OncoCoT: A Temporal-causal Chain-of-Thought Dataset for Oncologic Decision-Making

Peiru Yang, Yudong Li*, Shiting Wang, Xinyi Liu, Haotian Gan, Xintian Li, Qingyu Gao, Yongfeng Huang

Department of Electronic Engineering, Tsinghua University
liyudong@tsinghua.edu.cn

Abstract

Long Chain-of-Thought (CoT) reasoning has shown great promise in complex reasoning tasks, but its application to medical decision-making presents unique challenges. Unlike structured tasks relying on static verification frameworks, medical decision-making requires dynamic validation through longitudinal clinical outcomes, exhibiting temporal-causal dependencies that complicate the verification of reasoning processes. Therefore, we introduce a novel data construction framework specifically designed for medical decision-making. First, the framework analyzes real-world clinical cases to construct a timeline of medical events and identify critical decision points, including examination, diagnosis, and treatment. Subsequently, it employs a clinical causality-aware strategy to generate decision-making questions at the identified points, along with reasoning traces and corresponding answers. Finally, information drawn from future nodes serves as clinical logic-constrained criteria to re-evaluate and refine the soundness of the generated reasoning and responses. Building on this, we present OncoCoT, an oncologic decision-making dataset derived from clinical records over the past four years across eight common cancer types. Furthermore, we distill a subset of OncoCoT into a dedicated benchmark, OncoEval, to facilitate systematic evaluation of clinical reasoning capabilities in LLMs. Evaluation results show that existing state-of-the-art reasoning models, such as Deepseek-r1 and GPT-o3, exhibit limited capability in addressing clinical problems in OncoEval, highlighting the need for further improvement.

Datasets — https://github.com/ydli-ai/OncoCoT

Introduction

Long Chain-of-Thought (Long CoT) technology, exemplified by GPT-o1 (OpenAI 2025) and Deepseek-r1 (Guo et al. 2025), has significantly enhanced complex reasoning capabilities in structured tasks such as mathematical deduction and code generation through multi-step reasoning guidance. This progress primarily benefits from formalized verification standards and well-defined evaluation paradigms. However, Long CoT faces unique challenges when applied to medical decision-making, a domain with strong temporal dependencies in both task formulation and outcome validation. Medical decision-making is an evidence-based reasoning process that requires dynamic evaluation through longitudinal tracking of clinical outcomes rather than static, single-time-point verification frameworks. This challenge is exacerbated in oncology due to complex disease trajectories, diverse treatment options, and long-term clinical courses. Moreover, privacy constraints and high costs of expert annotation lead to severe shortages of high-quality data.

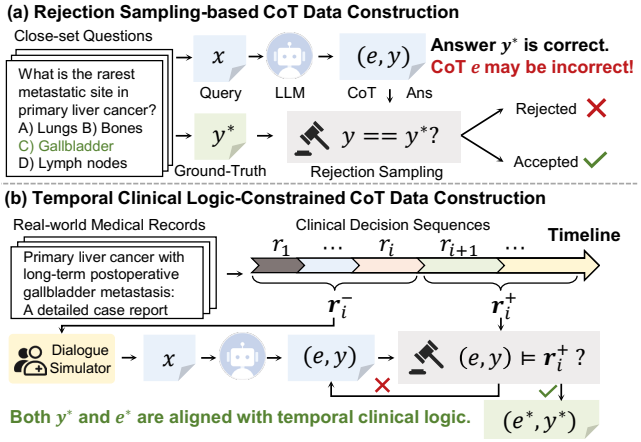


Figure 1: Comparison of two Chain-of-Thought data construction methods for clinical decision support: (a) Rejection sampling-based CoT for closed-set QA, which filters LLM-generated reasoning chains via ground-truth alignment; (b) Temporal clinical logic-constrained CoT that integrates the real-world medical timelines to ensure step-by-step reasoning adheres to diagnostic/treatment sequences.

While numerous datasets support medical tasks, few are designed to reflect the evolving nature of real-world clinical decision-making. Existing Medical datasets (Jin et al. 2020; Liu et al. 2023) typically convert medical exam data to QA format through rejection sampling. Although rich in medical knowledge, they are restricted to single-time-point entries, failing to simulate the evolving, longitudinal process of real-world clinical decision-making. Notably, Chen et al. (2024) pioneers the construction of medical Q&A into Long CoT formats, achieving preliminary multi-step diagnostic reason-

*Corresponding author.

ing. However, it remains constrained by the flat structure of exam questions, relying on static answer verification while overlooking the correctness of the CoT reasoning traces. In summary, existing medical datasets may fall short in evaluating the full capabilities of modern LLMs. These limitations underscore the need for clinically realistic, logically complex datasets to support medical decision-making tasks.

In this work, we present OncoCoT, building on a proposed data construction framework that is capable of simulating **temporal-causal** clinical decision-making process by analyzing the causal dependencies between medical events and decisions in real medical records. This framework systematically constructs long CoT reasoning data with clinical logic constraints. Specifically, we extract the timeline of medical events from the cases, and simulate clinician-patient interactions at specific timepoints while constraining reasoning paths to objectively reflect subsequent medical events. We focus exclusively on formulating questions around three critical types of decision points in clinical workflows: examination, diagnosis, and treatment, whereby our framework ensures that the constructed reasoning tasks impart meaningful medical decision-making. As shown in Fig. 1, existing medical data construction methods typically rely on multiple-choice questions to expand reasoning components, filtering data by verifying answer correctness, from which our approach fundamentally differs, constructing data from comprehensive real-world clinical cases. This paradigm shift ensures that our data align with authentic clinical pathways and enables temporal-causal evaluation of models.

Since our framework reveals the limitations of LLMs in clinical decision-making, we introduce OncoEval, a benchmark built from real-world diagnostic and treatment records covering eight common cancer types, serving as a comprehensive evaluation suite for clinical reasoning models. Our benchmark demonstrates that existing state-of-the-art deep reasoning models, including Deepseek-r1 and GPT-o1, exhibit limited capability in addressing our constructed problems, highlighting new directions for model development. To evaluate the effectiveness of our proposed OncoCoT dataset, we also set up an experiment to compare the models fine-tuned using different data construction methods. The results show that OncoCoT has the greatest improvement in model performance compared to other methods.

The contribution of this work can be summarized as:

- We propose a clinical logic-constrained CoT data construction framework based on spatiotemporal decision sequences, which designs reasoning tasks around critical medical decision nodes including examination, diagnosis, and treatment, ensuring the quality and rigor of the created reasoning data that reflects real-world clinical decision-making processes.
- We establish OncoEval, the first benchmark for common-cancer diagnosis and treatment with temporal-causal evaluation capability, and show that SOTA reasoning models still struggle on its tasks.
- Models fine-tuned on OncoCoT achieve the largest performance gains over those trained with alternative data-construction methods, validating its effectiveness for advancing clinical-reasoning models.

Related Work

CoT Dataset Construction

In the medical domain, integrating CoT reasoning into dataset construction enables model evaluation beyond decision accuracy, emphasizing plausibility and clinical validity of intermediate steps (Chen et al. 2024). A human–LLM hybrid pipeline has been proposed to build expert-verified CoT datasets for interpretable clinical AI (Ding et al. 2025), though its reliance on costly expert input limits scalability. Liu et al. (2024) present a hierarchical expert verification framework with a sparse mixture-of-experts model to improve interpretability in medical VQA. Nachane et al. (2024) propose CoT prompting to emulate step-by-step clinical reasoning. However, such datasets are often based on simplified QA tasks or exams, lacking authentic temporal and causal reasoning found in real-world clinical settings.

Medical LLMs and Benchmarks

Due to recent advances in LLM capabilities, many medical LLMs have emerged to support clinical tasks (Yu et al. 2019; Huang, Altosaar, and Ranganath 2019). Zhang et al. (2023b) introduce AlpaCare, leveraging GPT-4 to expand expert seed data and build the MedInstruct-52K dataset for improved medical QA. Other studies enhance medical LLMs by constructing doctor-patient dialogue datasets, integrating RAG, and applying RLAIF, achieving progress in dialogue, knowledge use, and response quality (Chen et al. 2023; Li et al. 2023; Zhang et al. 2023a; Devi et al. 2024). However, pretraining-based models remain costly, less adaptable, and limited by proprietary data, while fine-tuning methods often use rigid QA formats unsuited for complex real-world clinical tasks. Jin et al. (2021) propose MedQA, based on the USMLE, to evaluate clinical reasoning via multiple-choice questions. Other benchmarks such as MedMCQA, PubMedQA, MLEC, and MedNLI use exam questions and medical literature to test LLMs on knowledge recall and evidence-based reasoning (Pal, Umapathi, and Sankarasubbu 2022; Jin et al. 2019; Li, Zhong, and Chen 2021; Johnson et al. 2016; Shivade 2017). A more detailed discussion of related work can be found in the appendix.

Data Construction Framework

Motivation and Challenges

Existing medical datasets for LLMs primarily rely on multiple-choice or short-answer QA formats, mostly derived from textbook-style question banks, curated knowledge bases, or single-visit clinical records (Luo et al. 2022; Ben Abacha and Demner-Fushman 2019). While these benchmarks provide a convenient and quantifiable way to assess the knowledge storage and memory capabilities of LLMs, they offer only limited insights into reasoning and decision-making ability under real-world clinical conditions. Although some efforts have attempted to construct long CoT using multiple-choice or short-answer formats (Chen et al. 2024), they generally lack supervision for CoT

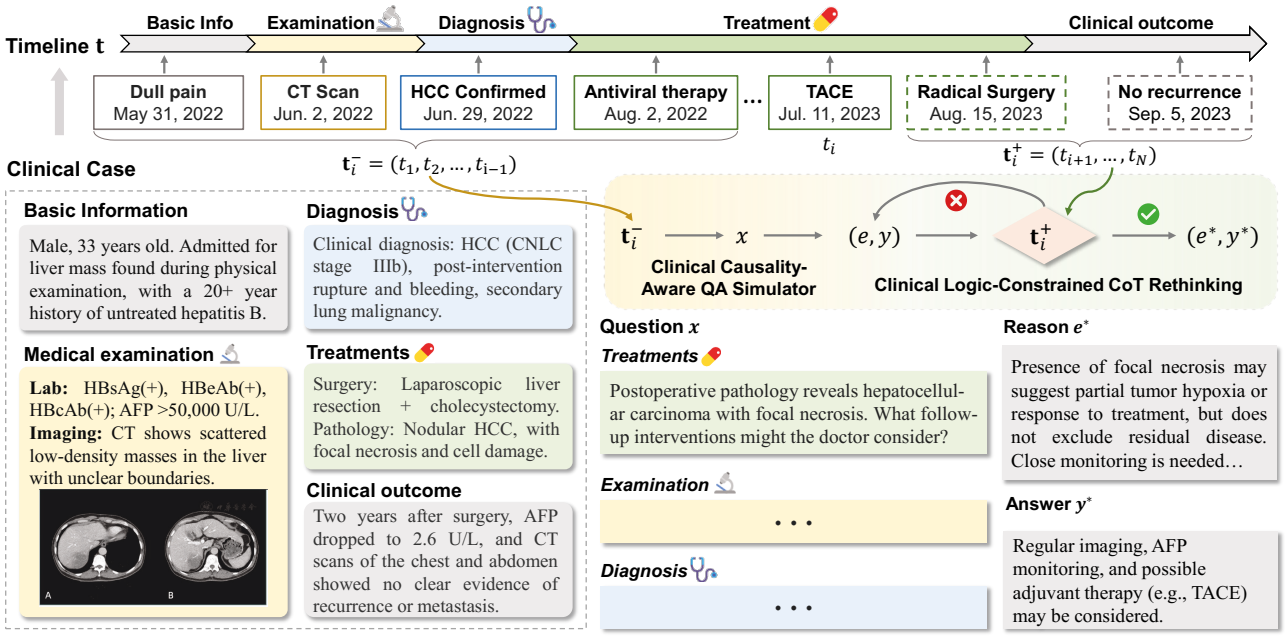


Figure 2: Workflow of the proposed clinical decision-support data-construction framework, analyzing temporal-causal dependencies between medical events to simulate medical reasoning in authentic clinical settings. It integrates timeline and decision points identification, clinical causality-aware QA simulator, and clinical logic-constrained CoT Rethinking.

quality and fail to capture the temporal-causal nature of clinical reasoning. In real clinical settings, the decision-making happens over time and builds on changing information, which static or isolated questions cannot capture. Therefore, our goal is to construct medical decision-making tasks that reflect the complexities of real-world long-term clinical workflows, as well as to test a model's ability to interpret evolving information and handle real-world clinical decision scenarios, including ordering examinations, making diagnoses, and proposing treatment plans.

Framework Overview

Our framework aims to construct temporally grounded and causally coherent reasoning data from real-world clinical cases, enabling more faithful evaluation and enhancement of language models in complex medical scenarios. Unlike traditional static QA benchmarks, we focus on simulating how physicians reason through evolving clinical information, emphasizing the temporal-causal and decision-centric nature of medical thinking.

Given a raw clinical case, we extract a full medical event timeline $\mathbf{t} = (t_1, t_2, \ldots, t_N)$, where each t_i denotes a medical event which depends on its preceding context $\mathbf{t}_i^- = (t_1, t_2, ..., t_{i-1})$, forming a directed temporal chain of decision-making. Let $\mathbf{t}_i^+ = (t_{i+1}, \ldots, t_N)$ denote the sequence of subsequent medical events of t_i. From this timeline, we identify a subset of clinical decision-relevant time points $\mathcal{T}_{\text{key}}$ as anchor points for QA pair construction.

At each decision-relevant point $t_i \in \mathcal{T}_{\text{key}}$, a proxy model M is placed in the same informational position as a clinician at that moment, with only $\mathbf{t}_i^-$ as its available context to generate a decision-specific question x, simulating a physician's query about how to proceed at this stage. A deep-reasoning model M_{cot} is then prompted to generate a reasoning-answer pair (e, y) in response to a decision-specific question x associated with t_i. Since the full case provides the actual decision t_i and the future events $\mathbf{t}_i^+$, it can be used to evaluate the model's reasoning and serve as reference for subsequent refinement.

By aligning language model behavior with the structure of real-world clinical workflows, our framework provides a systematic approach for constructing and supervising medical long CoT reasoning.

Timeline and Decision Points Identification

We begin by extracting a structured medical event timeline $\mathbf{t}$ defined in the previous section from raw clinical cases using the proxy LLM M. This process yields a temporally ordered sequence of decision points, denoted as $(t_1, t_2, \ldots, t_i, \ldots, t_N)$, where each t_i corresponds to a specific clinical observation or action that occurred during the case. Among the full set of timeline events, we identify a subset of clinically critical decision points that reflect key stages in the medical reasoning process: *examination*, *diagnosis*, and *treatment*. Each of these categories corresponds to a distinct type of clinical decision: selecting appropriate diagnostic tests (examination), interpreting findings to conclude (diagnosis), and proposing concrete medical interventions (treatment). With slight abuse of notation, we denote a selected decision point of interest as t_i, which will serve as the target for downstream simulation and reasoning.

Clinical Causality-Aware QA Simulator

Given a decision point t_i in $\mathbf{t}$ and its preceding medical context $\mathbf{t}_i^-$, we prompt the proxy LLM M to simulate the clinical reasoning process. Specifically, M first generates a decision-specific question x that reflects the type of decision required at t_i. This models the real-world process in which a clinician, faced with a complex case history, must decide on the next course of action to resolve clinical uncertainty. The subsequent medical events t_i and $\mathbf{t}_i^+$ serve as the ground-truth trajectory, representing what decisions were actually taken by a clinician in the original case.

Next, we prompt a deep reasoning LLM M_{cot} with x and its corresponding context $\mathbf{t}_i^-$ to produce a reasoning-answer pair (e, y), where e and y denote the step-by-step chain of thought and the preliminary answer to x generated by M_{cot}, respectively. However, since M_{cot} is not aware of the subsequent timeline $\mathbf{t}_i^+$, the initial (e, y) pair may diverge from clinically grounded logic or real-world outcomes. This motivates the need for a following correction process to ensure alignment with true clinical reasoning trajectories.

Clinical Logic-Constrained CoT Rethinking

Given the future medical events t_i and $\mathbf{t}_i^+$ and the initial reasoning path e, we prompt model M_{cot} again to perform deep reasoning under real-world clinical constraints. The goal is to generate a revised reasoning-answer pair (e^*, y^*) that aligns with the clinical logic embedded in t_i and $\mathbf{t}_i^+$.

This rethinking process is conducted iteratively, where M_{cot} is repeatedly guided to refine both the reasoning process e^* and the final decision y^* until they are jointly consistent with the ground-truth future trajectory. In this paper, we verify the consistency and correctness of the reasoning-answer pair (e^*, y^*) by prompting the proxy LLM M. At minimum, the model performs one round of rethinking to incorporate clinical logic constraints. This ensures that the model's reasoning is not only factually correct but also clinically plausible, reflecting the causal and temporal dependencies inherent in authentic medical decision-making.

Dataset and Benchmark

OncoCoT Dataset

Utilizing our proposed data construction framework, we build the first comprehensive dataset targeting real-world medical decision-making across multiple cancer types, designed for temporal-causal model training and evaluation. We systematically collected 734 Chinese case reports from an open-access public medical case platform, spanning the past 4 years. This data source encompasses eight prevalent cancer types: *Gastric Cancer*, *Lung Cancer*, *Hepatic Cancer*, *Breast Cancer*, *Prostate Cancer*, *Cervical Cancer*, *Pancreatic Cancer*, and *Colorectal Cancer*.

Each case report provides detailed longitudinal information including patient presentation, diagnostic workup, treatment progression, and clinical outcomes, forming the foundation for our CoT data construction process. Specifically, we focus on three critical decision nodes that are intimately connected to medical decision-making: examination, diagnosis, and treatment. Through our framework, we generate 2,320 question-answer pairs with CoT reasoning, each designed to capture the nuanced reasoning patterns observed in authentic clinical scenarios. This dataset aims to enhance model capabilities in medical decision-making by providing training scenarios that mirror actual clinical practices.

As shown in Fig.3-(a), our dataset contains rich metadata encompassing comprehensive patient information, pathology details, and clinical documentation. To ensure quality, we conduct statistical analysis of the data distribution. Fig.3-(b) demonstrates that our dataset encompasses diverse subtypes ranging from common presentations like *ductal adenocarcinoma* to rare variants such as *signet-ring cell carcinoma*, while covering different clinical stages (I-IV) to ensure comprehensive representativeness. Fig.4 reveals well-balanced demographic characteristics with case concentrations peaking in the 50-79 age range, aligning with typical cancer epidemiological patterns, and gender-specific distributions reflecting real-world prevalence. In addition, a more detailed breakdown of gender distribution across individual cancer types is provided in the Appendix. The majority of our collected cases span the recent three years, ensuring clinical relevance and contemporary practice validity.

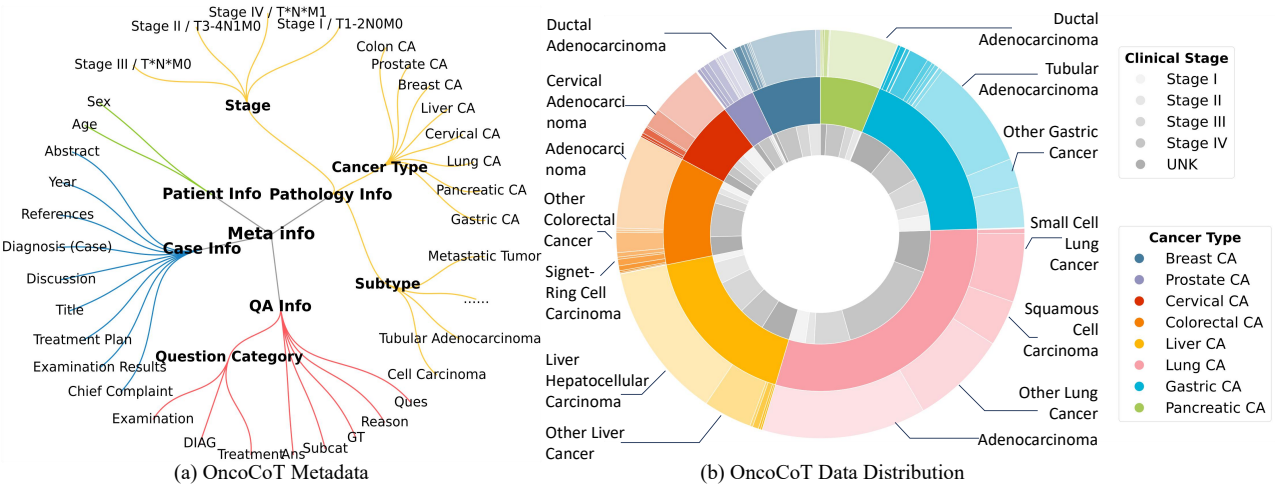
Ethical Considerations. Our proposed OncoCoT dataset is constructed exclusively from open-access academic case reports published in peer-reviewed medical journals, thereby eliminating privacy concerns and copyright infringement issues. The questions in our dataset are based on scenario reconstruction, while the ground-truth answers are derived from medical decisions in the cases through information extraction without introducing additional bias or subjective interpretation. Therefore, our dataset maintains ethical standards consistent with the original paper.

OncoEval Benchmark

Current medical LLM evaluation paradigms predominantly rely on multiple-choice questions and knowledge-based Q&A formats, which present limited challenges and fail to represent the complexity of real-world clinical applications. These conventional methods fail to capture the dynamic and multi-step reasoning processes inherent in clinical decision-making, particularly in cancer scenarios.

To address this evaluation gap and provide a more rigorous assessment of LLM performance in authentic clinical decision-making tasks, we construct OncoEval based on our proposed dataset. Building upon the case reports spanning the eight cancer types in our dataset, we select representative samples from recent 2 years (2023-2024) to construct our benchmark, ensuring temporal relevance and alignment with current clinical guidelines.

Our benchmark leverages the inherent complexity and diversity of our OncoCoT dataset to evaluate models across the full spectrum of oncological scenarios. The benchmark is structured around three distinct tasks. These tasks represent the most consequential and error-prone aspects of clinical decision-making, frequently cited in adverse event analyses and guideline development frameworks (Jackson and Feder 1998; Hooftman et al. 2024; Sandmann et al. 2025). The details of each task are shown as follows:



(a) OncoCoT Metadata (b) OncoCoT Data Distribution

Figure 3: Metadata of OncoCoT, showcasing cancer subtypes, clinical stages (I–IV), and diagnostic categories. It includes patient demographics, treatment plans, and pathology details for cancers such as breast, prostate, lung, and gastric cancer.

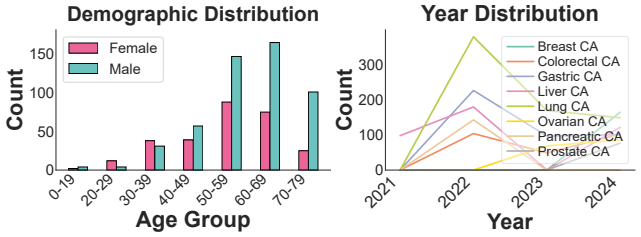


Figure 4: Age, gender, and year distribution of various cancer types, showing case counts across different age groups and disparities between female and male patients.

Examination Recommendation Task assesses the model's effectiveness in designing an examination workflow. This task challenges the model to propose a comprehensive set of necessary examinations, including their sequence and priorities, while avoiding redundant procedures to reduce patient burden and medical costs.

Clinical Diagnosis Task evaluates the model's clinical reasoning capabilities and is composed of two key parts: the primary diagnosis and the differential diagnosis. For the primary diagnosis, the model must accurately identify the disease, including its specific type and stage. For the differential diagnosis, the model is required to list other potential conditions that could explain the patient's symptoms but are excluded based on clinical evidence, a critical step in preventing misdiagnosis. All diagnostic conclusions must be supported by relevant clinical evidence from the case.

Treatment Planning Task evaluates the model's accuracy and safety in formulating treatment regimens. This task requires the model to devise plans that comply with clinical guidelines and are tailored to patient-specific conditions, while also providing clear decision-making reasoning to support its recommendations for therapeutic modalities, medications, dosages, and procedural interventions.

Evaluation Mechanisms

To comprehensively assess model performance on our benchmark, we designed a unified evaluation framework. As medical decisions rarely have a single, binary correct answer, we adopt a component-based comparison approach rather than seeking exact matches. In this method, both the model's output and the ground-truth reference answer are decomposed into a set of discrete, evaluable components, which we refer to as 'items'.

We then evaluate the model's performance using Precision, Recall, and Accuracy, calculated by comparing the items generated by the model to those mentioned in the reference answer. A True Positive (TP) represents an item that is present in both the model's output and the reference. A False Positive (FP) is an item suggested by the model but not found in the reference, while a False Negative (FN) is an item in the reference that the model failed to recommend. In this context, Precision measures the proportion of the model's correct recommendations, Recall is the proportion of ground-truth recommendations that the model successfully identified, and Accuracy provides a holistic score of the model's alignment with the reference.

Note that we do not include True Negatives (TN) in our evaluation, as the number of correctly omitted items in a medical scenario is practically infinite. This framework provides a consistent and objective assessment across all three tasks and is amenable to automated evaluation. We employ an evaluator LLM to perform the above extraction and comparison task. The effectiveness and consistency of this approach are also evaluated in the experiment section. To ensure scoring consistency and minimize model-subjective interpretation, the process is guided by detailed, structured scoring rubrics. See the Appendix in detail.

Model	Examination			Diagnosis			Treatment		
	Precision	Recall	Acc.	Precision	Recall	Acc.	Precision	Recall	Acc.
Closed-source Models									
GPT-o3	24.23	**72.01**	21.66	42.37	58.77	33.01	50.03	69.89	41.43
GPT-4.1	33.80	62.28	27.55	57.70	55.81	**39.81**	56.89	67.03	44.96
Gemini-2.5-Pro	33.23	65.21	27.78	55.67	54.93	38.75	56.72	62.66	42.73
Gemini-2.0-Flash	27.40	64.71	23.02	54.34	49.93	35.11	48.27	66.53	39.62
Claude-3.5	29.53	52.51	22.67	56.88	43.84	32.87	58.85	55.05	39.99
Claude-4	27.13	65.28	23.33	53.89	45.79	34.00	56.64	61.79	42.45
GLM-4-Plus	31.81	56.47	24.89	69.89	44.74	37.56	55.23	55.58	39.01
Qwen-Max	42.89	53.12	**30.25**	64.19	43.86	36.25	58.88	58.00	42.74
Grok-4	34.54	58.33	27.19	48.47	**61.02**	37.05	53.51	**73.31**	**45.17**
Open-source Models									
DeepSeek-r1	32.21	55.47	25.06	52.43	53.97	36.33	55.53	65.75	43.18
R1-Distill-Qwen-7B	28.07	38.98	18.47	62.71	38.31	31.46	46.38	43.49	29.00
R1-Distill-Qwen-14B	35.20	52.73	26.00	66.72	44.38	36.37	53.70	56.19	37.81
Qwen-QWQ-32B	33.41	59.08	26.57	52.36	51.66	35.26	55.23	66.15	42.94
Qwen2.5-7B	26.95	37.38	17.64	61.96	37.82	30.59	47.28	43.73	29.47
Qwen2.5-14B	34.86	52.11	25.64	66.13	42.98	35.22	55.38	58.06	39.33
Llama-3.1-8B-Instruct	27.42	49.90	20.78	69.85	40.48	34.58	52.93	42.50	31.11
GLM-4-9B-Chat	29.14	50.42	22.02	65.25	42.73	34.71	55.00	52.05	37.63
HuatuoGPT-o1-8B	**43.03**	40.76	25.21	**74.17**	41.78	36.92	**61.79**	49.00	38.02
BioMistral-7B	29.98	22.84	15.29	60.93	31.63	29.78	37.09	15.47	12.86

Table 1: Performance comparison of different models on OncoEval benchmark across three clinical decision-making tasks.

Experiments

Evaluation Results

To comprehensively evaluate different existing models' performance on our proposed benchmark, we select a diverse set of large language models for testing. Table 1 presents the evaluation results of models across three clinical decision-making tasks, including: *(1) Latest Proprietary models*: GPT-o3/4.1 (OpenAI 2025), Gemini series (Mallick and Kilpatrick 2025; Comanici et al. 2025), Claude series (Anthropic 2024, 2025), GLM-4-Plus (GLM et al. 2024), Qwen-Max (Yang et al. 2025), Grok-4 [1]; *(2) Representative Open-source models*: DeepSeek-r1 and its variants (Guo et al. 2025), Qwen series (Qwen-Team 2025, 2024), Llama-3.1-8B-Instruct (Grattafiori et al. 2024), GLM-4-9B-Chat (GLM et al. 2024); and *(3) Medical domain-specialized models*: HuatuoGPT-o1 (Chen et al. 2024), BioMistral-7B (Labrak et al. 2024). These models represent different architectures and training paradigms, providing a broad perspective on current capabilities in medical tasks.

The evaluation results demonstrate that our proposed OncoEval benchmark presents significant challenges for current state-of-the-art models. Even the best-performing models achieve accuracy scores below 46% across all tasks. This consistently low performance indicates the substantial limitations in current LLMs for complex cancer medical decision-making, highlighting the distinctive difficulty and value of our proposed benchmark.

[1]https://x.ai/news/grok-4

The results show that closed-source models generally outperform their open-source counterparts, demonstrating superior capabilities across most evaluation metrics. A notable exception emerges with HuatuoGPT-o1, a Chinese medical reasoning model that achieves the highest precision scores across three tasks. This performance is particularly significant because HuatuoGPT-o1 was post-trained on medical domain reasoning data with only 7B parameters, providing evidence for the importance of high-quality CoT reasoning dataset in enhancing clinical decision-making capabilities.

Analysis of the three clinical decision-making tasks reveals distinct patterns. In examination tasks, models exhibit high recall but low precision, indicating a tendency to over-recommend examinations rather than making selective, clinically appropriate choices. Diagnosis tasks present more balanced precision-recall distributions. Treatment tasks represent the highest complexity, requiring models to simultaneously integrate examination findings, diagnostic insights, and therapeutic knowledge, resulting in relatively unsatisfactory performance across all evaluated models.

These results underline the challenges in medical decision-making tasks, for the existing models' consistently suboptimal performance indicates that LLMs require fundamental improvements in medical reasoning capabilities under authentic clinical settings. It also confirms that our method provides a rigorous standard for evaluating model capabilities in complex medical scenarios. These findings provide valuable insights for developing more sophisticated medical AI systems that can better support real-world clinical cancer decision-making tasks.

Data Quality Analysis

To validate the effectiveness of our proposed OncoCoT dataset, we conduct a comparative analysis using R1-Distill-Qwen-7B as the base model. This experiment aims to demonstrate the performance gains achieved through training with our OncoCoT dataset compared to alternative data construction paradigms. We evaluate on R1-Distill-Qwen-7B model using four configurations: (1) *Baseline*: the original pre-trained model without fine-tuning; (2) *w/o CoT*: fine-tuning on Q&A pairs without CoT reasoning; (3) *R1-distill CoT*: fine-tuning on CoT data constructed using vanilla rejection sampling (Guo et al. 2025); and (4) *OncoCoT*: fine-tuned on our proposed OncoCoT dataset. All experiments maintain identical hyperparameters and training iterations to ensure fair comparison, with models evaluated on the same benchmark described in the previous section.

Configuration	Examination	Diagnosis	Treatment
Baseline	28.51	44.16	39.62
w/o CoT	30.12	41.78	41.33
R1-distill CoT	33.16	43.92	46.40
OncoCoT	**34.95**	**44.55**	**50.13**

Table 2: Performance comparison of different configurations across three clinical tasks. Results present average scores combining precision, recall, and accuracy metrics.

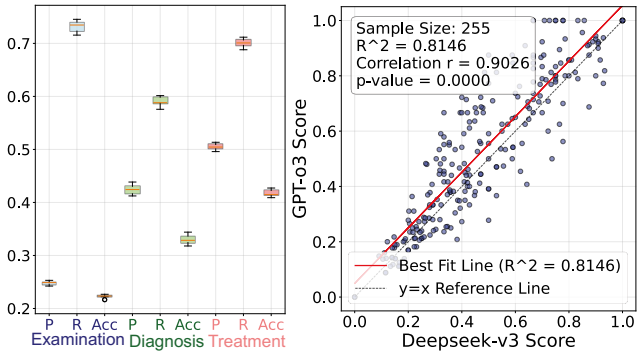
Table 2 presents the comparative performance. It can be seen that, under identical training data source, the model fine-tuned with OncoCoT data construction method consistently achieves the highest performance across all three tasks. This result demonstrates that our approach most effectively enhances model performance compared to existing data construction paradigms. The superior performance stems from our framework's incorporation of spatiotemporal clinical logic, which establishes stronger causal dependencies between medical events and constrains reasoning paths, thereby generating higher-quality CoT data.

Validation of Evaluation Method

In our work, we employ an LLM-based method to assess model capabilities. Although we have designed the evaluation to function as a task of information extraction and comparison to minimize the reliance on the evaluator model's intrinsic medical knowledge, the introduction of an external model inevitably introduces stochasticity into the evaluation process and challenges strict reproducibility. In this section, we present experiments to validate the reliability and consistency of our evaluation method.

First, to quantify the consistency of our evaluation model, DeepSeek-v3, we conduct a test-retest reliability analysis. Fig.5-(a) shows the result distributions from 10 repeated experiments in a box plot. The plot reveals minimal variance (*Std <0.01*) across all metrics, with tightly clustered distributions and no significant outliers. This indicates that our evaluation method is robust and demonstrates consistency.

Next, we assess the correlation of scores produced by different LLM evaluators to examine whether the evaluation results depend on a specific model. We compare the scores from DeepSeek-v3, Gemini-2.0-flash, and GPT-o3. We observe strong and statistically significant positive correlations among all evaluators (*Gemini vs. GPT-o3: r=0.7019,p<0.0001; Gemini vs. DeepSeek-v3: r=0.7495,p<0.0001; GPT-o3 vs. DeepSeek-V3: r=0.8078,p<0.0001*). Fig.5-(b) presents a scatter plot of the Accuracy scores assigned by GPT-o3 and DeepSeek-v3.



(a) Box plot of replicate evaluation results

(b) Scatter correlation between GPT-o3 and Deepseek-v3

Figure 5: Evaluation method validation: (a) Test-retest reliability analysis; (b) Inter-evaluator correlation analysis.

Given that all evaluator models perform their evaluation by comparing the target model's output against the human-curated ground truth, this high degree of inter-evaluator agreement demonstrates the robustness of our framework. The evaluation results are not contingent on a specific choice of evaluator but reflect a consistent measurement of the ground truth standard.

Conclusion

We present a temporal-causal data construction framework that constructs long CoT data aligned with the real-world comprehensive reasoning and decision patterns, and present the first-of-its-kind dataset OncoCoT using our proposed framework. The framework integrates timeline and decision point identification, a clinical causality-aware QA simulator, and clinical Logic-Constrained CoT rethinking. From our proposed dataset, we also distill OncoEval benchmark, the first medical benchmark for evaluating temporal-causal reasoning capabilities. As confirmed by correlation analysis between different evaluator LLMs, the LLM-based evaluation method proposed in OncoEval provides reliable and consistent results. Extensive experiments on OncoEval reveal the limitations of current LLMs supportive of medical reasoning tasks in handling complex real-world medical decision-making scenarios. Our experiments also verify the effectiveness of OncoCoT in training deep reasoning models, as models trained on OncoCoT achieve the highest improvement compared to other methods. Our finding points a new direction of endeavor to refine models that have support in medical question-answering tasks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 82090053; Tsinghua University Initiative Scientific Research Program of Precision Medicine under Grant number 2022ZLA007.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4.

Ben Abacha, A.; and Demner-Fushman, D. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1): 1–23.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Chen, Y.; Wang, Z.; Xing, X.; Zheng, H.; Xu, Z.; Fang, K.; Wang, J.; Li, S.; Wu, J.; Liu, Q.; et al. 2023. BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT. *CoRR*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Devi, S.; Dhar, G.; Bharadwaj, C.; and M, A. 2024. Retrieval Augmented MedLM. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, 1220–1221.

Ding, C.; Bian, M.; Chen, P.; Zhang, H.; Li, T.; Liu, L.; Chen, J.; Li, Z.; Zhong, Y.; Liu, Y.; et al. 2025. Building a Human-Verified Clinical Reasoning Dataset via a Human LLM Hybrid Pipeline for Trustworthy Medical AI. *arXiv preprint arXiv:2505.06912*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hooftman, J.; Dijkstra, A. C.; Suurmeijer, I.; van der Bij, A.; Paap, E.; and Zwaan, L. 2024. Common contributing factors of diagnostic error: A retrospective analysis of 109 serious adverse event reports from Dutch hospitals. *BMJ Quality & Safety*, 33(10): 642–651.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jackson, R.; and Feder, G. 1998. Guidelines for clinical guidelines: a simple, pragmatic strategy for guideline development.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-a.; Rouvier, M.; and Dufour, R. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *62th Annual Meeting of the Association for Computational Linguistics (ACL'24)*.

Li, J.; Zhong, S.; and Chen, K. 2021. MLEC-QA: A Chinese multi-choice biomedical question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8862–8874.

Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; and Zhang, Y. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*, 15(6): e40895.

Liu, J.; Wang, Y.; Du, J.; Zhou, J.; and Liu, Z. 2024. MedCoT: Medical Chain of Thought via Hierarchical Expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17371–17389.

Liu, J.; Zhou, P.; Hua, Y.; Chong, D.; Tian, Z.; Liu, A.; Wang, H.; You, C.; Guo, Z.; Zhu, L.; et al. 2023. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36: 52430–52452.

Luo, M.; Saxena, S.; Mishra, S.; Parmar, M.; and Baral, C. 2022. Biotabqa: Instruction learning for biomedical table question answering. *arXiv preprint arXiv:2207.02419*.

Mallick, S. B.; and Kilpatrick, L. 2025. Gemini 2.0Flash, Flash-Lite and Pro. Accessed: 2025-05-01.

Nachane, S.; Gramopadhye, O.; Chanda, P.; Ramakrishnan, G.; Jadhav, K.; Nandwani, Y.; Raghu, D.; and Joshi, S. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 542–573.

OpenAI. 2025. OpenAI o3 and o4-mini System Card.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Flores, G.; Chen, G. H.; Pollard, T.; Ho, J. C.; and Naumann, T., eds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, 248–260. PMLR.

Qwen-Team. 2024. Qwen2.5: A Party of Foundation Models.

Qwen-Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Sandmann, S.; Hegselmann, S.; Fujarski, M.; Bickmann, L.; Wild, B.; Eils, R.; and Varghese, J. 2025. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nature Medicine*, 1–1.

Shivade, C. 2017. MedNLI — A Natural Language Inference Dataset For The Clinical Domain.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yu, X.; Hu, W.; Lu, S.; Sun, X.; and Yuan, Z. 2019. BioBERT based named entity recognition in electronic medical record. In *2019 10th international conference on information technology in medicine and education (ITME)*, 49–52. IEEE.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Chen, G.; Li, J.; Wu, X.; Zhiyi, Z.; Xiao, Q.; et al. 2023a. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10859–10885.

Zhang, X.; Tian, C.; Yang, X.; Chen, L.; Li, Z.; and Petzold, L. R. 2023b. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.