

---

# Instruction Tuning Large Language Models to Understand Electronic Health Records

---

Zhenbang Wu<sup>1,2,3</sup>, Anant Dadu<sup>2,3</sup>, Mike Nalls<sup>2,3</sup>, Faraz Faghri<sup>2,3\*</sup>, Jimeng Sun<sup>1\*</sup>  
<sup>1</sup>UIUC, <sup>2</sup>National Institutes of Health, <sup>3</sup>Data Tecnica  
\*Co-corresponding authors

## Abstract

Large language models (LLMs) have shown impressive capabilities in solving a wide range of tasks based on human instructions. However, developing a conversational AI assistant for electronic health record (EHR) data remains challenging due to (1) the lack of large-scale instruction-following datasets and (2) the limitations of existing model architectures in handling complex and heterogeneous EHR data. In this paper, we introduce MIMIC-Inst<sub>r</sub>, a dataset comprising over 400K open-ended instruction-following examples derived from the MIMIC-IV EHR database. This dataset covers various topics and is suitable for instruction-tuning general-purpose LLMs for diverse clinical use cases. Additionally, we propose L<sub>l</sub>em<sub>r</sub>, a general framework that enables LLMs to process and interpret EHRs with complex data structures. L<sub>l</sub>em<sub>r</sub> demonstrates competitive performance in answering a wide range of patient-related questions based on EHR data. Furthermore, our evaluations on clinical predictive modeling benchmarks reveal that the fine-tuned L<sub>l</sub>em<sub>r</sub> achieves performance comparable to state-of-the-art (SOTA) baselines using curated features. The dataset and code are available at <https://github.com/zzachw/llemr>.

## 1 Introduction

EHRs document a patient’s medical history and care, including demographics, diagnoses, laboratory test results, medication prescriptions, and clinical notes [Evans, 2018]. Despite the potential benefits in supporting clinical decision-making and care coordination, EHR systems also lead to physician burnout due to challenges in navigating the user interface, the large volume of data that needs to be reviewed for each medical decision, and the extra clerical tasks directed to physicians [Melnick et al., 2020, Tajirian et al., 2020, DeChant et al., 2019]. Previous studies show that physicians spend an average of 3.17 hours daily on EHR systems [Overhage and McCallie, 2020]. This not only detracts from patient care but also reduces the time physicians can spend interacting directly with patients.

Advances in LLMs offer an opportunity to streamline EHR processes and ease the load on healthcare providers. LLMs have revolutionized natural language processing fields in tasks such as question answering [Wei et al., 2022a, Touvron et al., 2023], visual understanding [Liu et al., 2023, Li et al., 2023a, 2022, 2023b], reasoning [Wei et al., 2022b, Brown et al., 2020], and code generation [Black et al., 2022]. They have demonstrated remarkable capabilities in understanding complex inputs and following human instructions to solve diverse tasks. Recent works have further shown that LLMs can achieve expert-level performance on multiple-choice questions from medical licensing exams [Singhal et al., 2022, 2023]. However, despite these advances, developing a conversational AI assistant specifically for EHR data remains a significant challenge.

**Challenge 1: Lack of large-scale instruction-following data.** LLMs are typically fine-tuned on large-scale instruction-following datasets to understand user instructions and perform a variety of tasks [Wei et al., 2022a]. These datasets are created using manually defined templates or with

the assistance of LLMs. The construction process requires substantial efforts and becomes even more complex when data must be paired with patient EHRs. Thus, most prior works mainly focus on the clinical notes [Kweon et al., 2024, Lehman et al., 2022, Yue et al., 2021], as generating instruction-following data from free text is comparatively straightforward. However, a substantial amount of information exists solely within structured EHR data (e.g., relational tables). Although some question-answering (QA) datasets are based on structured EHR data [Pampari et al., 2018, Lee et al., 2023, Tang et al., 2023], they mainly focus on factoid extraction and lack alignment with real-world clinical decision-making, which often requires complex reasoning. Moreover, existing datasets are limited in size [Fleming et al., 2024], ranging from thousands to tens of thousands of examples, which is insufficient for effective LLM instruction tuning.

**Challenge 2: Limitations of existing model architectures in handling complex and heterogeneous EHR data.** Due to the complex schemas and various standardizations of EHR data [Gamal et al., 2021], most existing clinical predictive models depend heavily on manual data preprocessing [Harutyunyan et al., 2019, Choi et al., 2016a,b]. This preprocessing includes steps like feature selection, code mapping, unit standardization, value normalization, and imputation. Such manual processes demand significant time and expertise and may introduce human errors and biases into the data. While recent works have attempted to develop a unified foundation model for EHRs [Hur et al., 2024, Kim et al., 2024], these models are generally limited to specific tasks and lack interactive capabilities.

In this paper, we introduce `MIMIC-InstR`, a dataset of over 400K EHR-grounded instruction-following examples based on the publicly available MIMIC-IV EHR database [Johnson et al., 2023]. This dataset is divided into two parts: **(1) Schema alignment subset:** A set of 350K QA pairs was constructed from over 100 templates and subsequently paraphrased using GPT-3.5<sup>1</sup>. These questions query various information from the structured EHR data, such as patient demographics, diagnoses, treatment histories, and test results. They are designed to train LLMs on the ability to navigate and extract specific information from the complex and heterogeneous EHR data. **(2) Clinical reasoning subset:** Another set of 50K QA pairs was generated from discharge summaries with GPT-3.5. Discharge summaries capture the complexities of patient cases and the rationales behind medical decisions. This subset challenges LLMs to go beyond simple fact extraction, engaging in deeper clinical reasoning tasks such as understanding the progression of a patient’s condition, predicting possible complications, and suggesting appropriate follow-up actions.

To address the second challenge, we propose `LLeMR`, an instruction-tuned LLM for electronic medical records (EMRs)<sup>2</sup>. We adopt the Medical Event Data Standard [Arnrich et al.] and represent each patient’s EHR data as a stream of clinical events (e.g., procedures, prescriptions, and transfers). Each event is formatted as a triplet of timestamp, type, and value, such as (2024-03-11 13:32:26, Lab, Hemoglobin 12 g/dl), which can be combined into sentence-like representations. This format simplifies inputs and is robust to variations in EHR schemas and standards. While the concatenated event sequence can be directly input into LLMs, it often exceeds the context length of LLMs for patients with extensive event histories. Inspired by `REMed` [Kim et al., 2024], we utilize `ClinicalBERT` [Alsentzer et al., 2019] to first encode each event into an embedding, and then feed the sequence of event embeddings into the LLM with an additional mapping layer (i.e., a linear projection). This approach largely reduces the input length and speeds up the training. `LLeMR` is trained in a curriculum learning fashion: it initially learns to navigate through EHR data via 350K schema alignment examples, and then progresses to more complex reasoning with the 50K clinical reasoning examples.

`LLeMR` exhibits excellent ability to answer diverse inquiries about a patient. Further, our evaluation on standard clinical predictive benchmarks shows that the fine-tuned `LLeMR` achieves performance comparable to SOTA baselines using manually curated features.

In summary, this paper makes the following contributions:

- **Clinical instruction-following data.** We created a dataset of 400K instruction-following examples based on the MIMIC-IV database. This dataset enables instruction-tuning of general LLMs to better understand EHRs.

---

<sup>1</sup>We used Azure’s HIPAA-compliant platform in accordance with PhysioNet’s regulations.

<sup>2</sup>We use EMR and EHR interchangeably in this paper.

Table 1: Comparison between MIMIC-Instr and existing clinical question-answering / instruction-following datasets.

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
MIMIC-Instr	400K	MIMIC-IV	Question + EHR + Answer	Free Text

- **Foundation model for EHR.** We introduce L1emr, a general framework to empower LLMs to perform both information extraction and clinical reasoning on EHR data.
- **Open-source.** We will release the instruction-following data via PhysioNet<sup>3</sup>, and also share the code and model weights to facilitate future research.

## 2 Related Work

### 2.1 Clinical Instruction-Tuning Data

Instruction tuning fine-tunes a pre-trained LLM using pairs of instructions and responses. It generalizes LLMs’ capabilities beyond next token prediction to diverse new tasks described with instructions. Generally, the instruction tuning datasets are constructed either with manually-defined templates or LLMs like GPT. While there are increasing interests in adapting LLMs to the clinical domain, existing works mainly focus on broad clinical tasks with natural language inputs, such as answering medical licensing exam questions and consumer queries [Singhal et al., 2022, Johri et al., Tu et al., 2024, 2023], information extraction and text summarizing [Tran et al., 2023, Zhang et al., 2024], and ICD coding [Wang et al., 2024]. These tasks are typically based on natural language as inputs and thus the instruction tuning data is relatively easy to generate. However, existing works fall short in instruction tuning a LLM to understand EHR data. While there are some recent benchmarks evaluating the capability of LLM on EHR data, they either focus on information extraction tasks [Lee et al., 2023, Shi et al., 2024], or solely perform model evaluation [Kweon et al., 2024, Fleming et al., 2024, Zakka et al., 2024], as the datasets are too small to enable instruction tuning. To bridge this gap, we release a dataset of 400K instruction-response examples on patient EHR data covering a broad range of topics and can be used to instruction tune general-purpose LLMs to understand EHR data.

### 2.2 Foundation Model for EHR

With the wide adoption of EHR systems, there has been growing interest in utilizing deep learning models in interpreting and analyzing EHR data to assist clinical decision-making and improve patient outcome [Choi et al., 2016b, Tan et al., 2022, Choi et al., 2017, Zhang et al., 2021, Li et al., 2020, Xu et al., 2024]. However, real-world EHR data is often very messy and has complex schemas. Thus, most existing methods typically require manual feature selection and complex data preprocessing to harmonize the data. This process requires domain expertise and is very time-consuming. Yet, the developed model is often task-specific and the same process needs to be performed again for new tasks. In recent years, language models have demonstrated remarkable capability in understanding diverse text inputs [Brown et al., 2020]. Thus, some recent methods try to utilize them to encode clinical events and eliminate the need for feature selection and data preprocessing [Hur et al., 2024, Kim et al., 2024]. However, they still follow the task-specific supervised training paradigm and do not fully utilize the power of LLMs. In this work, we also follow recent trends in converting clinical events into text and utilize LLMs to interpret them. But we take a step further in tuning LLMs to follow instructions and generalize to unseen tasks.

<sup>3</sup>Under the PhysioNet Credentialed Health Data License: <https://www.physionet.org/content/ehr-ds-qa/view-license/1.0.0/>.

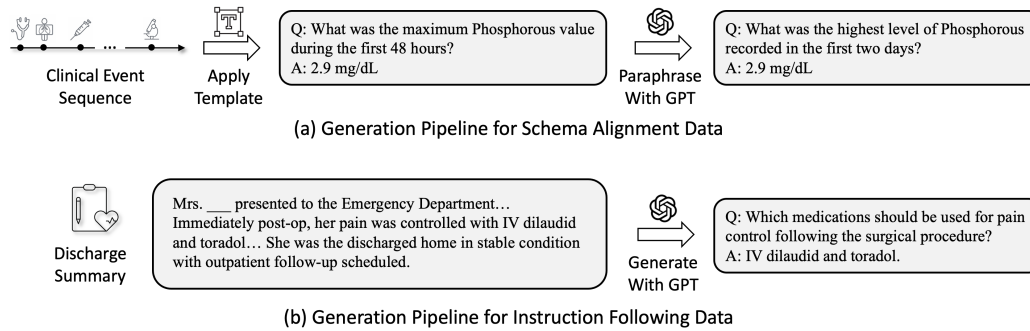


Figure 1: Illustration of the construction process of MIMIC-Inst.r, a dataset of 400K EHR-grounded instruction-following examples based on the publicly available MIMIC-IV database. It serves two sequential purposes: adapting LLMs to the EHR schema and teaching LLMs to perform in-depth clinical reasoning.

### 3 Preliminaries

EHR records comprehensive clinical information about a patient, including demographics, diagnoses, prescriptions, laboratory and microbiology tests, vital signs, and more. Most existing clinical predictive models rely on expert-defined features to construct a predictive pipeline [Harutyunyan et al., 2019]. This process is labor-intensive, requiring significant domain expertise, and is often repetitive across different tasks. To alleviate this, we adopt the Medical Event Data Standard [Arrrich et al.] and represent each patient’s EHR data as a sequence of events. Each event consists of a timestamp (e.g., 2024-03-11 13:32:26) denoting when the event occurred, a type (e.g., Lab) denoting the category of the event, and a value (e.g., Hemoglobin 12 g/dl) denoting the content of the event. This approach allows a unified representation across different EHR schemas.

Besides clinical event sequence, EHR data often includes] discharge summaries written by doctors or nurses at the time of discharge. Discharge summaries provide a comprehensive overview of a patient’s hospital stay, detailing the reason for admission, treatments provided, patient’s responses to treatment, and recommendations for follow-up care. This data is usually not used in clinical predictive modeling as it is only available at hospital discharge. In this work, we only leverage discharge summaries as a complementary source to generate instruction-following data.

### 4 Clinical Instruction-Following Data

With the growing digitization of healthcare, EHR data is now routinely collected [Evans, 2018]. However, clinical instruction-following data remains limited because its creation is time-consuming and requires significant domain expertise. Therefore, many existing works resort to medical exam questions to tune and evaluate LLM performance in the medical domain [Singhal et al., 2023, Tu et al., 2023]. Yet, these exam-style questions are quite different from how doctors interact with LLMs in real-world clinical practice. To adapt general-purpose LLMs for clinical use, we must first enable them to understand EHR schemas and to reason effectively over EHR data. Inspired by the success of recent works in utilizing GPT to generate instruction-following data [Liu et al., 2023, Li et al., 2023a], we created a clinical instruction-following dataset through a machine-human co-curation process. This dataset includes two subsets: information extraction data and clinical reasoning data, which are used at different training stages. An overview of the data generation process can be found in Figure 1.

#### 4.1 MIMIC-IV EHR Database Preparation

We construct our cohort from ICU patients in the MIMIC-IV [Johnson et al., 2023] database. This database contains 51K patients admitted to the ICU at Beth Israel Dea-

Table 2: Example questions from MIMIC-Instr.

<b>Schema Alignment</b>	What was the ethnicity of the individual?
	What was the patient’s Cr level at the time of discharge?
	What was the highest recorded Blood Oxygen Blood Gas level on day 13?
	Which organisms were detected in the MRSA SCREEN sample after 104.00 hours?
<b>Clinical Reasoning</b>	What was the average measurement of Blood Lymphocytes Hematology within the initial 12-hour period?
	What was the reason for the patient’s hypotension upon presentation to the emergency department?
	What interventions were performed for the patient’s pancreatic pseudocyst?
	What is the recommended follow-up plan for the patient’s abdominal pain and gastrointestinal symptoms?
	Why was a statin not started for the patient despite other medications being titrated to effect?
	What is the recommended dose for levalbuterol for the patient’s severe COPD exacerbation?

coness Medical Center. We filter out patients without discharge summaries, with more than two ICU stays per hospital admission, and with negative ICU or hospital length-of-stay. We then select the following tables from MIMIC-IV: hosp/patients, hosp/admissions, hosp/diagnosis, hosp/labevents, hosp/microbiologyevents, hosp/prescriptions, hosp/transfers, icu/icustays, icu/inputevents, icu/outputevents, icu/procedureevents. Note that icu/chartevents table is excluded due to two reasons: it contains dense bedside monitor data, which is better treated as a time series rather than as an event sequence, and it has substantial overlap with other tables, such as hosp/labevents [Johnson et al., 2023]. In the end, we have a total of 55846 admissions. We hold out 10% each for validation and testing.

## 4.2 Data for Schema Alignment

As described in Section 3, EHR data has a schema that is fundamentally different from general text. To bridge this gap, we first created a set of 350K instruction-tuning examples focused on clinical information extraction. Specifically, for each type of clinical event, we developed a set of question templates (e.g., “which {measurement\_name} performed on the {specimen\_name} were abnormal {time\_period}?”). These templates query diverse information from patient EHR data in the MIMIC-IV database. Each question template is paired with a manually crafted Python script that extracts the ground-truth answer from the corresponding EHR table.

Given a patient’s EHR data, we randomly select a template to generate a corresponding question-answer pair (e.g., Q: “Which Blood Gas measurement on the Blood specimen were abnormal at the 650.05 hour?” A: “Calculated Total CO2, pCO2, pO2.”). Since the generated QA pairs all follow some fixed template, which limits their effectiveness for training LLMs to interpret diverse instructions, we leveraged GPT-3.5 to paraphrase the generated QA pairs without altering their meanings (e.g., Q: “Show me the abnormal blood gas measurements at the 650.05 hours?” A: “The calculated total CO2, pCO2, pO2 were abnormal.”)

In this way, we generated 350K QA pairs focused on information retrieval. This set of instruction-tuning QA pairs primarily asks about the extraction and aggregation of specific factual information from EHR data, serving as a foundational step for enabling LLMs to perform deeper clinical reasoning on EHR data.

## 4.3 Data for Clinical Reasoning

Expectations for clinical assistant AI often go beyond information extraction to following various instructions and performing clinical reasoning. To align the model with this goal, we created diverse instruction-following data focused on clinical reasoning using GPT-3.5. Specifically, we prompted GPT-3.5 to generate questions and answers that resemble those doctors might ask in real-world clinical settings. We also manually created few-shot examples in the prompt to demonstrate how to generate high-quality QA pairs.

However, struggled to interpret raw clinical event sequences (converted to text) due to the unique structure of EHR data. So instead, we leveraged complimentary discharge summaries from the MIMIC-IV database as input to generate QA pairs. Compared to raw clinical event sequences, discharge summaries provide a more concise overview of the patient’s hospital trajectory and often

include the rationale behind treatments and plans for future care. This makes the QA data generated from discharge summaries better suited for clinical reasoning tasks. In this way, we generated another 50K QA pairs to equip the model with clinical reasoning abilities.

Note that since the discharge summaries are usually generated at the end of hospital admission and are not available for many real-time clinical predictive tasks, we only use discharge summaries to generate instruction-tuning data. The inputs to our foundation clinical model (introduced next) consist solely of clinical event sequences.

## 5 L1emr: A Foundation Model for EHR Data

With the generated data, we further propose L1emr, a simple yet effective model for EHR data.

### 5.1 Architecture

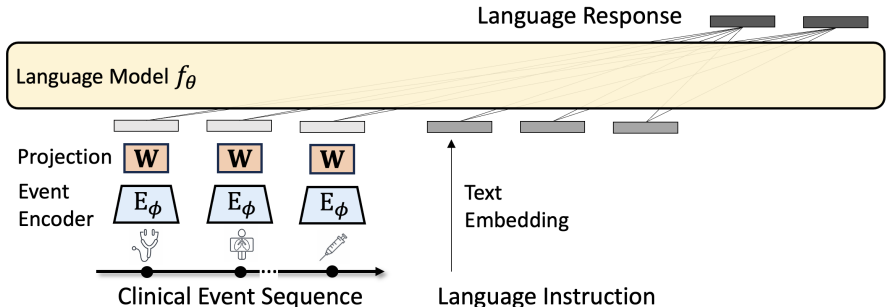


Figure 2: Model architecture of L1emr.

Given a patient’s EHR event sequence and a language instruction from the user, L1emr learns to generate a free-text response. Due to the high volume of events associated with MIMIC-IV ICU patients, instead of directly encoding raw events, L1emr utilizes an additional event encoder. Specifically, for each patient’s event sequence, ClinicalBERT [Alsentzer et al., 2019] is first used to encode each event into an embedding. Then, L1emr applies a linear projection layer to map the event embeddings to the word embedding space. The sequence of event embeddings is further concatenated with the token embeddings for the language instruction and fed into the backbone LLM (i.e., vicuna-7b-v1.5 [Chiang et al., 2023]).

### 5.2 Training with Curriculum Learning

Inspired by LLaVA [Liu et al., 2023], we adopt a two-stage curriculum training procedure to first bridge the schema gap between general text and EHR data, and then equip the model with clinical reasoning abilities.

#### 5.2.1 Stage 1: Training for Schema Alignment

We utilize the 350K QA pairs generated from templates and paraphrased by GPT-3.5. For each patient, given the event sequence and language instruction, we ask the model to generate the corresponding response. We keep both the event encoder and LLM backbone frozen and only update the projection matrix. In this way, we can teach LLMs to interpret the outputs from the event encoder and bridge the schema gap between general text and EHR data.

#### 5.2.2 Stage 2: Training for Clinical Reasoning

In this stage, we continue to tune the model to follow more complex instructions and perform clinical reasoning. We utilize the other 50K QA pairs for clinical reasoning and tune both the LLM and projection matrix. The weights of the event encoder are kept frozen. This allows L1emr to go beyond

Table 3: Evaluation on the performance of L1emr as a conversation 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
L1emr + Stage 1	<b>69.71 ± 6.32</b>	64.35 ± 7.21	67.03 ± 6.83
L1emr + Stage 1&2	<b>70.42 ± 5.88</b>	<b>76.23 ± 4.23</b>	<b>73.33 ± 5.30</b>

information extraction, enabling it to perform more advanced clinical reasoning tasks based on the given instructions.

## 6 Experiments

We conduct experiments to evaluate two key components: the quality of the generated clinical instruction-following data, and the performance of the proposed L1emr. We design experiments to answer the following research questions: (1) How well does L1emr perform as a conversational clinical AI assistant? (2) How does L1emr compare to SOTA methods on standard clinical predictive benchmarks?

### 6.1 Performance as Conversational AI Assistant

**Datasets.** To evaluate the performance of L1emr as a conversational clinical AI assistant, we construct a separate test set of 200 QA pairs. Specifically, we randomly select 100 ICU admissions from the previously held-out test set. These 100 admissions are unseen during training. Then, we follow the procedure described in Section 4 to generate 100 questions each for schema alignment and clinical reasoning.

**Metrics.** We follow Li et al. [2023a] and leverage GPT to quantify the quality of the generated responses. For each question, we generate a reference response with GPT-4. Then, we feed the question, the ground-truth answer, the GPT-4 generated answer, and the candidate LLM generated answer to GPT-3.5 and ask it to score both answers on a scale of 1 to 10, with higher scores indicating better performance. We prompt the GPT-3.5 to consider the helpfulness, relevance, accuracy, and level of detail of the responses. Next, we compute the relative score for the candidate LLM by normalizing against the GPT-4 reference score. We also report 95% confidence interval based on the 200 questions.

**Baselines.** We compare L1emr against seven popular open-sourced LLMs. For each baseline model, we concatenate the patient’s event sequence and the corresponding question and input them directly into the LLM. These baseline LLMs have a context length from 4K to 32K tokens. For patients with long event sequences, we adopt Fleming et al. [2024]’s multi-step refinement approach. We segment the EHR sequence into chunks and process them sequentially. LLMs will generate an initial response from the first chunk, then decide whether to update or maintain this response as they process each subsequent chunk.

**Results.** The results are reported in Table 3. Overall, we can see that L1emr outperforms all existing LLM baselines and matches 73% of the GPT-4 performance. Surprisingly, we find that stage-1 training alone yields performance improvements, despite only the linear projection layer is trained. This may be due to L1emr ’s architecture, which enables it to process the entire patient EHR sequence at once, whereas the baseline LLMs process it in chunks. Incorporating Stage 2 training further boosts performance, especially for clinical reasoning tasks.

In Table 4, we show examples of the generated responses and compare L1emr against its backbone LLM Vicuna [Chiang et al., 2023]. Generally, we find that L1emr tends to give more concise and

Table 4: Example of the generated response from L1emr.

Schema Alignment	
User	When was the first Blood Anion Gap Chemistry measured?
Gold	6.02 hours after admission.
Vicuna	The initial Blood Anion Gap Chemistry measurement was 14 mEq/L, which is within the normal range of 8-16 mEq/L.
L1emr	The measurement was taken at 6.02 hours.
User	Where was the patient admitted?
Gold	The patient was admitted to the emergency room.
Vicuna	The patient was admitted to the hospital with an emergency room complaint of agitation. The patient was transferred to the Medical Intensive Care Unit (MICU). The patient’s diagnosis is not provided in the given data.
L1emr	The patient was initially admitted to the Emergency Room.
Clinical Reasoning	
User	What was the rationale for not pursuing surgery for the patient with severe c. diff and septic shock?
Gold	Surgery was not considered for the patient due to being a poor surgical candidate with severe lung disease and because surgery was not aligned with the patient’s goals of care based on discussions with him and his family.
Vicuna	The patient was not pursued surgically due to concerns about the patient’s overall condition and the potential risks of surgery. The patient’s lab results showed abnormal values for various chemistry and hematology tests.
L1emr	The rationale for not pursuing surgery for the patient was likely due to the patient’s critical condition and the presence of multiple lung diseases.
User	What antibiotic was used for perioperative coverage during the aortic valve replacement surgery?
Gold	Cefazolin was used for perioperative antibiotic coverage during the aortic valve replacement surgery.
Vicuna	The antibiotic used for perioperative coverage during the aortic valve replacement surgery was likely Omeprazole.
L1emr	The data only includes various medications such as analgesics, sedatives, and other treatments, but no antibiotics were prescribed specifically for the surgery.

precise answers. In contrast, Vicuna [Chiang et al., 2023] tends to give broad responses or directly copy information from the input sequence. We also note that the last question contains information only from the discharge summaries. This can happen as the clinical reasoning data was generated from discharge summaries instead of EHR tables. This points out a potential future direction of further filtering the generated instruction-responses data.

## 6.2 Performance on Standard Clinical Predictive Benchmarks

**Datasets.** We leverage the held-out test set of 5585 ICU admissions from the MIMIC-IV database. The training and validation sets share the same cohorts as the ones used in instruction-tuning. Additional patient filtering is performed for each task introduced below.

**Tasks & Metrics.** We focus on four common clinical predictive tasks.

- **Mortality prediction** aims to predict whether the patient will pass away upon discharge using events from the first 48 hours of the hospital admission.
- **Length-of-stay prediction** aims to determine whether the patient’s hospital stay will be longer than 7 days using the first 48 hours of the hospital admission. For both mortality and length-of-stay prediction tasks, patients with hospital length-of-stay less than 48 hours are filtered.
- **Readmission prediction** aims to predict whether the patient will be readmitted back to the hospital within 14 days following current discharge using all events from the current admission. Patients who are deceased in the current hospital admission are filtered for this task.
- **Diagnosis classification** aims to classify which acute care conditions are present using all events from the current admission. We follow existing works [Harutyunyan et al., 2019] and define disease labels with 25 conditions that are common in adult ICUs, including 12 critical conditions, (e.g.,



Table 5: Results on the MIMIC-IV clinical predictive benchmark tasks.

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)	<b>0.8128 (0.01)</b>
L1emr (further-trained)	<b>0.8388 (0.01)</b>	<b>0.7251 (0.03)</b>	<b>0.7132 (0.01)</b>	0.8086 (0.01)

respiratory failure; 8 chronic conditions (e.g., diabetes); and 5 mixed (i.e., recurring or chronic with periodic acute episodes) conditions (e.g., cardiac dysrhythmias).

The first three tasks are binary classification and we calculate the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) scores. The last task is multilabel classification and we report the macro AUC-ROC. We report the average scores and standard deviation by performing bootstrapping (i.e., sampling with replacement) 1000 times.

**Baselines.** We compare L1emr with three groups of baselines: (1) general machine learning (ML) methods: RNN [Cho et al., 2014] and Transformer [Vaswani et al., 2017]; (2) ML models specifically designed for clinical predictive modeling: RETAIN [Choi et al., 2016a], GRASP [Zhang et al., 2021], GenHPF [Hur et al., 2024]; (3) Foundation models that alleviate the need for manual preprocessing: REMed [Kim et al., 2024]. The first two baseline groups rely on hand-crafted features and thus we follow existing works [Zhang et al., 2021, Jiang et al., 2024] and only input the procedure, lab, and prescription features; and truncate event sequence to 512.

**Results.** It is important to note that these clinical predictive tasks are quite different from the instruction-following tasks. Thus, we perform an additional supervised fine-tuning step for L1emr. A new classification head (i.e., linear layer) is added on top of L1emr and trained for each clinical predictive task.

The results on clinical predictive benchmarks can be found in Table 5. Both L1emr and the baseline methods are supervised trained for the benchmark tasks. We select the best model weights on the validation set and report the performance on the test set. First, we can see that baselines relied on manual data engineering perform quite well, reaching 0.80 AUC-ROC score for mortality prediction and diagnosis classification. Among them, RETAIN [Choi et al., 2016a], GRASP [Zhang et al., 2021], and GenHPF [Hur et al., 2024] perform slightly better by injecting various domain-specific inductive biases. Interestingly, despite being free from feature engineering, REMed [Kim et al., 2024] performs even better than many other baselines. This is probably because REMed [Kim et al., 2024] is able to take a broader range of events as input and utilize them as additional information. Lastly, we can see that the fine-tuned L1emr can perform better or on par with SOTA methods. This demonstrates the flexibility and adaptability of L1emr.

## 7 Conclusion

The remarkable abilities of LLMs to understand complex inputs and follow instructions for diverse tasks suggest their potential to simplify and enhance the analysis of EHRs. However, developing a conversational AI assistant for EHR data is difficult due to the following challenges: (1) the lack of large-scale instruction tuning data and (2) the limitation of model architectures in handling EHR data with complex schemas. In this paper, we introduce MIMIC-InstR, a dataset of over 400K open-ended instruction-tuning examples generated by GPT-3.5. This dataset covers a broad range of topics and can be used to instruction-tune general-purpose LLMs. Additionally, we propose L1emr, a generic framework designed to empower LLMs to encode EHR data with heterogeneous schema. Evaluation results show that L1emr exhibits excellent capabilities in answering diverse inquiries about a patient and performs on par with SOTA baselines when further fine-tuned for clinical predictive tasks.

## References

- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.
- Bert Arnrich, Edward Choi, Jason A Fries, Matthew B A McDermott, Jungwoo Oh, Tom J Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. MEDICAL EVENT DATA STANDARD (MEDS): FACILITATING MACHINE LEARNING FOR HEALTH.
- Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014.
- Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*, page 3512–3520, Red Hook, NY, USA, 2016a. Curran Associates Inc. ISBN 9781510838819.
- Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 08 2016b. ISSN 1067-5027. doi: 10.1093/jamia/ocw112. URL <https://doi.org/10.1093/jamia/ocw112>.
- Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, page 787–795, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098126. URL <https://doi.org/10.1145/3097983.3098126>.
- Cognitive. Dolphin-2.0-mistral-7b. <https://huggingface.co/cognitivecomputations/Dolphin-2.0-Mistral-7b>, 2023.
- P. F. DeChant, A. Acs, K. B. Rhee, T. S. Boulanger, J. L. Snowdon, M. A. Tutty, C. A. Sinsky, and K. J. Thomas Craig. Effect of organization-directed workplace interventions on physician burnout: A systematic review. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, 3(4):384–408, 2019. doi: 10.1016/j.mayocpiqo.2019.07.006.
- R. S. Evans. Electronic Health Records: Then, Now, and in the Future. *Yearbook of Medical Informatics*, 25(S 01):S48–S61, March 2018. ISSN 0943-4747. doi: 10.15265/IYS-2016-s006. URL <http://www.thieme-connect.com/products/ejournals/abstract/10.15265/IYS-2016-s006>. Publisher: Georg Thieme Verlag KG.
- Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju,

- Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i20.30205. URL <https://ojs.aaai.org/index.php/AAAI/article/view/30205>.
- A. Gamal, S. Barakat, and A. Rezk. Standardized electronic health record data modeling and persistence: A comparative review. *Journal of Biomedical Informatics*, 114:103670, 2021. doi: 10.1016/j.jbi.2020.103670.
- Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL <http://www.nature.com/articles/s41597-019-0103-9>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.
- Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyou Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. GenHPF: General Healthcare Predictive Framework for Multi-Task Multi-Source Learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, January 2024. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2023.3327951. URL <https://ieeexplore.ieee.org/document/10298642/>.
- Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. GRAPHCARE: ENHANCING HEALTH-CARE PREDICTIONS WITH PERSONALIZED KNOWLEDGE GRAPHS. 2024.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL <https://www.nature.com/articles/s41597-022-01899-x>.
- Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, and Roxana Daneshjou. CRAFT-MD: A Conversational Evaluation Framework for Comprehensive Assessment of Clinical LLMs.
- Junu Kim, Chaeun Shim, Bosco Seong Kyu Yang, Chami Im, Sung Yoon Lim, Han-Gil Jeong, and Edward Choi. General-Purpose Retrieval-Enhanced Medical Prediction Model Using Near-Infinite History, March 2024. URL <http://arxiv.org/abs/2310.20204>. arXiv:2310.20204 [cs].
- Sunjun Kweon, Jiyou Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Seunghyun Won, and Edward Choi. EHRNoteQA: A Patient-Specific Question Answering Benchmark for Evaluating Large Language Models in Clinical Settings, February 2024. URL <http://arxiv.org/abs/2402.16040>. arXiv:2402.16040 [cs].
- Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. EHRSQL: A Practical Text-to-SQL Benchmark for Electronic Health Records, December 2023. URL <http://arxiv.org/abs/2301.07695>. arXiv:2301.07695 [cs].
- Eric Lehman, Vladislav Lialin, Katelyn Y. Legaspi, Anne Janelle R. Sy, Patricia Therese S. Pile, Nicole Rose I. Alberto, Richard Raymond R. Ragasa, Corinna Victoria M. Puyat, Isabelle Rose I. Alberto, Pia Gabrielle I. Alfonso, Marianne Taliño, Dana Moukheiber, Byron C. Wallace, Anna Rumshisky, Jenifer J. Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. Learning to ask like a physician, 2022.

- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day, June 2023a. URL <http://arxiv.org/abs/2306.00890>. arXiv:2306.00890 [cs].
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, February 2022. URL <http://arxiv.org/abs/2201.12086>. arXiv:2201.12086 [cs].
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023b. URL <http://arxiv.org/abs/2301.12597>. arXiv:2301.12597 [cs].
- Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-62922-y. URL <http://www.nature.com/articles/s41598-020-62922-y>.
- Wing Lian, Bley Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and Teknium. Mistralorca: Mistral-7b model instruct-tuned on filtered openorca v1 gpt4 dataset. <https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca>, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023. URL <http://arxiv.org/abs/2304.08485>. arXiv:2304.08485 [cs].
- Edward R. Melnick, Liselotte N. Dyrbye, Christine A. Sinsky, Mickey Trockel, Colin P. West, Laurence Nedelec, Michael A. Tutty, and Tait Shanafelt. The association between perceived electronic health record usability and professional burnout among us physicians. *Mayo Clinic Proceedings*, 95(3):476–487, 2020. ISSN 0025-6196. doi: <https://doi.org/10.1016/j.mayocp.2019.09.024>. URL <https://www.sciencedirect.com/science/article/pii/S0025619619308365>.
- MosaicML. Introducing mpt-7b: A new standard for open-source, commercially usable llms. <https://mosaicml.com>, 2023. Accessed: 2023-05-05.
- J. Marc Overhage and Daniel Jr. McCallie. Physician time spent using the electronic health record during outpatient encounters: A descriptive study. *Annals of Internal Medicine*, 172(3):169–174, 2020. doi: 10.7326/M18-3684. URL <https://doi.org/10.7326/M18-3684>. Erratum in: *Ann Intern Med*. 2020 Oct 6;173(7):596.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrQA: A Large Corpus for Question Answering on Electronic Medical Records, September 2018. URL <http://arxiv.org/abs/1809.00732>. arXiv:1809.00732 [cs].
- Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D. Wang. EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records, February 2024. URL <http://arxiv.org/abs/2401.07128>. arXiv:2401.07128 [cs].
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large Language Models Encode Clinical Knowledge, December 2022. URL <http://arxiv.org/abs/2212.13138>. arXiv:2212.13138 [cs].
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguerre y

- Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards Expert-Level Medical Question Answering with Large Language Models, May 2023. URL <http://arxiv.org/abs/2305.09617>. arXiv:2305.09617 [cs].
- T. Tajirian, V. Stergiopoulos, G. Strudwick, L. Sequeira, M. Sanches, J. Kemp, K. Ramamoorthi, T. Zhang, and D. Jankowicz. The influence of electronic health record use on physician burnout: Cross-sectional survey. *Journal of Medical Internet Research*, 22(7):e19274, 2020. doi: 10.2196/19274. URL <https://www.jmir.org/2020/7/e19274>.
- Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3970–3980, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539089. URL <https://doi.org/10.1145/3534678.3539089>.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning, November 2023. URL <http://arxiv.org/abs/2311.10537>. arXiv:2311.10537 [cs].
- Migel Tissera. Synthia-7b-v1.3: Synthetic intelligent agent. <https://huggingface.co/migtissera/Synthia-13B>, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. BioInstruct: Instruction Tuning of Large Language Models for Biomedical Natural Language Processing, November 2023. URL <http://arxiv.org/abs/2310.19975>. arXiv:2310.19975 [cs].
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agueria y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards Generalist Biomedical AI, July 2023. URL <http://arxiv.org/abs/2307.14334>. arXiv:2307.14334 [cs].
- Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards Conversational Diagnostic AI, January 2024. URL <http://arxiv.org/abs/2401.05654>. arXiv:2401.05654 [cs].
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, 7(1):16, January 2024. ISSN 2398-6352. doi: 10.1038/s41746-023-00989-3. URL <https://www.nature.com/articles/s41746-023-00989-3>.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a. URL <https://openreview.net/forum?id=gEZrGCozdqR>.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022b. URL [https://openreview.net/forum?id=\\_VjQ1MeSB\\_J](https://openreview.net/forum?id=_VjQ1MeSB_J).
- Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May D. Wang, Joyce C. Ho, and Carl Yang. RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records, February 2024. URL <http://arxiv.org/abs/2403.00815>. arXiv:2403.00815 [cs, q-bio].
- Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering, 2021.
- Cyril Zakka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, Akshay Chaudhari, and William Hiesinger. Almanac Copilot: Towards Autonomous Electronic Health Record Navigation, May 2024. URL <http://arxiv.org/abs/2405.07896>. arXiv:2405.07896 [cs].
- Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: Generic framework for health status representation learning based on incorporating knowledge from similar patients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):715–723, May 2021. doi: 10.1609/aaai.v35i1.16152. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16152>.
- Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. AlpaCare: Instruction-tuned Large Language Models for Medical Application, May 2024. URL <http://arxiv.org/abs/2310.14558>. arXiv:2310.14558 [cs].

## Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]** See Section 1.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
  - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[Yes]**
  - (b) Did you describe the limitations of your work? **[Yes]**
  - (c) Did you discuss any potential negative societal impacts of your work? **[Yes]**
  - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]**
2. If you are including theoretical results...
  - (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
  - (b) Did you include complete proofs of all theoretical results? **[N/A]**
3. If you ran experiments (e.g. for benchmarks)...
  - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]**
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[Yes]**
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[Yes]**
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
  - (a) If your work uses existing assets, did you cite the creators? **[Yes]**
  - (b) Did you mention the license of the assets? **[Yes]**
  - (c) Did you include any new assets either in the supplemental material or as a URL? **[No]**
  - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[Yes]**
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[Yes]**
5. If you used crowdsourcing or conducted research with human subjects...
  - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[N/A]**
  - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]**
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[N/A]**

## Contents of Appendix

---

<b>A Datasheet for MIMIC-Instr</b>	<b>17</b>
A.1 Motivation . . . . .	17
A.2 Composition . . . . .	17
A.3 Collection process . . . . .	18
A.4 Preprocessing/cleaning/labeling . . . . .	19
A.5 Uses . . . . .	19
A.6 Distribution . . . . .	19
A.7 Maintenance . . . . .	19
<b>B Use of OpenAI API</b>	<b>20</b>
<b>C Additional Information on the MIMIC-Instr</b>	<b>20</b>
C.1 Questions Templates . . . . .	20
C.2 Prompts . . . . .	22
C.3 MIMIC-IV Preparation . . . . .	22
C.4 Examples . . . . .	24
<b>D Additional Experimental Setup for Conversational AI Assistant</b>	<b>26</b>
<b>E Additional Experimental Setup for Clinical Predictive Benchmarks</b>	<b>26</b>
<b>F Limitations and Broader Impacts</b>	<b>27</b>

---



## A Datasheet for MIMIC-Instr

### A.1 Motivation

- **For what purpose was the dataset created?**

We created MIMIC-Instr to enable instruction tuning LLMs to understand EHR data. The end goal is to develop conversational AI assistants to help physicians in information extraction and clinical reasoning on the EHR data.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The MIMIC-Instr dataset was created by the authors of this paper.

- **Who funded the creation of the dataset?**

To be added upon publication.

### A.2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

MIMIC-Instr contains natural language instructions (questions) and responses (answers). Each QA pair is matched to a patient record from the publicly available MIMIC-IV [Johnson et al., 2023] database.

- **How many instances are there in total (of each type, if appropriate)?**

There are 353,448 QA pairs in the schema alignment subset and 50,188 QA pairs in the clinical reasoning subset. Together, there are 403,636 QA pairs.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

All instances were provided.

- **What data does each instance consist of?**

Each instance consists of a subject id (integer), a hospital admission id (integer), a question (str), and an answer (str). The subject id and hospital admission id can be used to locate the corresponding patient record in the MIMIC-IV [Johnson et al., 2023] database.

- **Is there a label or target associated with each instance?**

The answer acts as the label.

- **Is any information missing from individual instances?**

No.

- **Is any information missing from individual instances?**

No.

- **Are there recommended data splits (e.g., training, development/validation, testing)?**

Yes. The data splits will be provided.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

The schema alignment subset was generated based on hand-crafted templates and then paraphrased by GPT-3.5. The clinical reasoning subset was directly generated by GPT-3.5. The questions and results can contain errors, noise, and redundancies.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

The dataset needs to be used with the publicly available MIMIC-IV [Johnson et al., 2023] database.

- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

No. The dataset was constructed based on the MIMIC-IV [Johnson et al., 2023] database which is de-identified. We used Azure's HIPAA-compliant OpenAI API in accordance with PhysioNet's regulations.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**  
No.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?**  
Yes. Some generated questions may ask about patient’s age, gender, and ethnicity. These demographic information were provided in the MIMIC-IV database and went through the de-identification process to avoid identifying individual patients.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**  
No.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**  
No. The source datasets are already de-identified.

### A.3 Collection process

- **How was the data associated with each instance acquired?**  
The schema alignment subset was generated based on template and then paraphrased by GPT-3.5. The clinical reasoning subset was directly generated by GPT-3.5.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**  
The question templates were designed manually using EHRSQL [Lee et al., 2023] as a reference. The GPT-3.5 API was from Azure’s HIPAA-compliant platform.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**  
Not applicable.
- **Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?**  
The MIMIC-Instr dataset was created by the authors of this paper. No crowd workers were involved.
- **Over what timeframe was the data collected?**  
The MIMIC-IV [Johnson et al., 2023] data was collected between 2008 and 2019. The MIMIC-Instr dataset was built upon MIMIC-IV in 2024.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?**  
Not applicable.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**  
The MIMIC-Instr dataset was built upon the MIMIC-IV [Johnson et al., 2023] database.
- **Were the individuals in question notified about the data collection?**  
Not applicable.
- **Did the individuals in question consent to the collection and use of their data?**  
Not applicable.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**  
Not applicable.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**  
Not applicable.

#### A.4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**  
Not applicable.
- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**  
Not applicable.
- **Is the software that was used to preprocess/clean/label the data available?**  
Yes. We utilized Python, Jupyter Notebooks, and OpenAI API.

#### A.5 Uses

- **Has the dataset been used for any tasks already?**  
Currently, the dataset is used for instruction tuning LLMs to understand EHR data.
- **Is there a repository that links to any or all papers or systems that use the dataset?**  
No.
- **What (other) tasks could the dataset be used for?**  
Question answering based on EHR data.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**  
This dataset was mainly generated with templates or by GPT-3.5. So it could contain errors and noise.
- **Are there tasks for which the dataset should not be used?**  
Not applicable.

#### A.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**  
Yes.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**  
The dataset will be released via PhysioNet upon publication.
- **When will the dataset be distributed?**  
Upon publication.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**  
The dataset will be distributed under the PhysioNet Credentialed Health Data License: <https://www.physionet.org/content/ehr-ds-qa/view-license/1.0.0/>.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**  
MIMIC-InstR needs to be used with the MIMIC-IV [Johnson et al., 2023] database which is under the PhysioNet Credentialed Health Data License: <https://www.physionet.org/content/ehr-ds-qa/view-license/1.0.0/>.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**  
No.

#### A.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**  
The authors of this paper.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**  
Contact the first/corresponding authors via email (to be added upon publication) or raise GitHub issue.
- **Is there an erratum?**  
No.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**  
Yes, we plan to update the datasets as we further improve the templates and OpenAI updates its API.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?**  
No.
- **Will older versions of the dataset continue to be supported/hosted/maintained?**  
Yes. This dataset has very simple schema, so the older versions can be easily supported.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**  
Contact the first/corresponding authors via email (to be added upon publication) or raise GitHub issue.

## B Use of OpenAI API

We used Azure’s HIPAA-compliant platform in accordance with PhysioNet’s regulations. We used the “gpt-35-turbo (0125)” version of GPT-3.5 and the “gpt-4 (0125-Preview)” version of GPT-4.

## C Additional Information on the MIMIC-Instr

### C.1 Questions Templates

We provide all the question templates used in generating the schema alignment subset in Table 6

Table 6: Questions templates used for generating the schema alignment subset of MIMIC-Instr. The definition of the {time\_period} keyword can be found in Table 7.

Table	Template
patients & admissions	What was the gender of the patient? What was the age of the patient? What was the race of the patient? What was the insurance of the patient? What was the marital status of the patient? What was the admission type of the patient? What was the admission location of the patient? What was the chief complaint of the patient?
diagnoses_icd	What were the billed diagnoses of the patient? What were the top five billed diagnoses of the patient?
labevents	What was the {measurement_name} at the {timestamp} hour? Was the {measurement_name} at the {timestamp} hour normal? What {category} measurements were performed on the {fluid} specimen at the {timestamp} hour? What {category} measurements on the {fluid} specimen were abnormal at the {timestamp} hour? What was the {first/last} {measurement_name} {time_period}?

	<p>When was the {first/last} {measurement_name} {time_period}?</p> <p>How many times did the patient have the {measurement_name} {time_period}?</p> <p>What was the {maximum, minimum, average} {measurement_name} {time_period}?</p> <p>Did the patient have any {measurement_name} {time_period}?</p>
microbiologyevents	<p>What microbiology tests were performed on the {spec_type_desc} specimen at the {timestamp} hour?</p> <p>What organisms were found on the {spec_type_desc} specimen at the {timestamp} hour?</p> <p>What were the antibiotics test results against the {org_name} on the {spec_type_desc} specimen at the {timestamp} hour?</p> <p>Did the patient have any microbiology test on the {spec_type_desc} specimen {time_period}?</p>
prescription	<p>What was the composition of the prescribed {drug} at the {timestamp} hour?</p> <p>What was the dose of the prescribed {drug} at the {timestamp} hour?</p> <p>What was the administration route of the prescribed {drug} at the {timestamp} hour?</p> <p>What was the administration duration of the prescribed {drug} at the {timestamp} hour?</p> <p>What drugs were prescribed at the {timestamp} hour?</p> <p>What was the composition of the {first/last} prescribed {drug} {time_period}?</p> <p>What was the dose of the {first/last} prescribed {drug} {time_period}?</p> <p>What was the administration route of the {first/last} prescribed {drug} {time_period}?</p> <p>What was the administration duration of the {first/last} prescribed {drug} {time_period}?</p> <p>When was the {first/last} {drug} prescription {time_period}?</p> <p>How many times did the patient have the {drug} prescription {time_period}?</p> <p>Was the patient prescribed with any {drug}?</p>
transfers	<p>Which unit was the patient transferred to at the {timestamp} hour?</p> <p>When was the patient discharged from the hospital?</p> <p>How long was the length of hospital stay of the patient in hours?</p> <p>How long was the length of hospital stay of the patient in days?</p>
inpuvents	<p>What was the amount of the IV administration {label} at the {timestamp} hour?</p> <p>What was the duration of IV administration {label} at the {timestamp} hour?</p> <p>What drugs were administered through IV at the {timestamp} hour?</p> <p>What was the amount of the {first/last} IV administration {label} {time_period}?</p> <p>What was the duration of the {first/last} IV administration {label} {time_period}?</p> <p>When was the {first/last} {label} IV administration {time_period}?</p> <p>How many times did the patient have the {label} IV administration {time_period}?</p> <p>Was the patient administered with any {label} through IV {time_period}?</p>
outputevents	<p>What was the amount of the output {label} at the {timestamp} hour?</p> <p>What was the amount of the {first/last} output {label} {time_period}?</p> <p>What was the total amount of the output {label} {time_period}?</p> <p>What was the {maximum, minimum, average} amount of the output {label} {time_period}?</p> <p>When was the {first/last} output {label} {time_period}?</p>

	How many times did the patient have the {label} output {time_period}? Did the patient have any {label} output {time_period}?
procedureevents	What procedures were performed at the {timestamp} hour? What procedures were performed {time_period}? What was the duration of the {first/last} {label} procedure {time_period}? When was the {first/last} {label} procedure {time_period}? How many times did the patient undergo the {label} procedure {time_period}? Did the patient undergo any {label} procedure {time_period}?

Table 7: Definition of the {time\_period} keyword.

Keyword	Realization
{time_period}	during the first 12 hours during the first 24 hours during the first 48 hours during the last 12 hours during the last 24 hours during the last 48 hours during day day during the entire stay

## C.2 Prompts

The prompts used for paraphrasing and generating the QA pairs can be found in Prompts 1 and 2.

**Prompt 1** Prompting GPT-3.5 to paraphrase the QA pairs generated with templates.

You are an AI assistant with expertise in medical knowledge.

Your input consists of a question-answer pair created using predefined rules.

Your primary task is to rephrase both the question and the answer to introduce variety in the wording while preserving their original meanings.

Objective:

1. Paraphrase both the question and the answer.
2. Ensure the paraphrased text is grammatically correct.
3. Adjust capitalization as needed
4. Maintain the original intent and meaning of the question-answer pair.
5. Format your response as follows:  
- Question: [Your paraphrased question]  
- Answer: [Your paraphrased answer]
6. Aim for brevity in both the question and answer.

Question: {input question}

Answer: {input question}

## C.3 MIMIC-IV Preparation

We construct our cohort using the ICU patients from the MIMIC-IV [Johnson et al., 2023] database. This database was constructed from the patients admitted to the ICU in the Beth Israel Deaconess Medical Center. This database covers 50920 patients with 66239 hospital admissions and 73181 ICU stays. We filter out patients without discharge summaries (909 admissions

**Prompt 2** Prompting GPT-3.5 to generate QA pairs based on the corresponding discharge summaries. You are an AI assistant specialized in analyzing ICU patients' data.

You are provided with a discharge summary of an ICU patient, which summarizes important clinical records and serves as an essential reference for the doctor's clinical decision-making.

Your task is to generate a question-answer pair inquiring about the patient.

Objective:

1. Formulate one question that a doctor will ask based on the provided discharge summary.
2. The answer should be found within the provided discharge summary.
3. Refrain from formulating questions that can be answered without referring to the provided discharge summary.
4. Avoid questions that include sensitive personal information or "\_\_\_".
5. Do not create questions that are too easy to answer. To answer your question, someone should have the clinical expertise equivalent to a doctor and must fully understand all provided discharge summaries.
6. Arrange your output in the following format:  
 - Question: [Your Question]  
 - Answer: [Your Answer]
7. Keep both the question and answer concise (within 256 tokens).

{input discharge summary}

removed), with more than two ICU stays per hospital admission (5,762 admissions removed), less than 18 years old (no admission removed), and with negative ICU/hospital length-of-stay (55 admissions removed). We then select the following tables from MIMIC-IV: hosp/patients, hosp/admissions, hosp/diagnosis, hosp/labevents, hosp/microbiologyevents, hosp/prescriptions, hosp/transfers, icu/icustays, icu/inputevents, icu/outputevents, icu/procedureevents. Note that icu/chartevents table is excluded due to two reasons: it contains bedside monitor data, which is very dense and should be better treated as timeseries instead of event sequence, and it has large overlapping information with other tables like hosp/labevents [Johnson et al., 2023]. Lastly, we remove patients without any generated QA, i.e. rejected by GPT-3.5 (2,987 admissions removed) and with more than 4,000 events (680 admissions removed). In the end, we have a total of 55846 admissions. We hold out 10% each for validation (5579) and testing (5585). The average number of events per patient admission is  $559 \pm 543$ .

We covert each event into text using the templates in Table 8.

Table 8: Templates used in converting event to text.

Table	Template
patients & admissions	gender: {gender}, age: {age}, race: {race}, marital status: {marital_status}, insurance: {insurance}, type: {admission_type}, location: {admission_location}, chief complaint: {chief_complaint}
diagnoses_icd	Billed diagnosis: {long_title}
labevents	{fluid} {label} {category}: {value} {normal/abnormal}
microbiologyevents	{test_name} on {spec_type_desc}, organism grew: {org_name}, antibiotic tested: {ab_name}, antibiotic sensitivity: {interpretation}, comments: {comments}
prescriptions	{drug} ({prod_strength}), prescribed dose: {dose_val_rx} {dose_unit_rx}, route: {route}, duration: {duration} hour
transfers	{eventtype} to {careunit}
inputevents	{label} {amount} {amountuom} for {duration} hour
outputevents	{label} {value} {valueuom}
procedureevents	{label} for {duration} hour

## C.4 Examples

Table 9: Randomly selected examples from the MIMIC-Insttr dataset.

Question	Answer
<b>Schema Alignment</b>	
What was the blood gas measurement obtained through intubation at 1.55 hours?	Intubated.
How was the Calcium Gluconate administered at 1.20 hour?	Intravenously (IV)
Was there any value in using an Assistance Device for the patient on day 4?	None.
What type of admission did the patient have?	Urgent
Did the patient receive any IV administration of NaCl 0.9% within the initial 48 hours?	Affirmative.
What type of insurance did the patient have?	The patient had Medicare coverage.
Did the patient receive any IV administration of NaCl 0.9% on day 13?	Negative.
What was the timing of the initial blood PT Hematology measurement within the past 48 hours?	The measurement was taken at 38.85 hours.
What was the highest volume of fluid drained by Chest Tube #1 throughout the patient's stay?	The maximum amount of fluid drained by Chest Tube #1 was 40.00 ml.
How much fluid did Chest Tube #1 produce in the initial 12-hour period?	The output from Chest Tube #1 was 80.00 ml.
How much Foley output was typically observed on the second day?	The average amount of Foley output on day 2 was 22.50 ml.
How old was the individual?	63 years old.
How much NaCl 0.9% was administered intravenously at 63.17 hours?	The IV administration of NaCl 0.9% was 100.00 ml.
How much was the lowest output from the Foley catheter in the past 48 hours?	The minimum output from the Foley catheter was 20.00 ml.
What was the primary reason for the patient's visit?	The patient experienced discomfort in the chest.
What was the blood bicarbonate chemistry measurement within the initial 24-hour period?	The blood bicarbonate chemistry measurement was 23.0 mEq/L.
What was the timing of the initial CT scan within the first 12 hours?	The CT scan was performed at 2.68 hours.
Where was the patient admitted?	The patient was admitted to the emergency room.
What was the sex of the individual?	Female
What was the timing of the most recent Ultrasound procedure on day 1?	It occurred at 4.68 hours.
<b>Clinical Reasoning</b>	
What was the reason for discontinuing anticoagulation therapy in the patient with pericardial tamponade post-AVR surgery?	Anticoagulation therapy was discontinued in the patient with pericardial tamponade post-AVR surgery due to a super-therapeutic INR of 6.8 and no clear indication for anticoagulation at that time.
What was the intervention performed during the endoscopy for the patient with a foreign body in the esophagus?	The foreign body of food was removed during the endoscopy, and no other intervention was done.
What medications were initiated for the patient in the ICU due to frequent irregular heart rate and atrial ectopies?	The patient was started on beta blocker and Amiodarone for frequent irregular heart rate and atrial ectopies while in the ICU.
What interventions were performed on the patient during the procedure in the Operating Room?	The patient underwent CABG x 4 (coronary artery bypass grafting) and AVR (aortic valve replacement) with Dr. [Doctor's Name].



What interventions were performed to manage the patient's ventricular tachycardia?	The patient's ventricular tachycardia was managed with Amiodarone boluses, electrolyte repletion, and defibrillation. EP recommended continuing Amiodarone for treatment.
What was the patient's hematocrit level upon admission to the Acute Care Trauma Surgery service in the intensive care unit?	The patient's hematocrit was 39.6 upon admission to the Acute Care Trauma Surgery service in the intensive care unit.
What was the reason for not starting the patient on coumadin despite having atrial fibrillation post-operatively?	The patient was not started on coumadin due to the limited time frame of atrial fibrillation.
What prompted the initiation of Zosyn and a bronchoscopy for the patient during his hospital stay?	The patient had multiple episodes of PO intolerance complicated by aspiration and desaturation, leading to the initiation of Zosyn and a bronchoscopy to further evaluate his airway.
What was the patient's initial presentation upon transfer from rehab that raised concern for aortic valve thrombosis?	The patient presented with dyspnea and weight gain, initially requiring BiPAP, which raised concern for aortic valve thrombosis.
What was the reason for the patient's emergent surgery in the operating room?	The patient underwent Emergent repair of type A aortic dissection with cardiac tamponade, ascending aorta and Hemiarch replacement with 28mm Gelweave graft under circulatory arrest.
What was the reason for the patient's admission to the ICU and what was the finding on EGD?	The patient was admitted to the ICU due to difficulty managing secretions, and EGD revealed a piece of chicken impacted in the mid esophagus that was advanced into the stomach.
What was the reason for holding the patient's Plavix for 7 days according to Neurosurgery?	Neurosurgery recommended holding her Plavix for 7 days due to stable SAH and no need for seizure prophylaxis.
What was the reason for the patient's transfer to the ICU from the rehabilitation facility?	The patient was transferred to the ICU from the rehabilitation facility for Tylenol overdose in fulminant liver failure most likely due to Tylenol overdose with a Tylenol level of 154 at 24 hours from ingestion event.
What was the reason for consulting Ophthalmology for the patient on post operative day 1?	Ophthalmology was consulted for the patient's complaints of bilateral floaters to assess for signs of hemorrhage or neovascularization.
What interventions were implemented to manage the patient's post-operative Atrial Fibrillation?	The patient was started on Amiodarone and Coumadin after experiencing several hours of post-operative Atrial Fibrillation on POD 4.
What specialties evaluated the patient upon arrival and found no acute interventions necessary?	Plastics and ophthalmology evaluated the patient upon arrival and found no acute interventions necessary.
What was the reason for the patient's elevated white blood cell count post-operatively?	The elevated white blood cell count post-operatively was thought to be due to the patient's decadron.
What was the reason for holding certain home medications, including methadone, for the patient during their ICU course?	The patient's home medications that can lead to sedation, including methadone, were held due to findings of generalized spike wave and left hemisphere discharges on EEG, as well as an event with right gaze deviation and body stiffening, unresponsiveness.
What was the patient's LVEF and LV systolic function according to the TTE results?	LVEF 47%, moderate regional LV systolic dysfunction in setting of overall mild global LV systolic dysfunction c/w prior myocardial infarction in the mid LAD territory.

What was the patient’s neurological status on the day of discharge to rehabilitation?	On [date of discharge], the patient was neurologically intact, afebrile, ambulating with assistance, tolerating a diet, voiding and stooling without difficulty, and his pain was well controlled with oral pain medications.
---	---

## D Additional Experimental Setup for Conversational AI Assistant

Prompts for instructing GPT-3.5 to evaluate the generated response can be found in Prompt 3. This prompt was adapted from the one used in LLaVA-Med [Li et al., 2023a].

---

**Prompt 3** Prompting GPT-3.5 to evaluate the generated response.

---

You are a helpful and precise assistant for evaluating the quality of responses.

Please assess the performance of two clinical AI assistants based on the question and the ground-truth answer provided below.

Your evaluation should consider helpfulness, relevance, accuracy, and level of detail.

Rate each AI assistant’s response with a single score on a scale of 1 to 10, where 10 represents excellent performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space.

In the subsequent line, provide a concise explanation of your evaluation.

Avoid any potential bias and ensure that the order in which the responses were presented does not affect your judgment.

Question  
 {input question}  
 End of Question

Ground-truth Answer  
 {input ground-truth answer}  
 End of Ground-truth Answer

Assistant 1 Answer  
 {input assistant 1 answer}  
 End of Assistant 1 Answer

Assistant 2 Answer  
 {input assistant 2 answer}  
 End of Assistant 2 Answer

---

## E Additional Experimental Setup for Clinical Predictive Benchmarks

We leverage the held-out test set of 5585 ICU admissions from the MIMIC-IV database. The training and validation set share the same cohorts as the ones used in instruction-tuning. Additional patient filtering is performed for each task introduced below.

**Mortality prediction** aims to predict whether the patient will pass away upon discharge using events from the first 48 hours of the hospital admission. Patients with hospital length-of-stay less than 48 hours are filtered. The prevalence rate is 0.10.

**Length-of-stay prediction** aims to determine whether the patient’s hospital stay will be longer than 7 days using the first 48 hours of the hospital admission. Patients with hospital length-of-stay less than 48 hours are filtered. The prevalence rate is 0.18.

**Readmission prediction** aims to predict whether the patient will be readmitted back to the hospital within 14 days following current discharge using all events from the current admission. Patients who are deceased in the current hospital admission are filtered for this task. The prevalence rate is 0.11.

**Diagnosis classification** aims to classify which acute care conditions are present using all events from the current admission. We follow existing works [Harutyunyan et al., 2019] and define disease labels with 25 conditions that are common in adult ICUs, including 12 critical conditions, (e.g., respiratory failure; 8 chronic conditions (e.g., diabetes); and 5 mixed (i.e., recurring or chronic with periodic acute episodes) conditions (e.g., cardiac dysrhythmias). The complete list of disease labels and prevalence rates can be found in Table 10.

Table 10: List of all 25 disease labels.

Disease	Category	Prevalence
Septicemia	acute	0.15
Fluid and electrolyte disorders	acute	0.38
Acute myocardial infarction	acute	0.08
Congestive heart failure	acute	0.25
Acute cerebrovascular disease	acute	0.09
Pneumonia	acute	0.13
Pleurisy; pneumothorax; pulmonary collapse	acute	0.09
Respiratory failure	acute	0.24
Other lower respiratory disease	acute	0.12
Other upper respiratory disease	acute	0.05
Other liver diseases	acute	0.13
Gastrointestinal hemorrhage	acute	0.06
Acute and unspecified renal failure	acute	0.27
Complications of surgical procedures	acute	0.21
Shock	acute	0.12
Diabetes mellitus without complication	chronic	0.18
Diabetes mellitus with complications	chronic	0.14
Disorders of lipid metabolism	chronic	0.41
Essential hypertension	chronic	0.41
Hypertension with complications	chronic	0.21
Coronary atherosclerosis and other heart disease	chronic	0.32
Conduction disorders	chronic	0.10
Cardiac dysrhythmias	chronic	0.36
Chronic obstructive pulmonary disease	chronic	0.14
Chronic kidney disease	chronic	0.20

## F Limitations and Broader Impacts

The MIMIC-Inst<sub>r</sub> dataset was generated based on hand-crafted templates and with the help of GPT-3.5. This means the dataset could (and probably would) contain error and noise. The inherent inaccuracies stemming from the data generation process could introduce biases or distort the models’ understanding of real-world clinical scenarios. Consequently, when applying these models in the real-world setting, one must carefully evaluate their performance and interpretability to ensure that they reliably support clinical decision-making. Furthermore, future work may focus on refining data generation methods to improve the quality and realism of training datasets, thus enhancing the models’ applicability and accuracy in practical healthcare applications.