EHRStruct: A Comprehensive Benchmark Framework for Evaluating Large Language Models on Structured Electronic Health Record Tasks

Xiao Yang[1], **Xuejiao Zhao**[2,3*], **Zhiqi Shen**[1]

[1] College of Computing and Data Science, Nanyang Technological University (NTU), Singapore
[2] Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), NTU, Singapore
[3] Alibaba-NTU Singapore Joint Research Institute (ANGEL), NTU, Singapore
yangxiao.cs@gmail.com, {xjzhao,zqshen}@ntu.edu.sg

Abstract

Structured Electronic Health Record (EHR) data stores patient information in relational tables and plays a central role in clinical decision-making. Recent advances have explored the use of large language models (LLMs) to process such data, showing promise across various clinical tasks. However, the absence of standardized evaluation frameworks and clearly defined tasks makes it difficult to systematically assess and compare LLM performance on structured EHR data. To address these evaluation challenges, we introduce EHRStruct, a benchmark specifically designed to evaluate LLMs on structured EHR tasks. EHRStruct defines 11 representative tasks spanning diverse clinical needs and includes 2,200 task-specific evaluation samples derived from two widely used EHR datasets. We use EHRStruct to evaluate 20 advanced and representative LLMs, covering both general and medical models. We further analyze key factors influencing model performance, including input formats, few-shot generalisation, and finetuning strategies, and compare results with 11 state-of-the-art LLM-based enhancement methods for structured data reasoning. Our results indicate that many structured EHR tasks place high demands on the understanding and reasoning capabilities of LLMs. In response, we propose EHRMaster, a code-augmented method that achieves state-of-the-art performance and offers practical insights to guide future research.

Code — https://github.com/YXNTU/EHRStruct
Project page — https://yxntu.github.io/proEHRStruct/
Extended version — https://arxiv.org/abs/2511.08206

1 Introduction

Structured electronic health records (EHRs) (Häyrinen, Saranto, and Nykänen 2008; Wei, Zhao, and Miao 2018) store patient information in relational tables, including diagnoses, medications, and laboratory results. Each entry in records corresponds to a specific clinical event or measurement and is often timestamped to capture longitudinal patient trajectories. In contrast to traditional methods such as SQL-based querying, large language models (LLMs) (Brown et al. 2020; Zhao et al. 2025b) offer greater flexibility, more powerful reasoning abilities, and a natural language interface, and have been increasingly adopted for structured EHR modeling (Li et al. 2024a).

The application of LLMs to structured EHR data has recently attracted growing attention, emerging as a promising frontier for clinical AI research. However, effectively leveraging such data remains challenging even for state-of-the-art LLMs, due to the need for tabular understanding, clinical reasoning, and alignment with user intent (Ren et al. 2025). Recent efforts have attempted to tackle these aspects from different angles: Yang et al. (Yang et al. 2024) tackle tabular understanding by automating the generation of standardized summary tables from structured clinical trial data. Zhu et al. (Zhu et al. 2024) focus on clinical reasoning tasks, developing a prompting-based approach for predicting outcomes such as mortality, length of stay, and readmission from longitudinal EHR records. Kwon et al. (Kwon et al. 2024) address alignment with user intent by verifying the semantic consistency between structured EHR tables and unstructured clinical notes.

Despite promising advances, existing work lacks a unified evaluation framework for assessing LLMs on structured EHR tasks (Lovon et al. 2025). This issue manifests in several ways. First, most studies focus on a limited set of tasks—such as disease prediction (Contreras et al. 2024; Hu et al. 2024b,a; Zhao et al. 2025a), mortality risk estimation (Wang et al. 2025), information extraction (Huang et al. 2024), and arithmetic reasoning over tabular data (Yang et al. 2024)—while leaving many clinically important use cases underexplored, including medication recommendation (Shool et al. 2025) and clinical named entity recognition (Monajatipoor et al. 2024; Li et al. 2024c). Second, even when addressing the same task (e.g., disease prediction), prior studies (Zhu et al. 2024; Contreras et al. 2024; Hu et al. 2024b,a) often use different datasets and evaluation protocols, limiting reproducibility and hindering fair model comparison. Third, there is no consensus on input formatting strategies or experimental setups, leading to inconsistencies in how structured data are presented to LLMs. Fourth, existing evaluation metrics provide limited interpretability, making it difficult to understand which specific reasoning capabilities contribute to model successes or failures.

To address these limitations, we introduce EHRStruct, a comprehensive benchmark specifically designed to sys-

*Corresponding Author

Task Scenarios	Task Levels	Task Categories	Task IDs	Metrics
Data-Driven	Understanding	Information retrieval	D-U1/U2	Accuracy
	Reasoning	Data aggregation	D-R1/R2/R3	Accuracy
		Arithmetic computation	D-R4/R5	Accuracy
Knowledge-Driven	Understanding	Clinical identification	K-U1	AUC[1]
	Reasoning	Diagnostic assessment	K-R1/R2	AUC
		Treatment planning	K-R3	AUC

Table 1: Overview of the 11 structured tasks in EHRStruct, categorized by scenario (Data-Driven vs. Knowledge-Driven) and cognitive level (Understanding vs. Reasoning).

tematically evaluate LLMs on structured EHR tasks. First, to expand task coverage, EHRStruct defines 11 diverse tasks across 6 categories, including information retrieval, data aggregation, and more. These categories are distilled from a thorough analysis of real-world clinical applications and prior research paradigms, ensuring that the selected tasks reflect both operational diversity and clinical relevance. Second, to mitigate dataset inconsistencies, we construct task-specific evaluation samples using standardized data sources from two complementary origins—the synthetic Synthea dataset (Walonoski et al. 2018) and the real-world eICU database (Pollard et al. 2018). For each task, we select representative and non-overlapping input–output pairs, validated through cross-checking by multiple domain experts, enabling reproducible and fair comparisons across models. Third, to resolve inconsistencies in input formatting and experimental protocols, we conduct a systematic exploration of input construction strategies and propose a unified framework that clearly distinguishes between different prompt structures for controlled experimentation. Finally, to enhance interpretability and support fine-grained diagnostic analysis, we further classify all tasks along two orthogonal dimensions: evaluation scenario—Data-Driven versus Knowledge-Driven, and cognitive complexity—understanding versus reasoning. This two-dimensional design enables deeper insights into the specific capabilities and limitations of LLMs.

To assess the practicality and effectiveness of our proposed benchmark, EHRStruct, we evaluate 20 representative general and medical LLMs. We primarily focus on zero-shot settings, while also testing 1-shot, 3-shot, and 5-shot configurations to investigate few-shot performance across tasks in our benchmark. In addition, we systematically examine the impact of input formatting, comparing 4 common prompt structures for structured data, and evaluate the effectiveness of task-specific finetuning. In addition to evaluating base LLMs, we assess 11 representative LLM-based enhancement methods designed to improve structured data performance, including 8 originally developed for non-medical tasks and 3 specifically tailored for medical tasks. Building on insights from our evaluation of both base LLMs and enhancement methods, we propose EHRMaster—a novel code-augmented framework tailored for structured EHR tasks. EHRMaster operates in three stages: it first generates a high-level solution plan based on the task definition, then aligns key concepts in the plan with relevant data fields, and finally determines whether to generate executable code or proceed with direct reasoning. By structuring execution around task semantics, EHRMaster achieves substantial performance gains across diverse benchmark tasks. Our main contributions are as follows:

- We introduce EHRStruct, a comprehensive benchmark featuring diverse clinically grounded tasks, standardized datasets, systematic input design, and interpretable evaluation to assess LLMs on structured EHR tasks.
- We use EHRStruct to systematically evaluate 20 general and medical LLMs and 11 LLM-based enhancement methods, providing extensive analysis and insights into task-specific performance.
- We propose EHRMaster, achieving state-of-the-art results on our benchmark and providing valuable inspiration for future research on structured EHR modeling.

2 Key Findings

Table 1 summarizes the benchmark tasks across defined scenarios, levels, and categories, with detailed classification criteria provided in the Appendix Section A.1. Our main findings are summarized as follows:

- **General LLMs Outperform Medical LLMs:** General LLMs consistently outperform medical models on structured EHR tasks. Among these, closed-source commercial models—especially the Gemini series—achieve the best overall performance.
- **LLMs Excel at Data-Driven Tasks:** Overall, LLMs perform better on Data-Driven tasks than on Knowledge-Driven ones.
- **Input Format Influences Performance:** Natural language inputs benefit Data-Driven reasoning tasks, while graph-structured prompts help with Data-Driven understanding. No format consistently improves Knowledge-Driven tasks.
- **Few-shot Improves Performance:** Few-shot prompting improves performance overall, with 1-shot and 3-shot settings typically outperforming 5-shot.

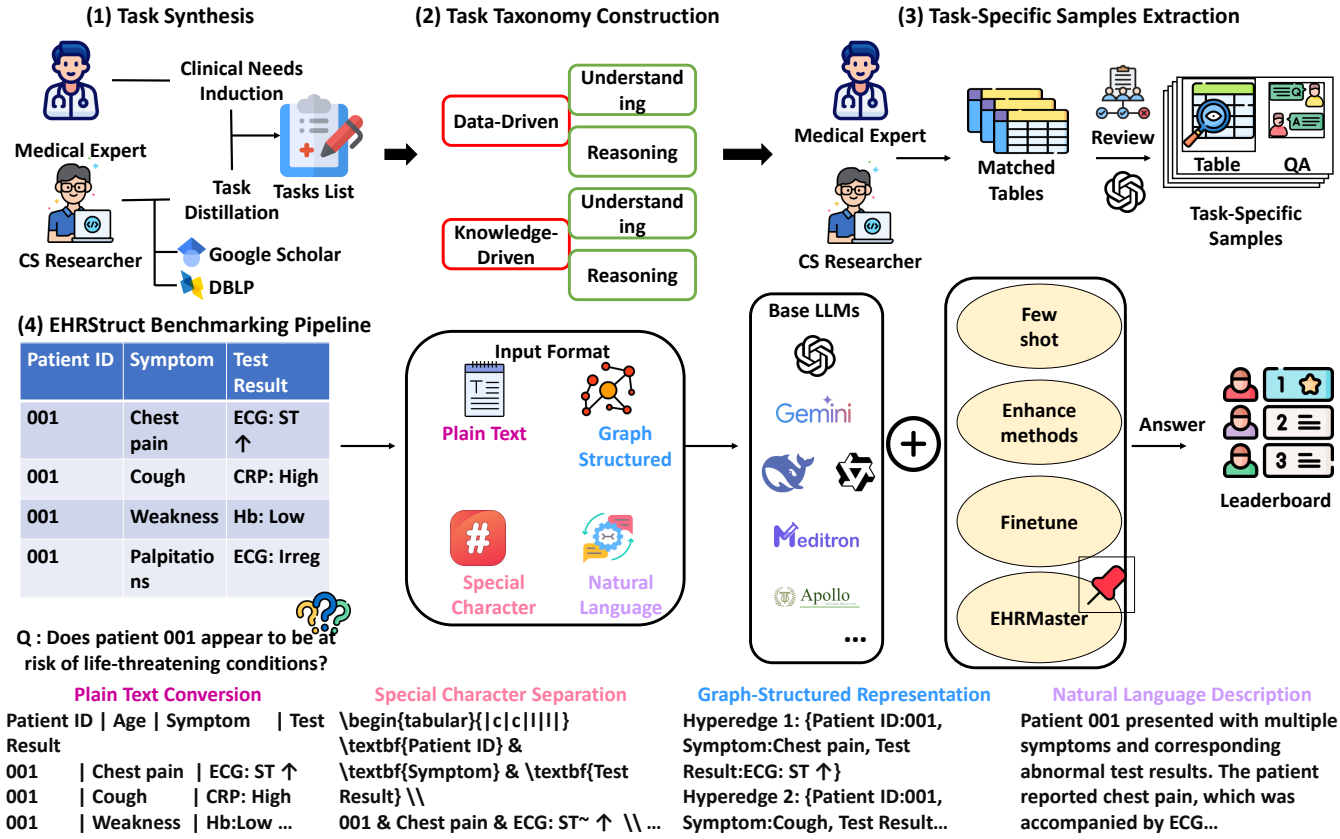


Figure 1: Overview of EHRStruct. The figure illustrates the four key components of the benchmark: (1) task synthesis through clinical needs induction and task distillation from prior research; (2) taxonomy construction based on clinical scenarios and reasoning levels; (3) task-specific sample extraction from real and synthetic EHR data; and (4) the model evaluation pipeline, including table input, format conversion, model inference, and answer evaluation.

- **Multi-task Fine-tuning Outperforms Single-task Fine-tuning:** While both strategies improve LLM performance, multi-task fine-tuning yields greater gains.
- **Enhancement Methods Are Scenario-Specific:** Non-medical enhancement methods underperform in Knowledge-Driven categories, while medical-specific methods struggle in Data-Driven scenarios.

3 EHRStruct

Figure 1 provides an overview of our benchmark framework, EHRStruct. In this section, we elaborate on its design, covering the data sources, task construction, evaluation setup, and our proposed EHRMaster.

3.1 Task Synthesis

Our benchmark defines six categories of tasks that reflect both emerging and established application scenarios in structured EHR modeling. Each task is initially proposed by the CS researcher through task distillation from prior work and existing modeling paradigms, then reviewed and validated by the medical expert to ensure clinical relevance. *Clinical identification* and *Treatment planning* are retained based on expert confirmation, as they are clinically important but remain underexplored in structured settings despite being common in unstructured EHR contexts. Their definitions are customized based on prior work in concept extraction (Ong et al. 2023; Chang and Sung 2024) and planning-based dialogue systems (Ullah et al. 2024; Tan et al. 2024; Zhao et al. 2021). The remaining four task types—*Information retrieval*, *Data aggregation*, *Arithmetic computation*, and *Diagnostic assessment*—are distilled from typical LLM applications to structured EHR data and reflect core reasoning patterns observed in prior studies. Together, these tasks are designed to cover a diverse range of real-world clinical needs in both operational and decision-support scenarios.

3.2 Task Taxonomy Construction

As summarized in Table 1, we organize the benchmark tasks along three axes: clinical scenario (Data-Driven vs. Knowledge-Driven), cognitive level (Understanding vs. Reasoning), and task category (six functional types). This taxonomy captures both the practical intent and the reasoning complexity of each task, supporting a comprehensive and interpretable evaluation framework (Fleishman 1975;

Types	Models	# Params
General Large Language Models	GPT-3.5 Turbo (OpenAI 2023)	Commercial
	GPT-4.1 (OpenAI 2025)	Commercial
	Gemini 1.5 (DeepMind 2024a)	Commercial
	Gemini 2.0 (DeepMind 2024b)	Commercial
	Gemini 2.5 (DeepMind 2025)	Commercial
	DeepSeek-V2.5 (Liu et al. 2024a)	236B
	DeepSeek-V3 (Liu et al. 2024b)	685B
	Qwen-7B (Team 2024)	7B
	Qwen-14B (Team 2024)	14B
	Qwen-32B (Team 2024)	32B
	Qwen-72B (Team 2024)	72B
Medical Large Language Models	Huatuo (Zhang et al. 2023)	7B
	HEAL (Han et al. 2023)	7B
	Meditron-7B (Chen et al. 2023)	7B
	MedAlpaca-13B (Yuan et al. 2024)	13B
	JMLR (Wang et al. 2024a)	13B
	PMC_LLaMA_13B (Wu et al. 2024)	13B
	Med42-70B (Christophe et al. 2024)	70B
	Apollo (Wang et al. 2024b)	70B
	CancerLLM (Li et al. 2024b)	70B

Table 2: List of 20 LLMs, including 11 general and 9 medical models, covering both open-source and commercial releases with parameter sizes ranging from 7B to 685B.

Zhao et al. 2017). Detailed descriptions are provided in Appendix Section A.2, and task instructions appear in Appendix Section A.3. Such a structured organization also enables fine-grained comparison of model performance across clinically and cognitively diverse settings.

3.3 Task-Specific Samples Extraction

Our benchmark is built from two representative structured EHR sources. The first is **Synthea** (Walonoski et al. 2018)[1], a synthetic dataset simulating realistic patient records without privacy concerns. The second is the **eICU Collaborative Research Database** (Pollard et al. 2018)[2], a real-world ICU dataset comprising clinically rich, multi-institutional structured tables. Together, they ensure coverage of both simulated and authentic clinical scenarios.

For each task, the CS researcher and the medical expert jointly screen and identify the most relevant tables based on the task definition and schema content. Once matched, we construct 100 evaluation samples per task per dataset, resulting in 2,200 annotated instances across 11 tasks. Representative data rows are selected to ensure clinical diversity, and GPT-4o is used to generate question–answer pairs conditioned on the task definition, table schema, and sampled content. All outputs undergo two-stage validation: medical reviewers assess the correctness and plausibility of answers, while technical reviewers verify that each question is faithful to the task objectives and input semantics.

[1]https://github.com/synthetichealth/synthea

[2]https://eicu-crd.mit.edu/

3.4 Evaluation Setup

We evaluate a diverse set of large language models (LLMs) on our benchmark, including both general-purpose and medical-domain models. Table 2 summarizes the 20 models included in our study, detailing their model types and parameter sizes. For each task ID, we evaluate all 20 LLMs using 200 question–answer pairs (100 from the synthetic Synthea dataset and 100 from the real-world eICU dataset), ensuring balanced coverage across data sources. We employ 4 distinct formats to transform structured EHR data into text inputs and report results separately for each data source. All evaluations use single-turn generation with consistent decoding parameters (e.g., temperature and maximum token limits) to ensure fair model comparisons.

Beyond these benchmark-wide evaluations, we conduct detailed few-shot and fine-tuning experiments on the Gemini series to investigate their potential on structured EHR tasks. We also reproduce 11 existing methods for structured data reasoning—8 from other domains and 3 from clinical settings—and systematically evaluate their performance on our benchmark for comparison. Finally, we evaluate our proposed EHRMaster method on the benchmark to demonstrate its effectiveness relative to both general-purpose LLMs and existing structured-data reasoning approaches.

3.5 EHRMaster Overview

Here, we briefly outline our proposed EHRMaster framework (see Appendix Section B for details). EHRMaster operates in three key stages:

Solution Planning: Generates a high-level natural-language solution plan based on the question, decomposing it into the required reasoning steps. **Concept Alignment:** Maps the abstract concepts in the plan to the corresponding fields and tables in the structured EHR data. **Adaptive Execution:** Selects between code-based execution and direct language reasoning, and applies the chosen strategy to retrieve evidence and derive the final answer.

For example, in the hospital cost estimation task, once the relevant fields have been identified, EHRMaster generates Python code to filter billing entries by admission and discharge timestamps and compute the total cost. In contrast, for tasks like assessing treatment effectiveness based on heterogeneous clinical events, EHRMaster may bypass code and instead use multi-step language reasoning grounded in the aligned data.

4 Results

In this section, we provide an overview of the results, analyses, and experimental findings for the LLMs evaluated in our benchmark. Detailed results and further analyses are presented in Appendix Section C.

4.1 Overall Benchmark Results

Table 3 reports the performance of all evaluated models on the synthetic Synthea dataset. We organize results by task scenario—Data-Driven and Knowledge-Driven—and

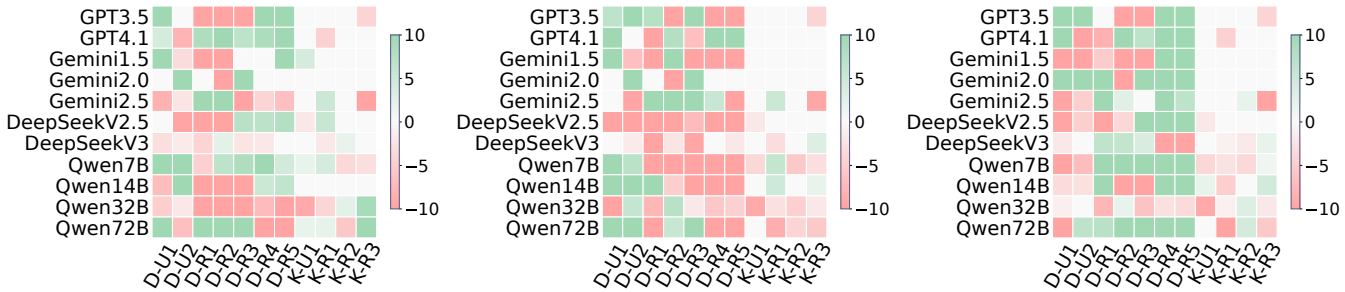


Figure 2: Relative Performance Gains from Different input formats across LLMs.

further distinguish tasks by level, specifically Understanding (U) and Reasoning (R). Although the specific question–answer pairs differ across datasets, we observe highly consistent performance patterns on the real-world eICU dataset as well, which is reported in Appendix Table 6.

First, general LLMs consistently outperform medical LLMs across nearly all task categories. This performance gap is particularly evident in the Knowledge-Driven scenario, where medical models frequently fail to produce valid outputs or achieve meaningful accuracy or AUC scores. Notably, none of the medical models rank among the top three performers for any task ID. In contrast, closed-source commercial models—especially the Gemini series—achieve the highest overall performance, demonstrating robust generalization across synthetic structured EHR tasks. These results suggest that general models benefit from broad pretraining on diverse text sources, which may indirectly support structured data understanding even without explicit domain-specific adaptation.

Second, model performance shows clear variation across task scenarios and levels, reflecting the different demands posed by each task category. In the Data-Driven scenario, strong general-purpose LLMs perform well on both understanding and reasoning tasks, suggesting that, when structured inputs are properly formatted, general models can interpret and reason over structured data with reasonable accuracy. In contrast, Knowledge-Driven tasks present substantially greater challenges. For the understanding task of Clinical Code Mapping (K-U1), general models achieve only moderate AUC scores (typically between 50–60%), while most medical-domain models fail to produce valid outputs. Performance drops even further on reasoning tasks such as Diagnostic Assessment (K-R1, K-R2) and Treatment Planning (K-R3), where many models struggle to generate meaningful predictions. This gap highlights the difficulty of integrating external clinical knowledge into structured data interpretation and underscores the need for models that better handle medical semantics and complex reasoning.

4.2 Few-shot Analysis

Figure 3 presents few-shot performance for a subset of Knowledge-Driven tasks, with full results provided in Appendix Figure 6. Overall, few-shot prompting is beneficial,

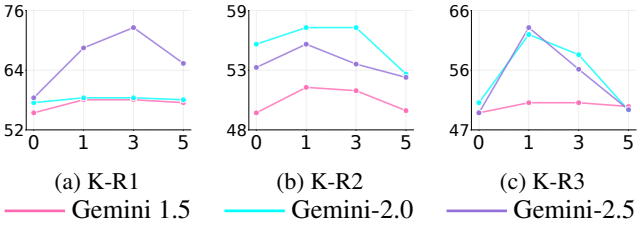


Figure 3: Performance of representative LLMs on two Scenarios under few-shot (1, 3, and 5-shot) learning settings.

with 1-shot and 3-shot settings typically outperforming 5-shot. Compared to Gemini 1.5, Gemini 2.0 and 2.5 exhibit more pronounced performance gains under few-shot settings, suggesting a greater sensitivity to in-context examples and a stronger generalization from limited demonstrations.

4.3 Effect of Input Formats

We examine 4 input formats for converting structured electronic health record data into text: plain text conversion, special character separation, graph-structured representation, and natural language description. These formats are used to test how different input styles affect model performance on our benchmark.

As shown in Figure 2, input format has a clear impact on performance. Natural language description improves results on Data-Driven reasoning tasks, especially for strong models such as the Gemini and GPT series. Graph-structured input is more effective for Data-Driven understanding tasks. Across all Knowledge-Driven tasks, however, no format yields consistent improvement. This shows that input format helps in certain settings, but deeper modeling efforts are needed for tasks requiring clinical knowledge.

4.4 Finetuning Analysis

To analyze the fine-tuning performance of large language models on structured EHR tasks, we construct additional training samples specifically for fine-tuning, which are completely separate from the evaluation samples used in benchmarking. Each fine-tuning dataset contains 30 task-specific question–answer table pairs following the same instruction format as in the evaluation, ensuring consistency of task

Types	Models	Data-Driven							Knowledge-Driven			
		U (%)		R (%)					U (%)	R (%)		
		D-U1	D-U2	D-R1	D-R2	D-R3	D-R4	D-R5	K-U1	K-R1	K-R2	K-R3
		ACC	ACC	ACC	ACC	ACC	ACC	ACC	AUC	AUC	AUC	AUC
General LLMs	GPT-3.5 Turbo	6	15	14	18	7	7	24	✗	58.1	55.4	52.9
	GPT-4.1	79	51	52	56	48	70	84	55	55.6	53.2	51
	Gemini 1.5	29	34	32	41	21	19	16	✗	55.6	✗	✗
	Gemini-2.0	64	43	21	30	24	54	67	52	57.7	56.2	51.6
	Gemini 2.5	98	58	92	82	83	✓	✓	✗	58.7	54.1	✗
	DeepSeek-V2.5	72	41	18	51	14	44	52	51	✗	✗	✗
	DeepSeek-V3	72	41	8	37	12	72	90	✗	52.8	✗	✗
	Qwen-7B	1	7	4	24	1	✗	✗	✗	✗	✗	✗
	Qwen-14B	4	30	19	17	11	16	4	✗	✗	✗	✗
	Qwen-32B	25	25	24	26	15	47	10	✗	58.3	51	✗
	Qwen-72B	15	6	27	48	20	41	29	✗	✗	✗	52.2
Medical LLMs	Huatuo	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	HEAL	✗	✗	1	8	✗	✗	✗	✗	✗	✗	✗
	Meditron-7B	✗	3	✗	6	✗	✗	✗	✗	✗	✗	✗
	MedAlpaca-13B	2	11	6	4	2	10	✗	✗	✗	✗	✗
	JMLR	1	3	11	10	6	7	3	✗	✗	✗	✗
	PMC_LLaMA_13B	6	6	15	13	10	8	✗	✗	✗	✗	✗
	Med42-70B	13	3	18	17	11	27	18	✗	✗	✗	✗
	Apollo	11	5	17	12	6	20	11	✗	✗	✗	✗
	CancerLLM	10	16	20	28	15	33	25	✗	✗	✗	✗

Table 3: Performance of LLMs on Structured EHR Tasks under the zero-shot setting(Synthea). ✗ indicates no valid output. ✓ indicates a perfect score of 100. 1st, 2nd, and 3rd denote the best, second-best, and third-best results, respectively.

structure while preventing data leakage. We adopt two fine-tuning strategies: single-task fine-tuning trains each task separately, while multi-task fine-tuning combines all tasks into one set. All experiments use Qwen-7B with LoRA fine-tuning, configured with a 10% validation split, a 0.0001 learning rate, three epochs, a batch size of 8, and LoRA settings of rank 8, alpha = 32, and dropout 0.05.

Figure 4 shows that fine-tuning improves model performance substantially across both Data-Driven and Knowledge-Driven tasks. Across all tasks, multi-task fine-tuning consistently outperforms single-task fine-tuning, likely because joint training helps the model learn shared structures and reasoning patterns relevant to structured EHR.

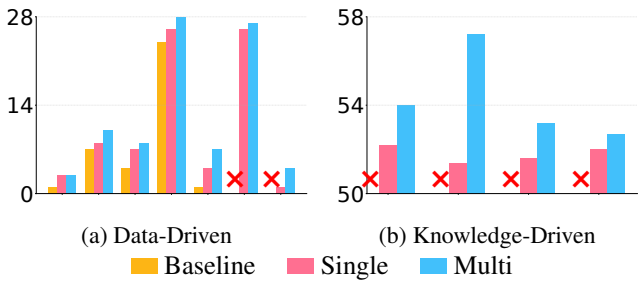
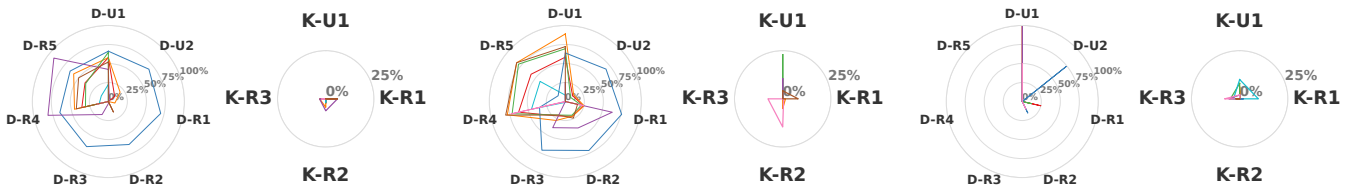


Figure 4: Finetuning results on all **targeted categories**. Single-task indicates separate finetuning on each task; multi-task indicates joint finetuning across all tasks.

4.5 Evaluating LLM-Based Enhancement Methods

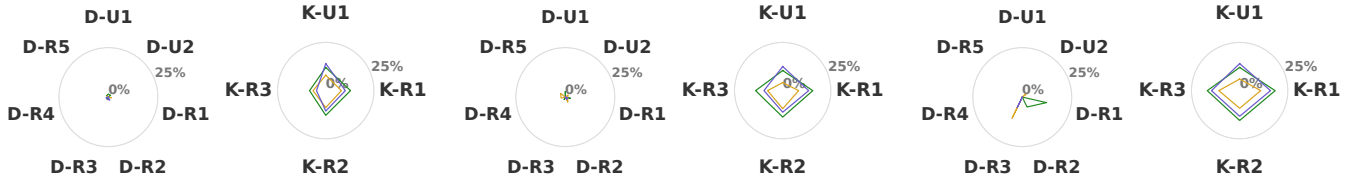
We reproduce and evaluate 11 representative LLM-based enhancement methods designed to enhance performance on structured data tasks. Among them, eight methods were originally developed for non-medical tasks—C.L.E.A.R. (Deng et al. 2024), TaT (Sun et al. 2025), TableMaster (Cao and Liu 2025), TIDE (Yang et al. 2025b), E^5 (Zhang, Gao, and Lou 2024), GraphOTTER (Li et al. 2024d), H-STAR (Abhyankar et al. 2024), and Table-R1 (Yang et al. 2025a)—while the remaining three are specifically designed for medical applications: LLM4Healthcare (Zhu et al. 2024), DeLLiriuM (Contreras et al. 2024), and EnsembleLLM (Hu et al. 2024b).

As shown in Figure 5, these methods exhibit clear specialization based on task scenarios. Non-medical methods generally achieve greater gains on Data-Driven tasks—such as field filtering or numeric reasoning—but show limited improvements on Knowledge-Driven tasks that require medical expertise. In contrast, medical methods perform better on Knowledge-Driven tasks—such as mortality prediction or treatment planning—yet struggle to generalize to Data-Driven scenarios. This divergence highlights a key limitation of existing enhancement approaches: none offer consistent improvements across the full spectrum of structured EHR tasks. These results underscore the need for unified solutions capable of both logical reasoning over structured tables and integration of medical domain knowledge.



(a) Gemini-1.5 (b) Gemini-2.0 (c) Gemini-2.5

Non-Medical : C.L.E.A.R, TaT, TableMaster, TIDE, E5, GraphOTTER, H-STAR, Table-R1,



(d) Gemini-1.5 (e) Gemini-2.0 (f) Gemini-2.5

Medical : llm4healthcare, DeLLiriuM, EnsembleLLM.

Figure 5: Comparison of relative gains for 11 SOTA methods across tasks. Relative gain is defined as the percentage of improvement each method achieves toward the maximum possible gain for each task, where 0% indicates no improvement and 100% represents the upper bound. In each subfigure, the left side shows Data-Driven tasks, and the right side shows Knowledge-Driven tasks.

Models	Methods	Data-Driven							Knowledge-Driven			
		D-U1	D-U2	D-R1	D-R2	D-R3	D-R4	D-R5	K-U1	K-R1	K-R2	K-R3
		ACC	ACC	ACC	ACC	ACC	ACC	ACC	AUC	AUC	AUC	AUC
Gemini 1.5	EHRMaster	**100**	**100**	**96**	**96**	**94**	**100**	**100**	**89**	**62.3**	54	**54.7**
	previous SOTA	76	79	80	78	73	85	93	57	61.3	**56.4**	54.2
Gemini 2.0	EHRMaster	**98**	**100**	**91**	**81**	**93**	80	87	**67**	**65.3**	**64.2**	56.2
	previous SOTA	96	82	81	80	78	**90**	**94**	63	64.3	62.2	**58.4**
Gemini 2.5	EHRMaster	**100**	**100**	**97**	**95**	**97**	**100**	**100**	**60**	59.3	55.1	**69.2**
	previous SOTA	**100**	89	94	85	85	100	100	57	**66.3**	**61.2**	58.4

Table 4: Performance of EHRMaster compared to LLM-based enhancement methods on benchmark.

4.6 EHRMaster Analysis

Table 4 reports the performance of EHRMaster compared to the state-of-the-art (SOTA) LLM-based enhancement baselines across all benchmark tasks. On Data-Driven tasks, EHRMaster consistently achieves perfect scores across Gemini models, demonstrating strong capabilities in structured reasoning and arithmetic operations. Significant gains appear on arithmetic-heavy tasks such as D-R4 and D-R5, where EHRMaster achieves 100% accuracy in most cases. For Knowledge-Driven tasks, EHRMaster delivers noticeable improvements on challenging tasks (e.g., K-R2 and K-R3), although gains vary across models and tasks. These results highlight EHRMaster's effectiveness in structured EHR reasoning, especially in Data-Driven tasks, while also offering competitive performance on complex clinical tasks.

5 Conclusion

In summary, we present EHRStruct, a benchmark for evaluating LLMs on structured EHR tasks, enabling systematic comparison across diverse scenarios. EHRStruct addresses the key challenges in applying LLMs to structured EHR data by providing a novel dataset, defining clear task specifications, and establishing a standardized evaluation framework. First, we construct a novel dataset containing 2,200 structured EHR tables, generated synthetically using Synthea and extracted from the real-world eICU database. This dataset spans a wide range of clinical scenarios to ensure robust evaluation. Second, we define 11 tasks across 6 categories, organized into Data-Driven and Knowledge-Driven scenarios, covering a broad spectrum of structured EHR applications, from data extraction to clinical decision support. Finally, we establish a standardized evaluation process, testing LLMs under various settings such as zero-shot, few-shot, and fine-tuning, and analyzing the impact of different input formats. This framework enables consistent and comprehensive assessment of LLM performance on structured EHR tasks. We also reproduce 11 SOTA LLM-based enhancement methods for structured data processing and evaluate them using our benchmark. Through this evaluation, we identified the limitations of the existing methods and EHRMaster, a novel code-augmented framework, to address these challenges and achieve superior performance.

Acknowledgments

This research is supported in part by the Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly (LILY), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Jinan-NTU Green Technology Research Institute (GreenTRI). This research is also partially supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through the Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

References

Abhyankar, N.; Gupta, V.; Roth, D.; and Reddy, C. K. 2024. H-STAR: LLM-driven Hybrid SQL-Text Adaptive Reasoning on Tables. *arXiv preprint arXiv:2407.05952*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, L.; and Liu, H. 2025. Tablemaster: A recipe to advance table understanding with language models. *arXiv preprint arXiv:2501.19378*.

Chang, E.; and Sung, S. 2024. Use of SNOMED CT in Large Language Models: Scoping Review. *JMIR Medical Informatics*, 12(1): e62924.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Christophe, C.; Kanithi, P. K.; Raha, T.; Khan, S.; and Pimentel, M. A. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Contreras, M.; Kapoor, S.; Zhang, J.; Davidson, A.; Ren, Y.; Guan, Z.; Ozrazgat-Baslanti, T.; Nerella, S.; Bihorac, A.; and Rashidi, P. 2024. DeLLiriuM: A large language model for delirium prediction in the ICU using structured EHR. *arXiv preprint arXiv:2410.17363*.

DeepMind. 2024a. Gemini 1.5: Unlocking multimodal understanding across trillions of tokens. *ArXiv*. Abs/2403.05530.

DeepMind. 2024b. Gemini 2.0 Flash [Large language model]. https://ai.google.dev/gemini-api.

DeepMind. 2025. Gemini 2.5 Flash [Large language model]. https://ai.google.dev/gemini-api.

Deng, I.; Dixit, K.; Gupta, V.; and Roth, D. 2024. Enhancing Temporal Understanding in LLMs for Semi-structured Tables. *arXiv preprint arXiv:2407.16030*.

Fleishman, E. A. 1975. Toward a taxonomy of human performance. *American Psychologist*, 30(12): 1127.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca–an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

Häyrinen, K.; Saranto, K.; and Nykänen, P. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5): 291–304.

Hu, D.; Liu, B.; Li, X.; Zhu, X.; and Wu, N. 2024a. Predicting Lung Cancer Patient Prognosis with Large Language Models. *arXiv preprint arXiv:2408.07971*.

Hu, D.; Liu, B.; Zhu, X.; and Wu, N. 2024b. The Power of Combining Data and Knowledge: GPT-4o is an Effective Interpreter of Machine Learning Models in Predicting Lymph Node Metastasis of Lung Cancer. *arXiv preprint arXiv:2407.17900*.

Huang, J.; Yang, D. M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L. J.; et al. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ digital medicine*, 7(1): 106.

Kwon, Y.; Kim, J.; Lee, G.; Bae, S.; Kyung, D.; Cha, W.; Pollard, T.; Johnson, A.; and Choi, E. 2024. EHRCon: Dataset for Checking Consistency between Unstructured Notes and Structured Tables in Electronic Health Records. *arXiv preprint arXiv:2406.16341*.

Li, L.; Zhou, J.; Gao, Z.; Hua, W.; Fan, L.; Yu, H.; Hagen, L.; Zhang, Y.; Assimes, T. L.; Hemphill, L.; et al. 2024a. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs). *arXiv preprint arXiv:2405.03066*.

Li, M.; Huang, J.; Yeung, J.; Blaes, A.; Johnson, S.; Liu, H.; Xu, H.; and Zhang, R. 2024b. Cancerllm: A large language model in cancer domain. *arXiv preprint arXiv:2406.10459*.

Li, M.; Zhou, H.; Yang, H.; and Zhang, R. 2024c. RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association*, 31(9): 1929–1938.

Li, Q.; Huang, C.; Li, S.; Xiang, Y.; Xiong, D.; and Lei, W. 2024d. GraphOTTER: Evolving LLM-based Graph Reasoning for Complex Table Question Answering. *arXiv preprint arXiv:2412.01230*.

Liu, A.; Feng, B.; Wang, B.; Wang, B.; Liu, B.; Zhao, C.; Dengr, C.; Ruan, C.; Dai, D.; Guo, D.; et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024b. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lovon, J.; Mouysset, M.; Oleiwan, J.; Moreno, J. G.; Damase-Michel, C.; and Tamine, L. 2025. Evaluating LLM Abilities to Understand Tabular Electronic Health Records: A Comprehensive Study of Patient Data Extraction and Retrieval. *arXiv preprint arXiv:2501.09384*.

Monajatipoor, M.; Yang, J.; Stremmel, J.; Emami, M.; Mohaghegh, F.; Rouhsedaghat, M.; and Chang, K.-W. 2024. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Ong, J.; Kedia, N.; Harihar, S.; Vupparaboina, S. C.; Singh, S. R.; Venkatesh, R.; Vupparaboina, K.; Bollepalli, S. C.; and Chhablani, J. 2023. Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding. *Journal of Medical Artificial Intelligence*, 6.

OpenAI. 2023. GPT-3.5 Turbo [Large language model]. https://chat.openai.com.

OpenAI. 2025. GPT-4.1 [Large language model]. https://platform.openai.com/docs/models/gpt-4.1.

Pollard, T. J.; Johnson, A. E.; Raffa, J. D.; Celi, L. A.; Mark, R. G.; and Badawi, O. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5(1): 1–13.

Ren, W.; Zhu, J.; Liu, Z.; Zhao, T.; and Honavar, V. 2025. A Comprehensive Survey of Electronic Health Record Modeling: From Deep Learning Approaches to Large Language Models. *arXiv preprint arXiv:2507.12774*.

Shool, S.; Adimi, S.; Saboori Amleshi, R.; Bitaraf, E.; Golpira, R.; and Tara, M. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1): 117.

Sun, Z.; Deng, N.; Yu, H.; and You, J. 2025. Table as Thought: Exploring Structured Thoughts in LLM Reasoning. *arXiv preprint arXiv:2501.02152*.

Tan, Y.; Zhang, Z.; Li, M.; Pan, F.; Duan, H.; Huang, Z.; Deng, H.; Yu, Z.; Yang, C.; Shen, G.; et al. 2024. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Computers in Biology and Medicine*, 172: 108290.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2412.15115*.

Ullah, E.; Parwani, A.; Baig, M. M.; and Singh, R. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1): 43.

Walonoski, J.; Kramer, M.; Nichols, J.; Quina, A.; Moesel, C.; Hall, D.; Duffett, C.; Dube, K.; Gallagher, T.; and McLachlan, S. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3): 230–238.

Wang, J.; Yang, Z.; Yao, Z.; and Yu, H. 2024a. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Wang, X.; Chen, N.; Chen, J.; Wang, Y.; Zhen, G.; Zhang, C.; Wu, X.; Hu, Y.; Gao, A.; Wan, X.; et al. 2024b. Apollo: A Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. *arXiv preprint arXiv:2403.03640*.

Wang, Z.; Zhu, Y.; Zhao, H.; Zheng, X.; Sui, D.; Wang, T.; Tang, W.; Wang, Y.; Harrison, E.; Pan, C.; et al. 2025. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, 2250–2261.

Wei, S.; Zhao, X.; and Miao, C. 2018. A comprehensive exploration to the machine learning techniques for diabetes identification. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 291–295. IEEE.

Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; and Wang, Y. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9): 1833–1843.

Yang, Y.; Krusche, P.; Pantoja, K.; Shi, C.; Ludmir, E.; Roberts, K.; and Zhu, G. 2024. Using Large Language Models to Generate Clinical Trial Tables and Figures. *arXiv preprint arXiv:2409.12046*.

Yang, Z.; Chen, L.; Cohan, A.; and Zhao, Y. 2025a. Table-r1: Inference-time scaling for table reasoning. *arXiv preprint arXiv:2505.23621*.

Yang, Z.; Du, Z.; Zhang, M.; Du, W.; Chen, J.; Duan, Z.; and Zhao, S. 2025b. Triples as the Key: Structuring Makes Decomposition and Verification Easier in LLM-based TableQA. In *The Thirteenth International Conference on Learning Representations*.

Yuan, D.; Rastogi, E.; Naik, G.; Rajagopal, S. P.; Goyal, S.; Zhao, F.; Chintagunta, B.; and Ward, J. 2024. A continued pretrained llm approach for automatic medical note generation. *arXiv preprint arXiv:2403.09057*.

Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Li, J.; Chen, G.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Zhang, Z.; Gao, Y.; and Lou, J.-G. 2024. e5: Zero-shot hierarchical table analysis using augmented LLMs via explain, extract, execute, exhibit and extrapolate. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1244–1258.

Zhao, X.; Chen, H.; Xing, Z.; and Miao, C. 2021. Brain-inspired search engine assistant based on knowledge graph. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4386–4400.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025a. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, 4442–4457.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025b. A smart multimodal healthcare copilot with powerful llm reasoning. *arXiv preprint arXiv:2506.02470*.

Zhao, X.; Xing, Z.; Kabir, M. A.; Sawada, N.; Li, J.; and Lin, S.-W. 2017. Hdskg: Harvesting domain specific knowledge graph from content of webpages. In *2017 ieee 24th international conference on software analysis, evolution and reengineering (saner)*, 56–67. IEEE.

Zhu, Y.; Wang, Z.; Gao, J.; Tong, Y.; An, J.; Liao, W.; Harrison, E. M.; Ma, L.; and Pan, C. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.