracted from each case report, a team of four board-certified physicians reviewed a total of 100 randomly selected cases, with each case reviewed by a single physician. The physician compared the extracted case prompt, reasoning, and diagnosis against the source article and answered three yes/no questions on hallucinations, faithfulness of the diagnosis, and faithfulness and relevance of the reasoning. The full questions can be found in the Supplementary Material.

2.2. *Evaluating Diagnostic Accuracy in Models*

We evaluate each model's diagnostic correctness using LLM-as-a-judge, in accordance to previous literature (e.g.,[3,23]). We adopt the same prompt used in prior work[3] (see Supplementary Material), which has been validated to have high concordance with human raters, and use *gpt-4o-mini* as the judge model for its speed and accuracy.

Differential diagnoses typically contain between five to ten candidate diagnoses which are followed-up in a clinical setting. Also following the evaluation schema established by prior work,[3] each model is evaluated a total of 10 times (temperature of 0.8 and top-p of 0.95), and the pass@N performance is recorded.

As an external validation to the `MedCaseReasoning` Test Set, we also evaluate models on case reports from NEJM Clinicopathologic Conferences. In particular, we use a subset of 302 cases from previous works[3,13,18] as a gold-standard dataset of complex diagnostic cases. We manually extracted each case report with a personal license and fed the case presentation portion as the case prompt without any additional modifications. As NEJM CPC is not an open-access publication, we are not able to open-source this validation set. Additionally, no NEJM case reports were included in `MedCaseReasoning` or used in training any models.

2.3. *Evaluating Reasoning Traces in Models*

While some case studies include comprehensive reasoning steps, it is not guaranteed that case reports include an exhaustive set of reasoning steps. As such, we focus our study on evaluating how well models *recall* clinician-produced reasons (as opposed to evaluating *precision*). We define this metric as the "Reasoning Recall". (Note: We were not able to evaluate reasoning recall for *OpenAI o3* and other frontier models where the full reasoning traces are not available via API call. While they can produce reasoning tokens when prompted (i.e. with <think> tags), the comprehensive reasoning that is used to inform a diagnosis is not accessible. We do not score such prompted explanations, which are produced separately from the hidden reasoning behind the diagnosis and would not be comparable to the native traces of open models.)

Definition 1 (Reasoning Recall). Let N be the total number of cases. For each case i, let

$$R_i = \{\text{groundtruth reasons from case report}\},$$
$$T_i = \{\text{reasons from model reasoning trace}\}.$$

For case i, the recall rate is

$$c_i = \frac{|R_i \cap T_i|}{|R_i|}.$$

The *Reasoning Recall* is the average of these rates:

$$\text{RR} = \frac{1}{N}\sum_{i=1}^{N} c_i = \frac{1}{N}\sum_{i=1}^{N} \frac{|R_i \cap T_i|}{|R_i|}.$$

For example, a reasoning recall of 50% indicates that on average, a model's reasoning trace will cover half of the reasons provided by the case report. We evaluate the reasoning trace associated with its best-of-10 response, where we use the trace from a randomly selected correct response (or incorrect if there are no correct responses). The evaluation is implemented using an LLM-as-a-judge, where *o4-mini* is instructed to return a JSON with a decision on whether each of the groundtruth reasons are found in the reasoning trace. The prompt used is found in the Supplementary Material.

For example, in Table 2 the reference reasoning has three statements (Wilson-disease recurrence, tacrolimus-induced psychosis, and substance-related causes, each ruled out), so $|R_i| = 3$. The judge checks each statement against the full trace: the base *LLaMA-3.1-8B-Instruct* trace addresses only Wilson disease ($c_i = 1/3$), while the fine-tuned trace addresses all three ($c_i = 1$).

Validating the reasoning judge. We validate our LLM-as-a-judge for determining reasoning recall with verification from a board-certified physician. The physician was given N=33 cases and was asked to verify the LLM judge's decisions on a total of 89 pairs of groundtruth reasons and model thinking traces. The cases were randomly sampled across all evaluated models. For example, for a given case report that contained three reasoning steps, the physician was asked to cross-check each step against the entire model reasoning trace to see if it was considered. The physician found 84/89 (94.4%, 95% CI of 87.5%-97.6%) of pairs of reasons and thinking traces to be correctly assessed by the model, and a total of 31/33 (93.9%, 95% CI of 80.4%-98.3%) cases to be judged correctly on every one of their reasoning statements (i.e., the physician agreed with all of the judge's found/not-found decisions for that case).

2.4. *Supervised Fine-Tuning on Case Report Reasoning*

In addition to evaluating model reasoning traces, we explore whether fine-tuning on extracted reasons directly can improve model performance on diagnostic accuracy and reasoning recall. One technical challenge is that the extracted diagnostic reasoning is formatted as an enumerated list of summary points and quotations, rather than a cohesive reasoning trace. Naturally, we can use LLMs to "stitch" the list of points into reasoning traces without adding new information. Given the case prompt, reasoning points, and final diagnosis, the stitching model rewrites all points as one narrative, reusing quotations nearly verbatim and adding no reasoning of its own. We provide the prompt used to perform this step in the Supplementary Material. The stitching process may introduce biases if a stronger model introduces its own priors into the reasoning traces. To control for this, each training

Table 3. Performance of models on `MedCaseReasoning`'s test set (N=897). Each model is evaluated on reasoning recall and diagnostic accuracy (broken down into Pass@1, Pass@5, and Pass@10). Each cell reports the point estimate, with its bootstrapped 95% confidence interval (lower–upper bound) in subscript. Reasoning recall is not available for *OpenAI o3* as the reasoning traces are not provided via API. We additionally perform supervised fine-tuning (SFT) on three open-source models on the training split of `MedCaseReasoning` and find significant improvements in both reasoning recall and diagnostic accuracy (average absolute gains of 10.4 and 13.5 percentage points in reasoning recall and Pass@10, respectively).

MedCaseReasoning Test Set				
Base Models	*Reasoning Recall*	*Pass@1 Acc.*	*Pass@5 Acc.*	*Pass@10 Acc.*
OpenAI o3	N/A	$\mathbf{0.470}_{0.440\text{-}0.500}$	$\mathbf{0.609}_{0.579\text{-}0.639}$	$\mathbf{0.645}_{0.618\text{-}0.675}$
DeepSeek R1	$\mathbf{0.642}_{0.616\text{-}0.667}$	$0.320_{0.291\text{-}0.349}$	$0.447_{0.417\text{-}0.478}$	$0.480_{0.450\text{-}0.510}$
QwQ-32B	$0.590_{0.560\text{-}0.619}$	$0.272_{0.245\text{-}0.302}$	$0.371_{0.341\text{-}0.400}$	$0.398_{0.368\text{-}0.428}$
MedReason-8B	$0.407_{0.383\text{-}0.431}$	$0.248_{0.224\text{-}0.275}$	$0.331_{0.303\text{-}0.363}$	$0.382_{0.353\text{-}0.412}$
LLaMA-3.1-8B-Instruct	$0.451_{0.428\text{-}0.475}$	$0.161_{0.138\text{-}0.184}$	$0.281_{0.252\text{-}0.311}$	$0.332_{0.304\text{-}0.360}$
m1-7b-23k	$0.495_{0.440\text{-}0.551}$	$0.155_{0.133\text{-}0.177}$	$0.238_{0.211\text{-}0.264}$	$0.291_{0.262\text{-}0.321}$
Qwen-2.5-7B	$0.324_{0.301\text{-}0.347}$	$0.174_{0.152\text{-}0.197}$	$0.252_{0.223\text{-}0.279}$	$0.287_{0.259\text{-}0.316}$
Fine-Tuned Models	*Reasoning Recall*	*Pass@1 Acc.*	*Pass@5 Acc.*	*Pass@10 Acc.*
MedReason-8B (SFT)	$\mathbf{0.522}_{0.498\text{-}0.545}$	$\mathbf{0.303}_{0.276\text{-}0.331}$	$\mathbf{0.430}_{0.400\text{-}0.459}$	$\mathbf{0.501}_{0.470\text{-}0.532}$
LLaMA-3.1-8B-Instruct (SFT)	$0.485_{0.460\text{-}0.510}$	$0.278_{0.249\text{-}0.307}$	$0.411_{0.378\text{-}0.443}$	$0.479_{0.448\text{-}0.509}$
Qwen-2.5-7B (SFT)	$0.486_{0.461\text{-}0.510}$	$0.249_{0.221\text{-}0.278}$	$0.363_{0.335\text{-}0.394}$	$0.425_{0.394\text{-}0.455}$

set is stitched once, before training, by the *base* checkpoint of the model to be fine-tuned, so the fine-tuned model never generates its own training data. Only the training split is stitched; evaluation uses the enumerated reference reasoning. An example of a stitched reasoning trace can be found in the Supplementary Material.

Using supervised fine-tuning (SFT), we train two popular open-source base models, namely *Qwen-2.5-7B-Instruct* and *LLaMA-3.1-8B-Instruct*. Additionally, to test the marginal value of our reasoning dataset, we include *MedReason-8B*, a model based on *LLaMA-3.1-8B-Instruct* that was previously fine-tuned on a synthetic medical reasoning dataset. We updated each model with full-weight fine-tuning for three epochs on 8 NVIDIA H100 GPUs with a learning rate of 2e-5 and batch size of 256. All other hyperparameters are default according to the *verl* Python package implementation of SFT (*version v0.3.0.rc0*). We focus on 7–8B models to match existing open medical reasoning models (*MedReason-8B*, *m1-7b-23k*) and because models of this size can be run locally by health systems; larger models are left to future work.

3. Results

3.1. *Dataset Quality: Physician Validation of Extracted Cases*

We performed a human validation of the model-generated case prompts, diagnostic reasons, and final diagnoses on the `MedCaseReasoning` dataset. Our study included four U.S. board-certified

physicians reviewing a total of 100 cases. Each physician reported spending an average of 10-20 minutes per case. The figures below are the proportion of the 100 reviewed cases for which the reviewing physician answered "yes" to each question, with bootstrapped 95% confidence intervals; they are not an inter-rater agreement statistic, since each case was reviewed by a single physician. We found that 98% of the generated cases were free of hallucinations in the case prompt or diagnostic reasoning. We also report that 92% of final diagnoses were reported to have been faithful to the article and reasonably inferrable from the details of the case prompt. Finally, 93% of diagnostic reasoning steps were faithful to the case report and clinically relevant. This review assessed the faithfulness and relevance of the extracted reasoning, not its completeness: it did not ask whether every alternative discussed in the source article was captured. The results are found in the Supplementary Material.

3.2. *Benchmark Results: Diagnostic Accuracy*

We evaluate the following combination of frontier reasoning models and popular open-sourced models: *OpenAI o3*, *DeepSeek R1*, *QwQ-32B*, *MedReason-8B*, *m1-7b-23k*, *LLaMA-3.1-8B-Instruct*, and *Qwen-2.5-7B-Instruct*.

The models are each evaluated on pass@1, pass@5, and pass@10 accuracy, listed in Table 3 (NEJM CPC results in the Supplementary Material). Overall, we found that diagnostic reasoning remains a difficult task for top-performing LLMs. *OpenAI o3* performs significantly better than the rest of the models, with 64.5% pass@10 accuracy on the `MedCaseReasoning` test set, whereas *DeepSeek R1* achieves 48.0%. In comparison, *OpenAI o3* and *DeepSeek R1* achieve similar scores of 62.3% and 43.7% pass@10 accuracy on NEJM CPC, signaling the validity of `MedCaseReasoning` as an open-access alternative evaluation set for diagnostic cases.

We find that supervised fine-tuning significantly improves model performance across both test sets. For example, on the `MedCaseReasoning` test set, *MedReason-8B* improves by 31% (11.9 percentage points, from 38.2% to 50.1%) pass@10 accuracy, outperforming *DeepSeek R1*. Of note, this model also improves by 18% (5.3 percentage points, from 29.5% to 34.8%) on NEJM CPC and outperforms *QwQ-32B*. This provides evidence of the generalizability of `MedCaseReasoning` training data, as NEJM CPC consists of out-of-distribution, hand-crafted diagnostic cases that are often much longer in length.

3.3. *Benchmark Results: Reasoning Recall*

We evaluate six of seven models that provide reasoning traces on reasoning recall and find that the top model, *DeepSeek R1*, covers 64.2% of the reasoning steps provided in case reports. We also found that *MedReason-8B*, a model explicitly trained on synthetically generated medical reasoning traces, did not have substantially higher reasoning recall, and even had lower recall than *LLaMA-3.1-8B-Instruct*'s base model. However, after SFT on `MedCaseReasoning`, each of the base models improved significantly on reasoning recall. For example, *MedReason-8B* improved by 28% (11.5 percentage points) and *Qwen-2.5-7B-Instruct* improved by 50% (16.2 percentage points).

Qualitatively, the reference reasoning that base models missed was often candidate diagnoses never considered or exclusionary symptoms for common diagnoses (we leave quantifying these error types to future work). For example, one case report focused on a case of Adult-onset Still's

disease (AOSD), which presented very similarly to Rocky Mountain Spotted Fever (RMSF). The model (*m1-7b-23k*) made a diagnosis for RMSF, missing reasoning about the patient presentation from the case report: "We feel that our patient's symptoms were not due to RMSF, because the rash was salmon colored and worsened with the fever spikes.". We found a significant correlation between model performance and reasoning recall (Pearson r=0.710, p=0.0485), indicating the value of measuring reasoning steps as a proxy for model performance itself. Furthermore, we also observe a significant correlation between the length of the model's reasoning trace and the reasoning recall (r=0.790, p=0.0196). Since recall does not penalize extraneous content, and the scored trace is selected conditional on correctness, both correlations should be interpreted with caution (Section 5).

4. Discussion

`MedCaseReasoning` addresses a critical gap in current medical benchmarks: the assessment of diagnostic accuracy without scrutinizing the underlying reasoning process. SFT with the `MedCaseReasoning` training dataset demonstrates that smaller models, such as *Llama 3.1 8B* and *Qwen 2.5 7B*, can achieve diagnostic accuracy comparable to or exceeding that of larger models like *QwQ-32B* and *DeepSeek-R1* after SFT. While prior studies ([22,23]) have explored learning reasoning from synthetic traces generated by more powerful models, our research is the first to demonstrate the efficacy of training directly on diagnostic reasoning derived from clinician-written case reports.

Compared to established benchmarks like MedQA, where leading models such as *GPT-4o* have already achieved over 90% accuracy, diagnostic performance on `MedCaseReasoning` currently peaks at 64.5%, and it correlates strongly with performance on NEJM CPC (Figure 2), while being open-access and substantially larger.

5. Limitations

Our study presents several limitations. First, in order to enable scalable, systematic evaluation, we use LLMs to score model adherence to reasoning traces and diagnostic accuracy. LLMs are known to hallucinate facts and exhibit certain biases, which may be propagated when LLMs are used as judges as well. Our validation of the reasoning judge relied on a single physician and 33 cases (89 reasoning statements); a larger sample annotated by multiple physicians, with inter-annotator agreement, would give a more precise estimate of judge reliability. Some case reports may lack sufficient detail for a definitive diagnosis or present trivial cases. The QA conversion process can also introduce variability, where case details may be inadvertently left out or hallucinated. While we implemented a clinician-validated filtering pipeline for the test set to ensure reasoning statements are grounded in the case presentation, some intractable cases (eg. ones where the diagnosis cannot be made without certain information) or trivial cases (eg. ones where the diagnosis is given away in the prompt) still persist. In particular, some reference statements depend on findings withheld from the case prompt (e.g., histopathology), which our filters do not explicitly remove, and extraction may be incomplete, since physicians checked faithfulness and relevance but not completeness. Newer extraction models may reduce both errors.

Second, `MedCaseReasoning` captures the case presentation at a single timestep before asking for a final diagnosis. It does not reflect the iterative, multi-stage nature of real-world clinical diagnosis, which involves refining differential diagnoses based on evolving information from tests,

imaging, and treatment responses. Case reports also over-represent rare or complex presentations, so performance should not be read as everyday diagnostic accuracy.

Third, our reasoning recall metric only captures clinical reasoning provided within case reports. Diagnostic reasoning is inherently subjective, and while our extensive training corpus aims to encompass diverse diagnostic standards, the alignment metric should be interpreted as adherence to observed clinical reasoning patterns across a diverse range of clinicians rather than an absolute, single gold standard. Reasoning recall also measures coverage only and does not penalize unsupported or extraneous reasoning; consistent with this, trace length correlates with recall ($r = 0.790$), and a precision-style or length-controlled metric would complement it. Scoring a correct generation whenever one exists also couples recall to pass@10; recall on a fixed generation would separate the two.

Fourth, some open-access case reports may appear in LLM pretraining data, which could inflate scores, although low frontier-model accuracy, enrichment for post-2020 reports, and agreement with the license-restricted NEJM CPC suggest memorization does not dominate.

The models we train and open source are for research purposes only and are not intended to be used for diagnosis of real patients. Accordingly, the misuse of these models risks inaccurate medical diagnoses and potential patient harm.

6. Conclusion

We introduced `MedCaseReasoning`, an open-access benchmark and training set of 13,989 diagnostic cases built from clinician-written case reports, with reference reasoning that lets models be evaluated on how they reach a diagnosis, not only on the diagnosis. Current reasoning LLMs leave substantial room for improvement on both: the best model reaches 64.5% pass@10 accuracy, and the best open model recalls 64.2% of the reference reasoning. Fine-tuning 7–8B models on reasoning traces derived from `MedCaseReasoning` improves both diagnostic accuracy and reasoning recall, and these gains transfer to the held-out NEJM CPC benchmark. Because the pipeline can be rerun on new case reports, `MedCaseReasoning` can grow over time.

Data and Code Availability

Dataset: https://huggingface.co/datasets/zou-lab/MedCaseReasoning
Code and supplement:
https://github.com/kevinwu23/Stanford-MedCaseReasoning

Institutional Review Board (IRB)

This research used only publicly available, open-access case reports and involved no patient recruitment or interaction; physician review used no identifiable data beyond the published literature.

References

1. AMA, AMA: Physician enthusiasm grows for health care AI undefined american medical association (2025).

2. E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, A. Rodman and J. H. Chen, Large language model influence on diagnostic reasoning, *JAMA Network Open* **7**, p. e2440969 (2024).
3. D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni *et al.*, Towards accurate differential diagnosis with large language models, *Nature* , 1 (2025).
4. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *arXiv [cs.CL]* (2020).
5. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, Measuring massive multitask language understanding, *arXiv* (2020).
6. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. A. Y. Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam and V. Natarajan, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).
7. Y. Zuo, S. Qu, Y. Li, Z. Chen, X. Zhu, E. Hua, K. Zhang, N. Ding and B. Zhou, MedXpertQA: Benchmarking expert-level medical reasoning and understanding, *arXiv [cs.AI]* (January 2025).
8. J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang and W. Yin, Large language models for mathematical reasoning: Progresses and challenges, *arXiv* (2024).
9. H. Ding, Z. Fan, I. Guehring, G. Gupta, W. Ha, J. Huan, L. Liu, B. Omidvar-Tehrani, S. Wang and H. Zhou, Reasoning and planning with large language models in code development, in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (ACM, New York, NY, USA, August 2024).
10. V. L. Patel, J. F. Arocha and J. Zhang, Thinking and reasoning in medicine, *The Cambridge handbook of thinking and reasoning* **14**, 727 (2005).
11. Q. Jin, F. Chen, Y. Zhou, Z. Xu, J. M. Cheung, R. Chen, R. M. Summers, J. F. Rousseau, P. Ni, M. J. Landsman, S. L. Baxter, S. J. Al'Aref, Y. Li, A. Chen, J. A. Brejt, M. F. Chiang, Y. Peng and Z. Lu, Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine, *NPJ Digit. Med.* **7**, p. 190 (July 2024).
12. PubMed Central, PMC open access subset https://pmc.ncbi.nlm.nih.gov/tools/openftlist/, Accessed: 2025-5-15.
13. Z. Kanjee, B. Crowe and A. Rodman, Accuracy of a generative artificial intelligence model in a complex diagnostic challenge, *Jama* **330**, 78 (2023).
14. Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen and X. Lu, PubMedQA: A dataset for biomedical research question answering, *arXiv [cs.CL]* (September 2019).
15. S. L. Fleming, A. Lozano, W. J. Haberkorn, J. A. Jindal, E. P. Reis, R. Thapa, L. Blankemeier, J. Z. Genkins, E. Steinberg, A. Nayak, B. S. Patel, C.-C. Chiang, A. Callahan, Z. Huo, S. Gatidis, S. J. Adams, O. Fayanju, S. J. Shah, T. Savage, E. Goh, A. S. Chaudhari, N. Aghaeepour, C. Sharp, M. A. Pfeffer, P. Liang, J. H. Chen, K. E. Morse, E. P. Brunskill, J. A. Fries and N. H. Shah, MedAlign: A clinician-generated dataset for instruction following with electronic medical records, *arXiv [cs.CL]* (August 2023).
16. S. Bedi *et al.*, MedHELM: Holistic evaluation of large language models for medical tasks, *arXiv preprint arXiv:2505.23802* (2025).
17. B. K. Nayak, The significance of case reports in biomedical publication, *Indian J. Ophthalmol.* **58**, 363 (September 2010).
18. Gemini, Gemini: A family of highly capable multimodal models, *arXiv [cs.CL]* (December 2023).
19. E. Goh, R. J. Gallo, E. Strong, Y. Weng, H. Kerman, J. A. Freed, J. A. Cool, Z. Kanjee, K. P. Lane, A. S. Parsons, N. Ahuja, E. Horvitz, D. Yang, A. Milstein, A. P. J. Olson, J. Hom, J. H. Chen and A. Rodman, GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized

controlled trial, *Nature Medicine* **31**, 1233 (2025).
20. E. Strong, A. DiGiammarino, Y. Weng, A. Kumar, P. Hosamani, J. Hom and J. H. Chen, Chatbot vs medical student performance on free-response clinical reasoning examinations, *JAMA Intern. Med.* **183**, 1028 (September 2023).
21. T. Savage, A. Nayak, R. Gallo, E. Rangan and J. H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, *npj Digital Medicine* **7**, p. 20 (2024).
22. J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou and B. Wang, HuatuoGPT-o1, towards medical complex reasoning with LLMs, *arXiv [cs.CL]* (December 2024).
23. J. Wu, W. Deng, X. Li, S. Liu, T. Mi, Y. Peng, Z. Xu, Y. Liu, H. Cho, C.-I. Choi, Y. Cao, H. Ren, X. Li, X. Li and Y. Zhou, MedReason: Eliciting factual medical reasoning steps in LLMs via knowledge graphs, *arXiv [cs.CL]* (April 2025).