

Instruction Tuning Large Language Models to Understand Electronic Health Records

Zhenbang Wu^{1,2,3}, Anant Dadu^{2,3}, Michael Nalls^{2,3}, *Faraz Faghri^{2,3}, *Jimeng Sun¹

¹ University of Illinois Urbana-Champaign

² National Institutes of Health

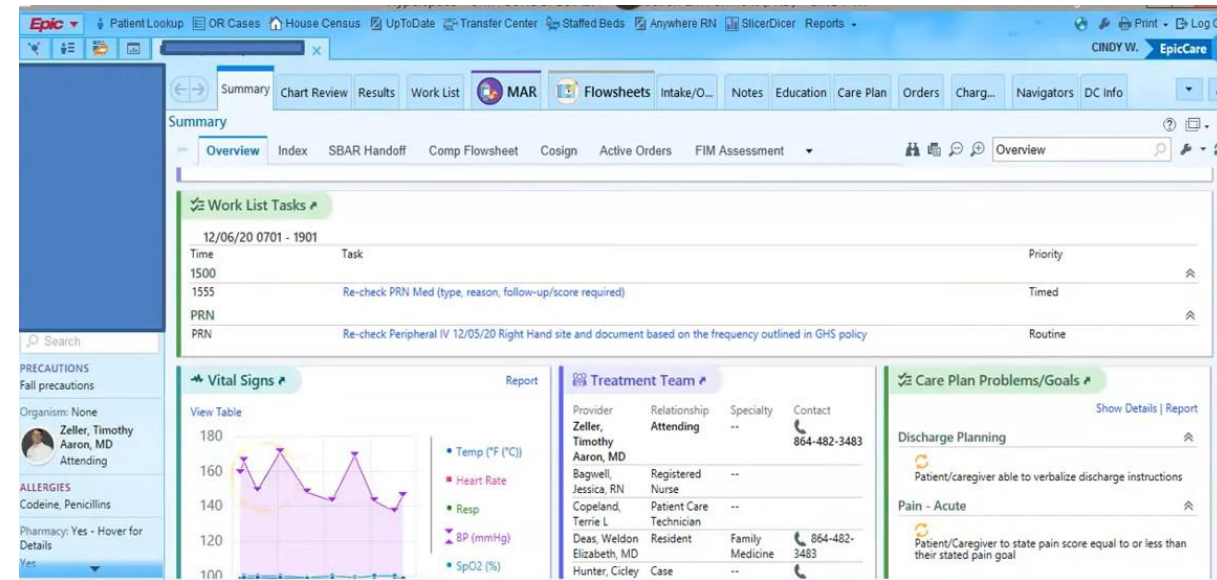
³ Data Tecnica

* co-corresponding

NeurIPS 2024 Datasets and Benchmarks Track (Spotlight)

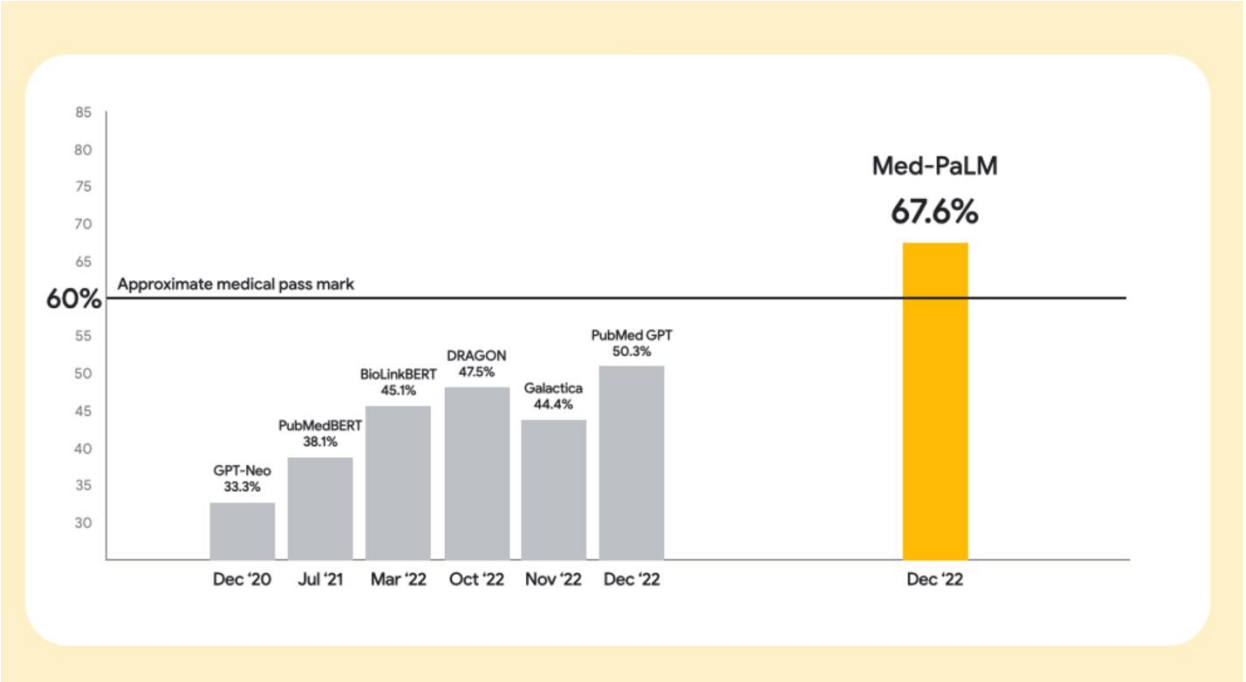
Background: EHR System

- EHR documents a patient's medical history and care
- Challenges of existing EHR systems:
 - navigating the user interface
 - vast amount of data that needs to be reviewed
 - extra clerical tasks directed to physicians
- There were 165.3 alerts/patient per day, but only 4.5% were important
- Physicians spent an average of 3.17 hours on EHR systems each day



Opportunity: Streamline EHR with LLM

- LLMs can understand complex inputs and follow human instructions to solve diverse tasks.
- LLMs also encode clinical knowledge



Example of USMLE-style question

A 32-year-old woman comes to the physician because of fatigue, breast tenderness, increased urinary frequency, and intermittent nausea for 2 weeks. Her last menstrual period was 7 weeks ago. She has a history of a seizure disorder treated with carbamazepine. Physical examination shows no abnormalities. A urine pregnancy test is positive. The child is at greatest risk of developing which of the following complications?

- A. Renal dysplasia
- B. Meningocele
- C. Sensorineural hearing loss
- D. Vaginal clear cell carcinoma

Goal & Challenges

Goal:

- Develop a conversational AI assistant for EHR data

Capabilities:

- Information extraction:

E.g., “What was the recorded Blood Oxygen level on admission?”

- Clinical reasoning:

E.g., “What is the recommended follow-up plan for the patient’s abdominal pain and gastrointestinal symptoms?”

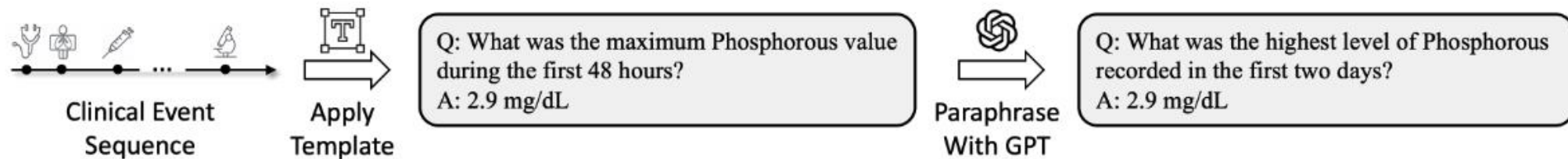
Challenge 1: Lack of Large-Scale Instruction-Following Data

Dataset	Size	Source	Format	Answer Type
MedQA [Jin et al., 2020]	13K	US medical licensing exam	Question + Answer	Multi Choice
MedMCQA [Pal et al., 2022]	6K	AIIMS and NEET PG entrance exams	Question + Answer	Multi Choice
PubMedQA [Jin et al., 2019]	0.5K	PubMed literature	Question + Context + Answer	Multi Choice
MMLU clinical [Hendrycks et al., 2021]	1K	US Medical Licensing Examination	Question + Answer	Multi Choice
EHRSQL [Lee et al., 2023]	24K	MIMIC-III	Question + Answer	SQL
EHRNoteQA [Kweon et al., 2024]	0.9K	MIMIC-IV	Question + Note + Answer	Free Text
MedAlign [Fleming et al., 2024]	0.9K	EHRs (Stanford University)	Question + EHR + Answer	Free Text

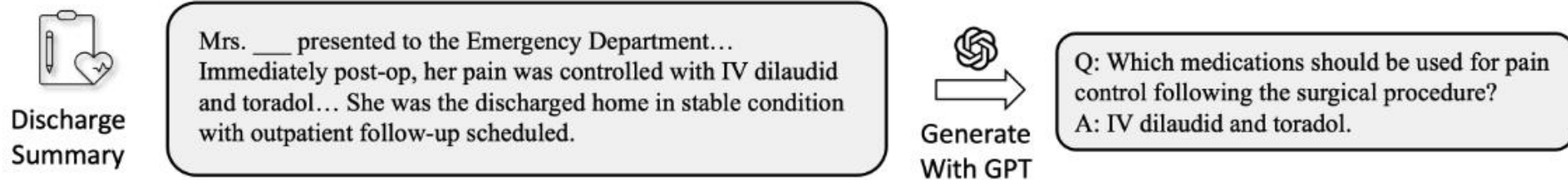
Challenge 2: Heterogeneous EHR Data

- Complex preprocessing
 - Code mapping
 - Concept standardization
 - Unit normalization
- Feature selection
 - Manually define a subset of features out of hundreds/thousands of events

Solution 1: A dataset of 400K EHR Instruction-Following Data



(a) Generation Pipeline for Schema Alignment Data

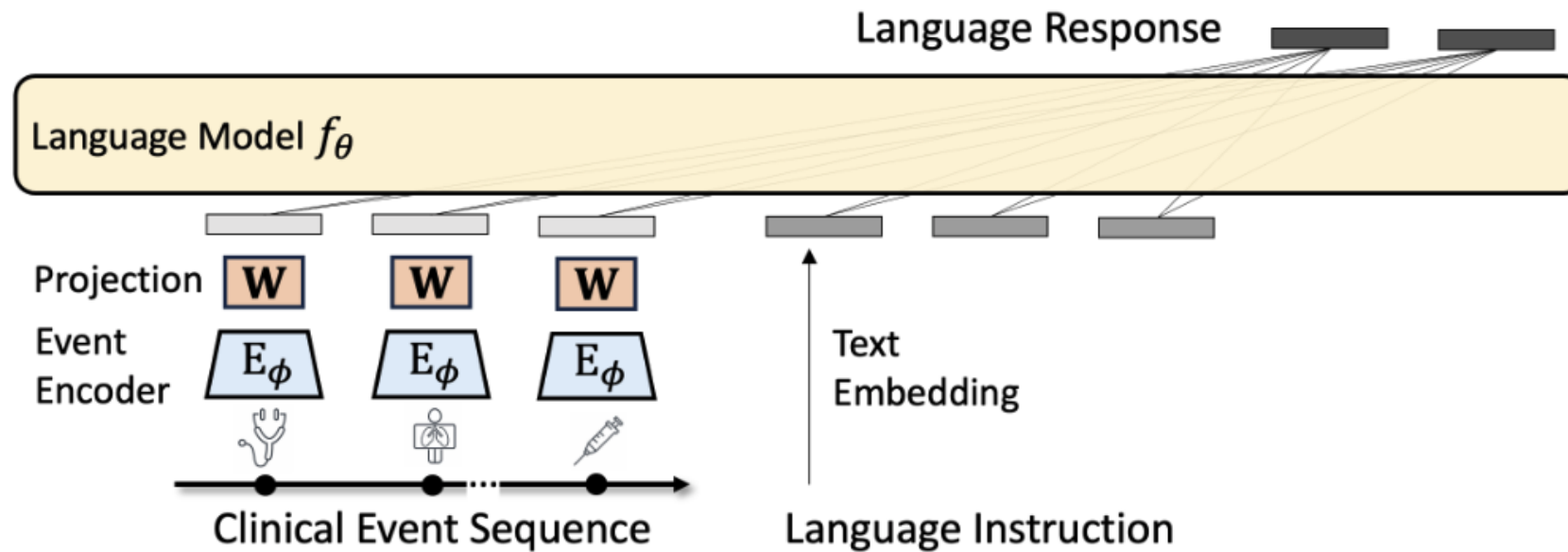


(b) Generation Pipeline for Instruction Following Data

Preliminary: Medical Event Data Standard

- Represent each patient's EHR data as a sequence of event
- Each event consists of a timestamp, a type, and a value
- Convert event to text:
 - 10 min, vital signs for heart rate 75 bpm
 - 20 min, vital signs for blood oxygen 97%
 - 235 min, lab measure for white blood cells, value 3.7 K/uL

Solution 2: A Foundation Model for EHR Data



Two-Stage Training

- Stage 1: Training for Schema Alignment
 - Only train the projection matrix with 350K QA pairs generated from the template and then paraphrased by GPT-3.5
- Stage 2: Training for Clinical Reasoning
 - Train both the LLM and the projection matrix with 50K QA pairs for clinical reasoning

Results 1: Llemr as a Conversational Clinical AI Assistant

Model	Schema Alignment	Clinical Reasoning	Overall
Llama-2-7b-chat-hf [Touvron et al., 2023]	47.66 ± 15.31	47.55 ± 11.73	47.60 ± 9.62
SynthIA-7B-v1.3 [Tissera, 2023]	47.18 ± 5.84	49.16 ± 4.99	48.17 ± 3.83
Mistral-7B-OpenOrca [Lian et al., 2023]	51.75 ± 8.20	51.18 ± 7.67	51.46 ± 5.60
Llama-3-8b-Instruct [Touvron et al., 2023]	56.18 ± 7.08	55.07 ± 7.25	55.62 ± 5.05
MPT-7b-8k-instruct [MosaicML, 2023]	68.13 ± 8.95	53.90 ± 4.92	61.01 ± 5.19
vicuna-7b-v1.5 [Chiang et al., 2023]	66.81 ± 5.61	62.40 ± 4.59	64.60 ± 3.63
dolphin-2.0-mistral-7b [Cognitive, 2023]	63.06 ± 5.36	72.66 ± 7.47	67.86 ± 4.64
Llemr + Stage 1	69.71 ± 6.32	64.35 ± 7.21	67.03 ± 6.83
Llemr + Stage 1&2	70.42 ± 5.88	76.23 ± 4.23	73.33 ± 5.30

Results 2: Llemr on Standard Clinical Predictive Benchmarks

Method	Mortality	Readmission	Length-of-Stay	Diagnosis
RNN [Cho et al., 2014]	0.8002 (0.02)	0.6643 (0.01)	0.6833 (0.03)	0.7735 (0.01)
Transformer [Vaswani et al., 2017]	0.8241 (0.03)	0.7006 (0.01)	0.6990 (0.01)	0.8025 (0.02)
RETAIN [Choi et al., 2016a]	0.8302 (0.02)	0.6994 (0.01)	0.7015 (0.01)	0.8073 (0.02)
GRASP [Zhang et al., 2021]	0.8362 (0.01)	0.7155 (0.01)	0.7100 (0.03)	0.8005 (0.02)
GenHPF [Hur et al., 2024]	0.8258 (0.02)	0.7102 (0.01)	0.6993 (0.02)	0.8103 (0.03)
REMed [Kim et al., 2024]	0.8346 (0.02)	0.7193 (0.02)	0.7018 (0.01)	0.8128 (0.01)
Llemr	0.8388 (0.01)	0.7251 (0.03)	0.7132 (0.01)	0.8086 (0.01)

Conclusion

Zhenbang Wu
CS Ph.D. Student @ UIUC
zw12@illinois.edu

- Goal
 - Develop a conversational AI assistant for EHR data
- Challenges
 - Lack of large-scale instruction-following data
 - Heterogeneous EHR data
- Method
 - MIMIC-Instr, a dataset of over 400K open-ended instruction-tuning data generated by GPT-3.5
 - Llemr, a generic framework designed to empower LLMs to encode EHR data with heterogeneous schema
- Result
 - Outperforms SOTA LLMs in answering diverse inquiries about a patient
 - Performs on par with SOTA baselines when further fine-tuned for clinical predictive tasks