AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

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
* Corresponding Author

Motivation & Problem

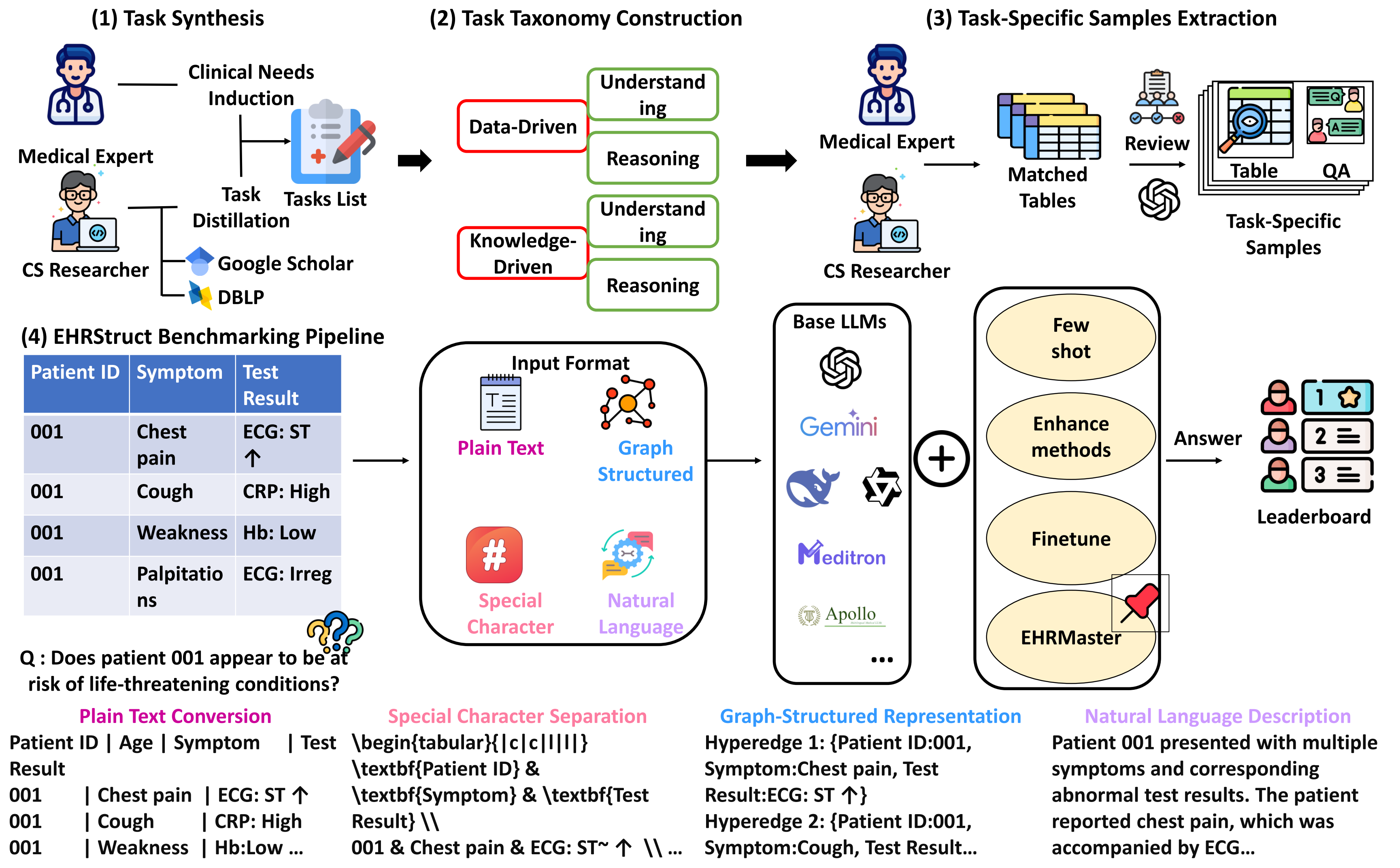
Structured electronic health records (EHRs) are central to clinical decision-making.
Large language models (LLMs) are increasingly applied to structured EHR modeling. However, existing evaluations remain fragmented:

- **Limited task coverage.**
- **Inconsistent datasets and protocols.**
- **Unclear reasoning requirements.**

What is EHRStruct?

EHRStruct is a comprehensive benchmark framework for evaluating LLMs on structured EHR data. It is built around four components:

- **Task synthesis**
- **Task taxonomy construction**
- **Task-specific sample extraction**
- **Model evaluation pipeline**



Patient ID	Symptom	Test Result
001	Chest pain	ECG: ST ↑
001	Cough	CRP: High
001	Weakness	Hb: Low
001	Palpitations	ECG: Irreg

Q : Does patient 001 appear to be at risk of life-threatening conditions?

Plain Text Conversion
Patient ID | Age | Symptom | Test Result
001 | Chest pain | ECG: ST ↑
001 | Cough | CRP: High
001 | Weakness | Hb:Low ...

Special Character Separation
\begin{tabular}{|c|c|l|l|}
\textbf{Patient ID} &
\textbf{Symptom} & \textbf{Test Result} \\
001 & Chest pain & ECG: ST~ ↑ \\ ...

Graph-Structured Representation
Hyperedge 1: {Patient ID:001, Symptom:Chest pain, Test Result:ECG: ST ↑}
Hyperedge 2: {Patient ID:001, Symptom:Cough, Test Result...

Natural Language Description
Patient 001 presented with multiple symptoms and corresponding abnormal test results. The patient reported chest pain, which was accompanied by ECG...

Task Taxonomy

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

Experimental Setup

- **Models**: 20 representative LLMs (general-purpose and medical-domain).
- **Evaluation axes**: task taxonomy, few-shot settings, input formats, finetuning, and enhancement methods.

Resources

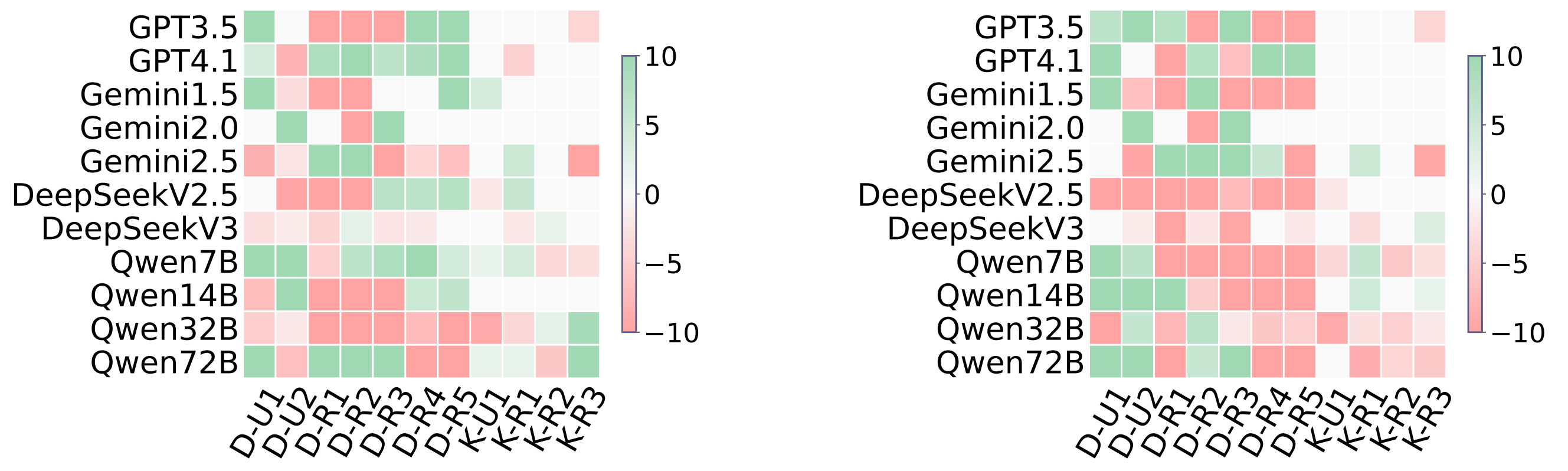
Project Page
Official Code
Full Paper

Key Findings

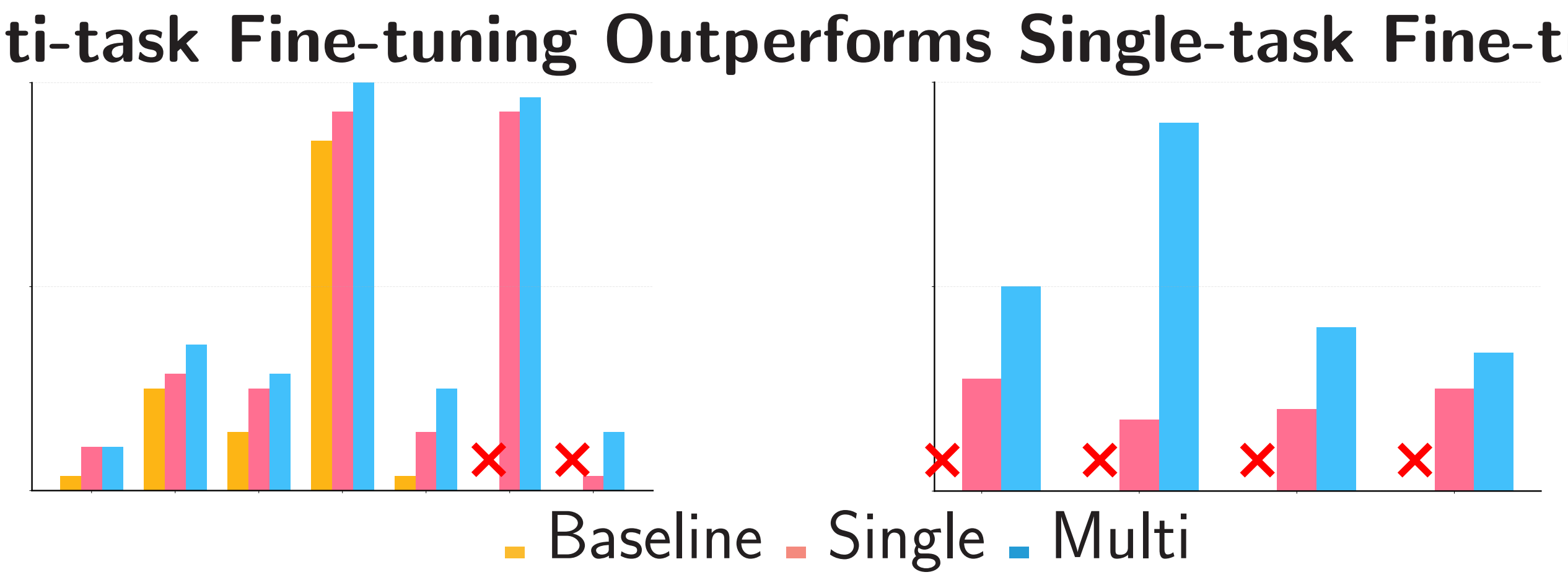
1. **General LLMs Outperform Medical LLMs.**
2. **LLMs Excel at Data-Driven Tasks.**

Types	Models	Data-Driven U (%) D-U1 ACC	D-U2 ACC	Data-Driven R (%) D-R1 ACC	D-R2 ACC	D-R3 ACC	D-R4 ACC	D-R5 ACC	Knowledge-Driven U (%) K-U1 AUC	Knowledge-Driven R (%) K-R1 AUC	K-R2 AUC	K-R3 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	✗	✗	✗	✗

3. **Input Format Influences Performance.**



4. **Multi-task Fine-tuning Outperforms Single-task Fine-tuning.**



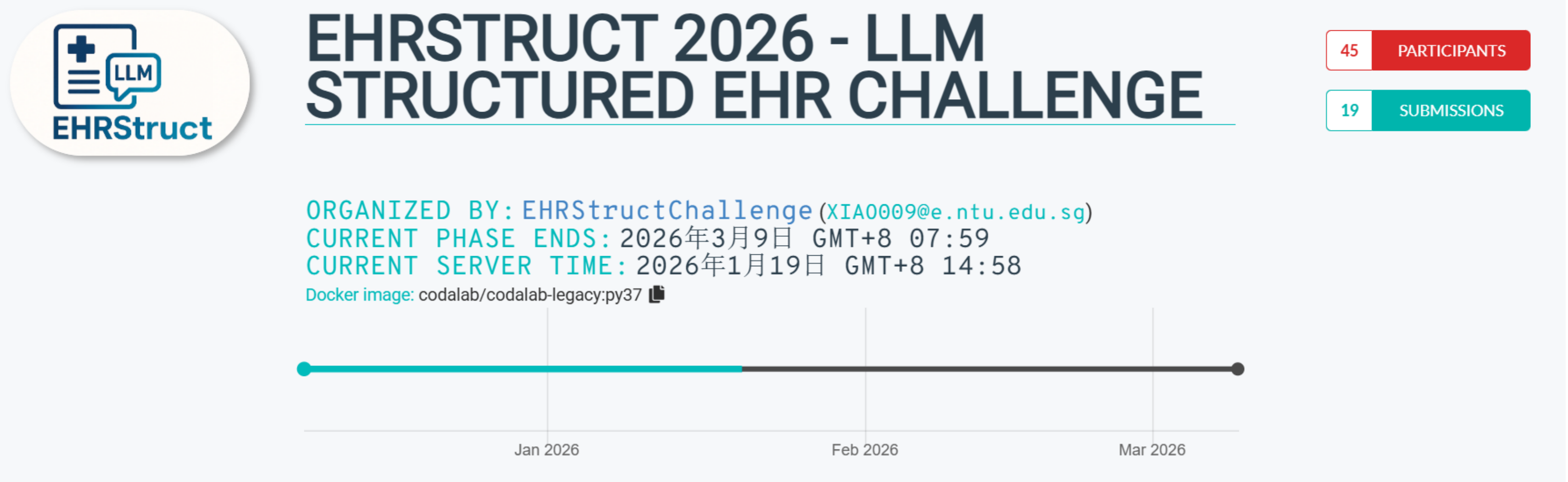
EHRMaster: A Three-Step Code-Augmented Framework

EHRMaster is a code-augmented framework tailored for structured EHR tasks.

- **Solution Planning**: Generates a high-level solution plan based on the task definition.
- **Concept Alignment**: Aligns the abstract concepts in the plan with relevant fields and tables in the structured EHR data.
- **Adaptive Execution**: Selects between code-based execution and direct language reasoning to derive the final answer.

Models	Methods	Data-Driven D-U1 ACC	D-U2 ACC	D-R1 ACC	D-R2 ACC	D-R3 ACC	D-R4 ACC	D-R5 ACC	Knowledge-Driven K-U1 AUC	K-R1 AUC	K-R2 AUC	K-R3 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

Take the Challenge on Codabench


EHRStruct
EHRSTRUCT 2026 - LLM STRUCTURED EHR CHALLENGE

45 PARTICIPANTS
19 SUBMISSIONS

ORGANIZED BY: EHRStructChallenge (XIA0009@e.ntu.edu.sg)
CURRENT PHASE ENDS: 2026年3月9日 GMT+8 07:59
CURRENT SERVER TIME: 2026年1月19日 GMT+8 14:58
Docker image: codalab/codalab-legacy:py37