DiaLLM: EHR-Enhanced Clinical Conversational System for Clinical Test Recommendation and Diagnosis Prediction

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have led to remarkable progresses in medical consultation. However, existing medical LLMs overlook the essential role of Electronic Health Records (EHR) and focus primarily on diagnosis recommendation, limiting their clinical applicability. We propose DiaLLM, the first medical LLM that integrates heterogeneous EHR data into clinically grounded dialogues, enabling clinical test recommendation, result interpretation, and diagnosis prediction to better align with real-world medical practice. To construct clinically grounded dialogues from EHR, we design a Clinical Test Reference (CTR) strategy that maps each clinical code to its corresponding description and classifies test results as "normal" or "abnormal". Additionally, DiaLLM employs a reinforcement learning framework for evidence acquisition and automated diagnosis. To handle the large action space, we introduce a reject sampling strategy to reduce redundancy and improve exploration efficiency. Furthermore, a confirmation reward and a class-sensitive diagnosis reward are designed to guide accurate diagnosis prediction. Extensive experimental results demonstrate that DiaLLM outperforms baselines in clinical test recommendation and diagnosis prediction. Our code is available for review[1].

1 Introduction

Rapid advancements in LLMs (Touvron et al., 2023; Wu et al., 2023) expanded opportunities to improve diagnostic assistance and patient interactions in the healthcare domain(Biswas, 2023; Li et al., 2023; Shah, 2024; Singhal et al., 2023), and clinical conversational systems (Wang et al., 2023a; Yang et al., 2024b) have emerged as a promising approach to enhance clinical reasoning and assist doctors with diagnosis. However, as shown in Table 1, existing studies primarily rely on synthetic dialogues generated from medical knowledge graph (Wang et al., 2023a) or open medical Question Answer (QA) (Labrak et al., 2024) and are difficult to work on real-world healthcare settings which use Electronic Health Records (EHR). Furthermore, the diagnosis workflow involves multiple critical sub-tasks, including the inquiry of clinical lab tests and the interpretation of their results before giving the diagnosis results, which are often neglected in current approaches (Zhou et al., 2023; Liu et al., 2024). Consequently, existing explorations in the clinical conversational system are still far from practical healthcare scenarios, primarily due to challenges in understanding EHR and adapting to real-world diagnosis workflows (Li et al., 2024).

Patient Name	Gender	Alert
Smith		Surfa-Drugs
Phone	Age	
365-5372-2186	27	

Date	Code System	Code	Results	Units
20190104	LOINC	48643-1	84.34	mL/min/{1.73_m2}
20190104	ICD-10	R10.12		
20190110	LOINC	48647-5	90.86	mL/min/{1.73_m2}
20190110	LOINC	8462-4	93	mm[HG]
20190115	LOINC	2160-0	0.73	mg/dL
20190115	LOINC	29463-7	175	lb_a
20190117	ICD-10	Q65.89		
20190117	ICD-10	M87.00		

Figure 1: Structured representation of EHR data, illustrating patient demographics, symptoms, clinical tests (e.g., lab test, vital signs), and physician decisions over time. Symptoms and diagnoses are encoded using ICD codes, while clinical tests follow the LOINC system.

How to learn from EHR data? EHR data is a comprehensive digital record encompassing a patient's medical history, treatments, test results, and clinical decisions. As shown in Figure 1, EHR captures multiple clinical visits and can be structured as a dialogue for clinical conversational systems. However, its heterogeneity and domain-specific nature pose a fundamental challenge, limiting compatibility with existing LLMs (Li et al., 2024). Unlike general NLP tasks, EHR data includes nu-

[1] https://anonymous.4open.science/r/DiaLLM-E3B3

Model	Main Data Source	Function	Training Method	Language
ChatDoctor (Li et al., 2023)	Medical Consultation Website	Diagnosis Assistance	SFT	English
DoctorGLM (Xiong et al., 2023)	Physician-Patient Dialogues	Diagnosis Assistance	SFT	Chinese
BenTsao (Wang et al., 2023a)	Medical knowledge Graphs	Diagnosis Assistance	SFT	Chinese
Zhongjing (Yang et al., 2024b)	Proprietary data, Crawled data	Diagnosis Assistance	SFT + RLHF	Chinese
Biomistral(Labrak et al., 2024)	Medical QA	Diagnosis Assistance	SFT	English
Meditron (Chen et al., 2024)	PubMed articles	Diagnosis Assistance	SFT	English
DiaLLM	EHR data	Diagnosis Assistance, Test Results Analysis, Clinical Test Ordering	PPO	English

Table 1: Comparison of Data Sources, Function, Training Method, and Language across Popular Medical Conversational Models.

merical values, categorical attributes from clinical tests, and domain-specific terminologies such as ICD codes and LOINC lab identifiers (Sui et al., 2024; Liang et al., 2024). These complexities require precise numerical reasoning and contextual understanding. While LLMs benefit from massive text-based data and are trained using next-token prediction (Zhao et al., 2023), they exhibit significant weaknesses in understanding numeric and specialized clinical knowledge (Sui et al., 2024).

Which Service should Clinical Conversational System Provide? The second challenge lies in the alignment between LLM-supported conversational system and clinical workflows. As shown in Table 1, existing works primarily focus on diagnosis assistance using "symptom-diagnosis" conversational data, which oversimplifies the diagnostic process and limits practical applicability. In a typical diagnostic scenario, a patient presents symptoms, prompting the clinician to iteratively gather information through inquiries and clinical lab tests. This process follows a cycle of 'evidence acquisition, results interpretation, and diagnosis confirmation', with the patient's medical trajectory information promptly recorded in the EHR data. Simulating clinicians' evidence acquisition and automated diagnosis process within LLMs still remains unexplored (Zhou et al., 2023).

Motivation of DiaLLM. In this work, we propose a novel conversational agent, named DiaLLM, which provides an EHR-grounded transformation pipeline and explicitly models clinicians' reasoning processes for evidence acquisition and automated diagnosis.

Technically, the EHR-grounded transformation aims to convert EHR data into dialogues that are aligned with common-sense knowledge and interpretable to LLMs. It first converts the EHR data into a single or multi-turn dialogue dataset, structured according to the patient's clinical visit timeline. Then, we translate the heterogeneous dialogue into clinically-grounded text by designing a Clinical Test Reference (CTR) strategy. The CTR facilitates (1) the translation of standardized clinical codes (such as ICD-9/10, LOINC) into clinically-grounded text; and (2) the interpretation of clinical test results, conditioned on the patient's age and gender.

On top of transformed EHR data, DiaLLM models evidence acquisition and automated diagnosis with a reinforcement learning framework. A policy network selects clinical tests (i.e., take actions) or terminates the process to make a diagnosis. Upon termination, a supervised classification model is invoked for disease diagnosis. To handle the large action space of clinical tests, we introduce a novel rejection sampling strategy (Bardenet et al., 2014; Fan et al., 2023; Mandel et al., 2016) that pre-filters redundant or unnecessary tests, ensuring only clinically relevant ones are selected. Meanwhile, since patients can have multiple diagnoses and disease distributions exhibit a long-tail pattern across populations, we propose a new confirmation reward and a class-sensitive classification reward to enhance diagnosis prediction.

Evaluation. We propose a comprehensive evaluation framework for medical LLMs, assessing both single-turn and multi-turn consultations. Experimental results show that DiaLLM outperforms both general-purpose and medical-specific LLMs in clinical test selection and diagnosis prediction. Our ablation study reveals that integrating the EHR-grounded transformation pipeline and specialized reward modeling significantly improves clinical test comprehensiveness, result interpretation, and early diagnosis accuracy.

2 Related Works

Representation Learning for EHR data. The majority of existing studies enhance medical code rep-

resentations by incorporating external relational information through medical ontologies (Choi et al., 2017; Panigutti et al., 2020) and knowledge graphs (Jiang et al., 2023b; Wang et al., 2023b). However, these established medical knowledge sources are restricted to specific diseases (Si et al., 2021). Recent works have attempted to derive clinical concept embeddings from large-scale medical text corpora (Ye et al., 2021), learning relational graphs through self-supervised learning (Yao et al., 2024), contrastive learning (Cai et al., 2022), or generating medical concepts directly using LLMs (Ma et al., 2024). However, these approaches often lack clinically grounded annotations, making the learned embeddings incompatible with LLM inputs.

Another line of research improves clinical prediction performance by learning from clinical test results. Existing works derive partition functions to learn from numerical and categorical clinical test results (La Cava et al., 2019), employing methods such as XGBoost (Chen and Guestrin, 2016), additive models (Hastie, 2017), piecewise linear functions (Montomoli et al., 2021), and logic-based rule learning (Ren et al., 2024). However, these approaches can be susceptible to biases introduced by patient population variations and may not align with established guidelines for interpreting test results (Ren et al., 2024).

Diagnosis-oriented Conversational System. Early research primarily focused on ICU-based temporal EHR data (Yoon et al., 2019; Fansi Tchango et al., 2022; Qin et al., 2024; He and Chen, 2022) or specific diagnosis categories (He and Chen, 2022; Fansi Tchango et al., 2022), and proposed to model evidence inquiry and diagnosis as a Markov decision process (MDP) (Tang et al., 2016). With the rise of foundation models, clinical conversational systems have been explored by tuning on various medical corpora, including clinical conversations collected from online medical consultation website (Li et al., 2023), symptom-diagnosis dialogues (Toma et al., 2023), medical question-answering pairs (Han et al., 2023), and knowledge graph-generated dialogues (Yang et al., 2024b; Wang et al., 2023a). However, these curated datasets deviate from real-world data distribution, lack essential clinical test support and do not interpret lab test results. These shortcomings limit their real-world applicability. For a comprehensive review, see (Zhou et al., 2023). In contrast, our approach constructs single-turn and multi-turn dialogue data leveraging real-world EHR data, and facilitates lab test requesting and diagnosis prediction.

3 Methodology

3.1 Diagnostic Conversational System Setup

Task Definition and Notations. DiaLLM encapsules the evidence acquisition and diagnosis automation with the following steps:

1. **Initial Query:** The patient provides demographic information $\mathbf{d}$ and symptoms $\mathbf{s}$.
2. **Clinical Test Recommendation:** The LLM agent suggests initial clinical tests $\mathbf{c_0}$.
3. **Test Result Analysis and Follow-ups:** Upon receiving the patient's initial clinical test results $\mathbf{v_0}$, the agent conducts a clinically grounded analysis and iteratively recommends additional follow-up tests $\mathbf{c_{t>0}}$ and interprets new results $\mathbf{v_{t>0}}$ to gather further evidence. This process continues until a conclusive diagnosis is achieved.
4. **Diagnosis Prediction:** The system ultimately predicts the diagnosis $\mathbf{y}$.

Problem Formulation. This process can be modeled as a Markov Decision Process (MDP) $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma)$, where: $\mathcal{S} = \mathcal{S}' \cup \{\mathbf{s}_\perp\}$ is the state space, with $\mathbf{s}_\perp$ as the terminal state. $\mathcal{A} = \mathcal{A}' \cup \{\mathbf{a}_\perp\}$ is the action space, with $\mathbf{a}_\perp$ as the stop action. γ is the discount factor. Each dialogue consists of at most T turns, where T represents the maximum number of visits recorded in the EHR data. At turn t, the state $\mathbf{s}_t \in \mathcal{S}'$ encodes sociodemographics, acquired evidence and dialogue history $\mathbf{h}_t$:

$$\mathbf{s}_t = \{\mathbf{d}, \mathbf{s}, \mathbf{c_t}, \mathbf{v_t}, \mathbf{h}_t\}. \quad (1)$$

At turn t, the agent selects an action $\mathbf{a_t}$, determining whether to request further tests or stop. The clinical test recommendation model $\pi_\theta(\mathbf{a_t}|\mathbf{s_t})$ governs test selection, while the diagnosis prediction model $\pi_\phi(\mathbf{y}|\mathbf{s}_\perp)$ makes the final diagnosis.

Overview of DiaLLM. In this work, we present DiaLLM designed to enhance evidence acquisition, result interpretation, and the diagnostic workflow. It consists of two key components: (1) a transformation strategy that constructs dialogues from heterogeneous EHR data, aligning them to common-sense texts for the ease of understanding by LLMs; and (2) a reinforcement learning framework to empower the base LLM with evidence

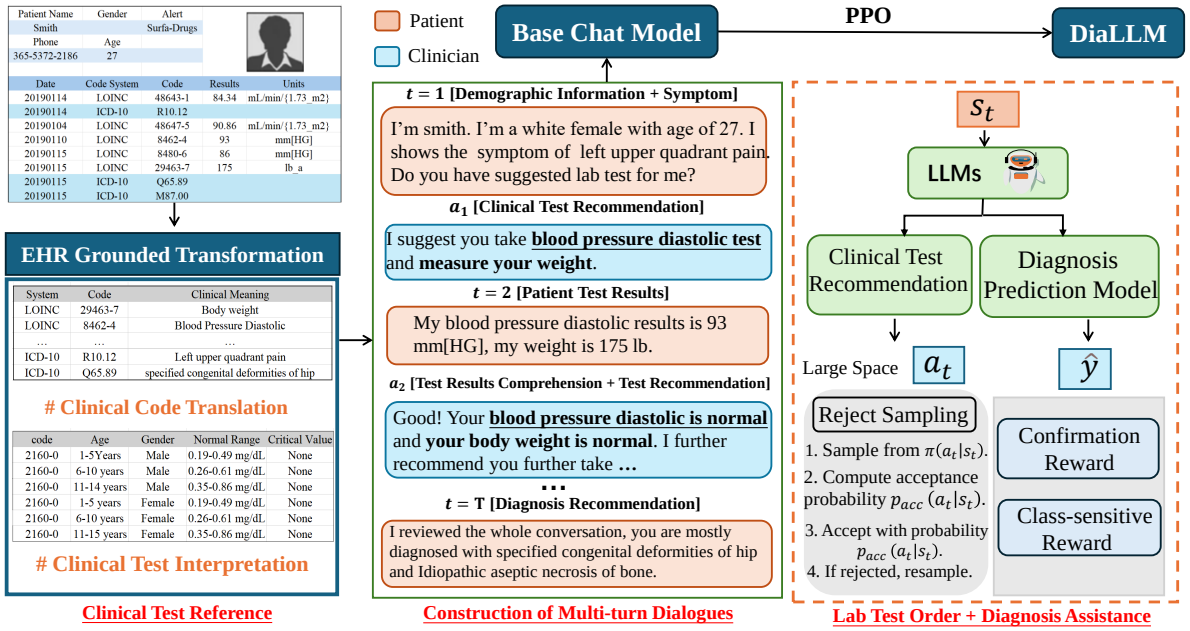


Figure 2: DiaLLM operates in two stages: (I) Dialogue Data Construction, where EHR data is transformed into clinically grounded dialogue data, and (II) Reward Modeling, which optimizes clinical test selection and diagnosis prediction. In Stage I, DiaLLM converts clinical test codes and results into texts using our constructed Clinical Test Reference. In Stage II, we adopt a novel rejection sampling strategy to handle the large action space and incorporate two reward signals to facilitate diagnosis prediction learning.

acquisition and automated diagnosis capabilities, for which we design a reject sampling strategy and several reward signals to facilitate learning. Both components will be detailed later in this section.

3.2 Dialogue Data Construction from EHR.

We introduce an EHR transformation strategy to convert structured EHR into clinically grounded dialogues. This process begins by segmenting patient visit records into conversational episodes when clinical visit intervals are within one week. A major challenge lies in representing medical terminologies (e.g., codes and lab test identifiers) and interpreting lab results. While text-driven LLMs excel in natural language understanding and reasoning, they lack direct exposure to structured EHR data. To address this, we manually curate a *Clinical Test Reference* data to translate heterogeneous EHR data in a clinically meaningful manner.

The transformation comprises two components: (1) Clinical Code Translation that converts specialized medical terms, including symptom codes $\mathbf{s}$, diagnosis codes $\mathbf{y}$, and clinical test codes $\mathbf{c_t}$ into clinically grounded and common-sense language that is easily understandable by LLMs. An example of the clinical code translation data is provided in Table 6 in the Appendix. (2) Clinical Test Interpretation that transforms heterogeneous lab test results using domain knowledge, including grounded reference ranges conditioned on gender and age, and classifies test results as 'normal' or 'abnormal'. An example of the clinical test interpretation data is shown in Table 7 in the Appendix.

Details of Clinical Test Reference database. For clinical code translation, symptom and diagnosis codes are commonly recorded using the ICD system, such as ICD-9 and ICD-10. To standardize code descriptions, we build a comprehensive codebase that includes ICD-9, ICD-10, and a mapping from ICD-9 to ICD-10 for standardization. Clinical tests are generally coded using the **LOINC system**. We annotate 735 clinical test code descriptions (including both lab test and vital sign codes) [2] which cover most of diseases to ensure comprehensive coverage and standardization.

For clinical test interpretation, we annotate 262 commonly used clinical tests based on established medical guidelines[3]. This resulted in 1,163 annota-

[2]LOINC Organization: https://loinc.org/

[3]LOINC Organization: https://loinc.org/, Mayo Clinic Laboratory: https://www.testcatalog.org/show/NAS, Corewell Health: https://corewellhealth.testcatalog.org/show/LAB299-1

tions [4] categorized as follows: (1) Normal Ranges that define reference ranges considering age and gender influences. (2) Critical Values that identifies thresholds indicating life-threatening conditions requiring immediate intervention. (3) Demographic Variability that includes age- and gender-specific ranges to account for physiological differences. (4) Units of Measurement that ensures consistency across international and clinical standards.

Example. Given the EHR example: "Female, Age 27, LOINC 2160 0.48 mg/dL," the CTR translates it to: "Given the patient demographic information (Age: 27, gender: Female). These lab tests show normal results: Creatinine in Serum or Plasma".

3.3 Learning for Lab Test and Diagnosis.

On top of transformed EHR data, DiaLLM models lab test acquisition and automated diagnosis using a reinforcement learning framework. At each step, the model will select follow-up lab tests (i.e., take actions) or termination of this process to make a diagnosis. To facilitate learning, we adopt rejection sampling to reduce the complexity. During fine-tuning with PPO, both a confirmation-based reward and a Class-sensitive Diagnosis reward are utilized.

Rejection Sampling for Lab Test Selection. In the presence of cost pressure, the goal of the agent is to balance timely and accurate evidence acquisition with cost-effective feature selection. The vast number of potential lab tests leads to a large action space, making direct reinforcement learning (RL) inefficient. To mitigate this, we introduce a strategy based on rejection sampling (Bardenet et al., 2014; Mandel et al., 2016) which pre-filters redundant or unnecessary lab tests before execution, ensuring that only clinically relevant tests are selected.

At each decision step t, the actor model $\pi(a_t|s_t)$ samples a candidate lab test c_t from the full action space. Instead of executing the sampled action immediately, we introduce a rejection sampling mechanism that determines whether the test should be conducted based on its informativeness, cost, and redundancy. The probability of accepting a test $p_{accept}(c_t|s_t)$ is defined as:

$$p_{\text{accept}}(c_t \mid s_t) = \frac{H(y \mid s_t) - H(y \mid s_t, v_t)}{\max_{c_t} H(y \mid s_t) - H(y \mid s_t, v_t)} \cdot 1[c_t \notin C_{\text{prev}}], \tag{2}$$

[4] Each clinical test may correspond to multiple reference ranges conditioned on patient age and gender.

where $H(y \mid s_t, v_t)$ denotes the entropy of predicted diagnosis given the current sate s_t and value v_t of candidate test c_t, capturing how much uncertainty is reduced by performing c_t. The redundancy filter $1[c_t \notin C_{\text{prev}}]$ prevents reordering previously conducted tests.

Confirmation Reward. As the conversation progresses, the agent should systematically acquire more evidence to refine its belief in the correct diagnosis while minimizing uncertainty. Inspired by potential-based reward shaping (Hu et al., 2020), we formalize the confirmation reward R_{co} as:

$$R_{\text{Co}}(s_t, a_t, s_{t+1}) = 1_{s_{t+1} \neq s_\perp} \cdot \left(CE(\hat{y}_{t+1}, y) - CE(\hat{y}_t, y)\right),$$

where $1_{s_{t+1} \neq s_\perp}$ is an indicator that the terminal state has not been reached yet.

Class-sensitive Diagnosis Reward. This reward is designed to provide feedback to the agent by evaluating the quality of its final predicted diagnosis with respect to the ground truth diagnosis y once the interaction process is completed. To directly optimize diagnosis prediction while addressing the issue of class imbalance (Puthiya Parambath et al., 2014; Elkan, 2001), we introduce a weighted classification reward:

$$R_{\text{Cl}}(s_T) = \sum_{y_i \in y} w_{\text{Cl}}(y_i) \cdot \text{CE}(\hat{y}_i, y_i), \tag{3}$$

where $w_{\text{Cl}}(y_i) = \frac{1}{p(y_i)}$ assigns a higher weight to rare diagnoses based on the inverse of the class frequency. This reward adjusts the importance of correctly predicting minority classes, ensuring that the agent focuses on both high-risk and low-risk diagnoses effectively.

4 Experiments

4.1 Dataset and Task Description

To critically evaluate the performance of LLMs on EHR data, we design the dataset with two key considerations: (I) *Unexposed*: Ensuring that the dataset has not been previously used by most LLMs. (II) *Predictive*: Focusing on diseases that can be predicted solely from EHR data, reflecting their clinical applicability. Detailed data statistical analysis is listed in Table 8 in Appendix A.2.1.

NHANES Dataset. Following (Kachuee et al., 2019), we construct a diabetes dataset from the public National Health and Nutrition Examination

Survey (NHANES) [5]. The NHANES dataset does not include clinical visit time information, and we construct a single-turn dialogue based on this data. The dataset includes 8,897 patients and 45 heterogeneous features, such as demographic data, lab results (e.g., total cholesterol, triglycerides), physical measurements (e.g., weight, height), and responses to questionnaires (e.g., smoking, alcohol consumption).

TriNetX Dataset. TriNetX[6] is a global health research network providing access to large-scale de-identified patient EHR data. TriNetX includes records from 33,105 de-identified patients, spanning data from 1982 to 2023, collected across more than 100 community hospitals and 500 outpatient clinics. To satisfy goal (I) and (II), we extract data on metabolic, respiratory, and circulatory diseases, referred to as TriNetX-Metabolic, TriNetX-Respiratory and TriNetX-Circulatory, respectively.

4.2 Baselines and Evaluation Metrics

We compare DiaLLM with (1) pretrained foundation models[7] including `Mistralv0.3-7B Instruct` (Jiang et al., 2023a), `Llama3.1-8B Instruct` (Dubey et al., 2024), `Qwen2.5-7B Instruct` (Yang et al., 2024a); and (2) medical-specialized models include `BioMistral` (Labrak et al., 2024), `Meditron-7B` (Chen et al., 2024), `Meditron3-8B` (Chen et al., 2023), `Chatdoctor` (Li et al., 2023). To evaluate zero-shot performance, we incorporate diagnosis labels in the prompt. For foundation models, we also test an alternative approach where we extract their generated embeddings and train an MLP for prediction.

We evaluate the prediction performance through two dimensions: (1) Coverage Ability. We use the recall@5 and F1 to measure the coverage ability. Since there are no lab test requests in single-turn dialogues, we evaluate lab test performance on the multi-turn dialogue data using recall@5. (2) Early-Prediction Ability. We utilize Mean Reciprocal Rank (MRR) to indicate the effectiveness of early diagnosis prediction. Details about evaluation metrics are shown in A.2.2 in Appendix.

4.3 Implementation Details

During implementation, we select Llama3.1-8B as the backbone foundation model, and append a mean-pooling layer and a two-layer MLP to it for lab test recommendation and diagnosis prediction. During tuning, r is set to 16 for the LoRA adapters (Hu et al., 2021). Batch size is set to 4 and learning rate is $1e-4$. Training epoch is set to 5. We randomly run each experiment twice and report the mean. We run experiments on four A100, and train:eval:test is set to 8:1:1.

4.4 Main Results and Analysis

Results on Single-turn Prediction. For single-turn diagnosis, additional lab test query is not needed. For space limitation, experimental results are presented in Table 2 and Table 4 in Appendix A.2.3. Several key observations can be made from the following dimensions: (1) **Superiority of DiaLLM.** It is obvious that DiaLLM consistently outperforms all baselines, demonstrating its effectiveness. This improvement is mainly due to the EHR transformation pipeline and rewarding models for PPO, which incorporate clinically grounded knowledge, thereby enhancing the model's decision-making abilities. (2) **Impact of Tuning.** Comparing base models (e.g., Mistralv0.3-7B, Llama3.1-8B, Qwen2.5-7B) with their supervised fine-tuned counterparts, we observe substantial performance gains across all metrics, hightlighting the importance of task-specific fine-tuning in improving decision-making in clinical dialogue tasks. (3) **Performance of clinical-Specific Models.** The performance of BioMistral and ChatDoctor is significantly lower than that of fine-tuned general-purpose models. This suggests that current biomedical LLMs are primarily trained for general clinical tasks, such as clinical QA, rather than being optimized for real medical data. (4) **Comparison with Fine-Tuned Baselines**. Among fine-tuned models, those on Llama3.1-8B and Qwen2.5-7B perform competitively. However, DiaLLM consistently outperforms them, indicating that its enhanced clinical grounding and policy optimization contribute to superior prediction.

Results on Multi-turn Prediction. For multi-turn dialogue data, models can query additional lab tests before giving the final prediction. As shown in Table 3, performance of DiaLLM significantly exceeds baseline models, similar to previous single-turn cases. The performance of general-used LLMs, e.g., Mistralv0.3-7B, Llama3.1-8B, Qwen2.5-7B still falls short in comparison to fine-tuned models and DiaLLM, indicating that they lack the clinical specificity necessary for multi-turn dialogues. Medical-specific LLMs like BioMis-

[5] https://wwwn.cdc.gov/nchs/nhanes/
[6] https://trinetx.com/
[7] For simplicity, we use shortened names later.

Model	TriNetX-Metabolic			TriNetX-Respiratory			TriNetX-Circulatory		
	Recall@5	F1	MRR	Recall@5	F1	MRR	Recall@5	F1	MRR
Mistralv0.3-7B	3.64	11.89	10.12	4.54	7.84	12.09	10.98	8.67	17.57
Llama3.1-8B	5.34	9.28	10.95	7.49	12.05	15.48	17.02	10.89	22.97
Qwen2.5-7B	8.58	10.51	12.11	10.03	10.64	16.25	5.84	7.82	9.68
BioMistral	39.67	13.82	26.10	25.95	15.50	33.78	20.01	16.62	31.08
ChatDoctor	19.16	11.57	20.82	35.88	9.71	26.29	10.03	10.03	18.34
Meditron-7B	6.84	9.31	11.66	11.58	10.34	16.80	15.07	14.71	18.88
Meditron3-8B	5.99	6.46	10.74	11.98	13.28	15.42	10.13	16.04	19.34
Mistralv0.3-7B$_{mlp}$	66.72	50.92	45.61	67.71	54.95	45.64	67.51	55.95	45.74
Llama3.1-8B$_{mlp}$	68.11	55.01	45.76	67.66	56.80	45.77	66.72	56.23	45.76
Qwen2.5-7B$_{mlp}$	69.16	52.95	45.72	67.66	55.21	45.71	68.86	55.18	45.79
DiaLLMs	**79.54**	**76.21**	**47.33**	**78.19**	**74.82**	**48.18**	**77.75**	**74.36**	**48.14**

Table 2: Performance Comparison of DiaLLM and Baselines on the Constructed Single-turn Dialogue.

Model	TriNetX-Metabolic				TriNetX-Respiratory				TriNetX-Circulatory			
	Diagnosis			Lab Test	Diagnosis			Lab Test	Diagnosis			Lab Test
	Recall@5	F1	MRR	Recall@5	Recall@5	F1	MRR	Recall@5	Recall@5	F1	MRR	Recall@5
Mistralv0.3-7B	6.89	8.07	13.23	0.58	14.44	16.02	11.98	1.96	18.00	16.93	25.39	0.21
Llama3.1-8B	12.44	11.84	14.79	8.52	18.22	17.84	26.08	0.75	12.72	13.24	21.24	1.87
Qwen2.5-7B	7.56	16.68	17.86	0.61	5.56	8.90	15.03	1.33	8.22	9.23	15.23	1.75
Meditron-7B	4.22	8.31	13.28	1.21	8.89	13.04	14.67	0.79	12.22	12.43	21.90	1.29
Meditron3-8B	16.22	21.25	21.57	2.90	17.33	15.38	24.32	1.37	8.22	15.25	17.26	1.91
BioMistral	9.37	13.54	16.62	2.33	12.67	19.51	22.98	1.70	4.44	13.37	10.56	0.79
Chatdoctor	6.87	9.71	18.92	1.77	2.89	9.66	9.44	0.95	16.00	19.22	22.08	1.18
Mistralv0.3-7B$_{mlp}$	57.90	56.03	37.42	10.45	59.76	59.14	38.90	10.51	59.56	58.90	39.93	10.39
Llama3.1-8B$_{mlp}$	57.31	55.92	37.43	10.64	61.22	59.58	39.42	10.97	61.21	59.47	40.13	10.16
Qwen2.5-7B$_{mlp}$	58.01	56.72	38.37	11.04	61.91	60.21	39.72	11.13	60.89	58.54	38.46	10.30
DiaLLMs	**70.22**	**73.59**	**41.88**	**11.43**	**70.89**	**75.12**	**40.52**	**12.31**	**71.11**	**75.39**	**43.34**	**11.55**

Table 3: Performance Comparison of DiaLLM with Baselines on the Constructed Multi-turn Dialogue.

tral, Meditron, and ChatDoctor underperform compared to general-purpose models such as Llama3.1-8B and Qwen2.5-7B. This is likely due to their training on open medical QA datasets or synthetic medical dialogues, which fail to capture the complexity of real-world clinical practice.

4.5 Ablation Study

To evaluate the effectiveness of different components in DiaLLM, we conduct an ablation study to quantify the contributions of the EHR transformation pipeline and the rewards in PPO. For space limitation, experimental results on single-turn and multi-turn dialogue data are shown in Figure 3, Figure 4 and Figure 5 in Appendix A.2.4.

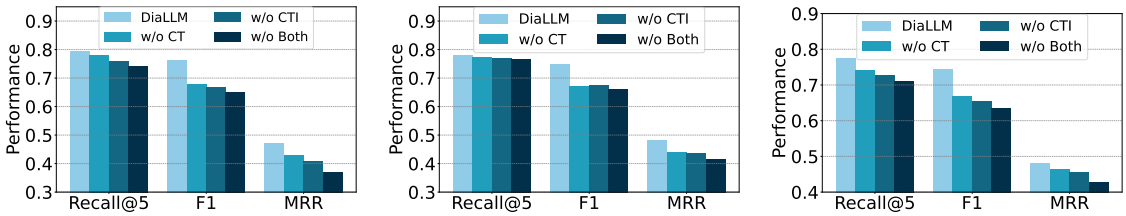
4.5.1 Ablation Study on EHR Transformation

We assess the impact of different EHR transformation components by progressively removing them: (1) DiaLLM (w/o CT): Removes code translation (CT). (2) DiaLLM (w/o CTI): Removes clinical test interpretation (CTI). (3) DiaLLM (w/o Both): Removes all EHR transformation components.

These results highlight the importance of integrating clinical grounded knowledge into LLMs. DiaLLM (w/o Both) results in a noticeable performance drop in both single-turn and multi-turn dialogue data, confirming that the EHR transformation strategy enhances the model's ability to understand EHR data and improves diagnostic predictions. Besides, DiaLLM (w/o CTI) exhibits lower performance than DiaLLM (w/o CT) across most datasets, highlighting the necessity of understanding clinical terminology and numerical values.

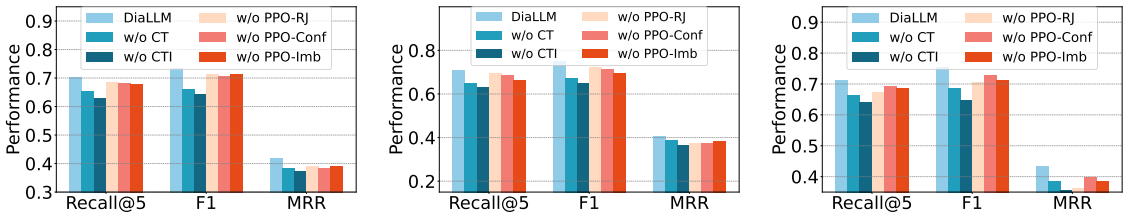
4.5.2 Ablation Study on Reward Modeling

We analyze the influence of specific PPO mechanisms: (1) DiaLLM (w/o PPO-RJ): Removes reject sampling. (2) DiaLLM (w/o PPO-Conf): Removes confirmation reward. (3) DiaLLM (w/o PPO-Imb): Removes imbalanced reward. Reward modeling is primarily effective for multi-turn dialogue data,



(a) TriNetX-Metabolic Dataset. (b) TriNetX-Respiratory Dataset. (c) TriNetX-Circulatory Dataset.

Figure 3: Ablation Study on Single-turn Dialogue Data.



(a) TriNetX-Metabolic Dataset. (b) TriNetX-Respiratory Dataset. (c) TriNetX-Circulatory Dataset.

Figure 4: Ablation Study on Multi-turn Dialogue Data.

and experimental results are shown in Figure 4.

Each PPO component contributes to model performance, as elaborated in the experimental results. In the TriNetX-Respiratory multi-turn dialogue dataset, removing the imbalanced reward reduces the F1-score to 0.6954, while removing the confirmation reward lowers it to 0.7129. These declines highlight the importance of addressing class imbalance (PPO-Imb) and incorporating confidence-based rewards (PPO-Conf) to improve classification quality, balance predictions, and enhance decision reliability. Furthermore, the EHR transformation component has a greater impact on model performance than reward modeling in multi-turn dialogue data, showing the importance of aligning medical terms with common-sense knowledge.

4.6 Case Study

We qualitatively evaluate the diagnostic recommendations of various models using the same example dialogue, as presented in Table 9 in the Appendix A.2.5. The selected case presents a complex scenario requiring not only diagnosis prediction but also the integration of additional clinical test evidence. DiaLLM initially recommended tests for carbon dioxide, chloride, etc. Upon receiving the results, it identified abnormalities in carbon dioxide and chloride levels and suggested further tests for glucose and sodium. Based on these analysis, DiaLLMs can recommend relevant diagnoses.

In contrast, all baseline medical LLMs lacked the ability to inquire about appropriate clinical tests or effectively analyze test results, highlighting a significant gap in diagnostic reasoning. BioMistral has a narrow diagnostic scope, potentially missing relevant diagnoses. While Meditron-7B and 8B considered a broader range of conditions, their diagnostic results diverged from the ground truth. ChatDoctor's diagnoses often lacked relevance to the patient's symptoms and lab results, suggesting limited comprehension of the medical history. Among these, DiaLLM demonstrated the most comprehensive and accurate diagnostic approach, effectively inquiring, analyzing lab results, and adjusting recommendations, showcasing superior practical medical value.

5 Conclusion

In this work, we introduced DiaLLM, a clinical dialogue model that enhances medical conversational systems through a combination of clinical-grounded EHR transformation and PPO optimization. Experimental results consistently show that DiaLLM outperforms existing models across a variety of metrics, demonstrating its superior prediction ability. By integrating domain-specific knowledge and a structured decision-making process, DiaLLM offers a significant improvement over both general-purpose and medical-specific LLMs in complex medical dialogues.

Limitations

Automation of CTR Database In this work, we recruited three medical students to annotate clinical test code descriptions and clinical value reference ranges based on medical guidelines and publicly available clinical resources (e.g., Mayo Clinic). Our constructed Clinical Test Reference data are crucial for diagnosis prediction using EHR data grounded in clinical knowledge. However, manual annotation is time-consuming and prone to variability. Future work will focus on automating this process to enhance efficiency and consistency.
Clincial Conversational System with Multimodal Data In this work, we focus on diseases that can be solely predicted through EHR data. Our approach focuses on EHR data and lab tests with numerical or textual features. With advances in multimodal learning, integrating imaging-based diagnostic tests, such as CT scans, could further enrich the model's clinical reasoning capabilities. We also leave this extension for future exploration.

Ethic Statement

Data Collection This study utilizes publicly available EHR data and de-identified in-hospital records approved for research use. To safeguard patient privacy, we conducted a thorough review to ensure the dataset contains no sensitive information. We adhere to strict ethical standards in data handling and acknowledge the research community's commitment to maintaining data integrity and privacy.
Compensation and Ethical Collaboration with Medical Professionals We collaborated with two senior medical graduates to identify diseases predictable solely from EHR data and to evaluate the constructed CTR. Each was compensated 200 RMB per hour, aligning with local salary standards to ensure fair remuneration for their expertise.
Concerning the Trustworthiness of DiaLLM While DiaLLM demonstrate strong diagnostic reasoning capabilities, its reliability in real-world clinical settings is not yet fully established. Potential risks include hallucinations, biases from training data, and inconsistencies in medical reasoning. To enhance trust, further research is needed to improve model transparency, robustness, and validation through expert evaluation. Developing rigorous verification mechanisms and integrating clinician oversight will be essential for safe deployment in medical practice.

References

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. 2014. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *International conference on machine learning*, pages 405–413. PMLR.

Som S Biswas. 2023. Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5):868–869.

Derun Cai, Chenxi Sun, Moxian Song, Baofeng Zhang, Shenda Hong, and Hongyan Li. 2022. Hypergraph contrastive learning for electronic health records. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 127–135. SIAM.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zeming Chen, Angelika Romanou, Antoine Bonnet, Alejandro Hernández-Cano, Badr Alkhamissi, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, et al. 2024. Meditron: Open medical foundation models adapted for clinical practice.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Charles Elkan. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885.

Arsene Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaetan Marceau Caron, and Joumana Ghosn.

2022. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24502–24515.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Trevor J Hastie. 2017. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Weijie He and Ting Chen. 2022. Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4695–4703.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yujing Hu, Weixun Wang, Hangtian Jia, Yixiang Wang, Yingfeng Chen, Jianye Hao, Feng Wu, and Changjie Fan. 2020. Learning to utilize shaping rewards: A new approach of reward shaping. *Advances in Neural Information Processing Systems*, 33:15931–15941.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023b. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Mohammad Kachuee, Orpaz Goldstein, Kimmo Karkkainen, Sajad Darabi, and Majid Sarrafzadeh. 2019. Opportunistic learning: Budgeted cost-sensitive learning from data streams. *arXiv preprint arXiv:1901.00243*.

William La Cava, Christopher Bauer, Jason H Moore, and Sarah A Pendergrass. 2019. Interpretation of machine learning predictions for patient outcomes in electronic health records. In *AMIA Annual Symposium Proceedings*, volume 2019, page 572. American Medical Informatics Association.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. 2024. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6555–6565.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, et al. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy Chang, and Wei Wang. 2024. Clibench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. 2016. Offline evaluation of online reinforcement learning algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jonathan Montomoli, Luca Romeo, Sara Moccia, Michele Bernardini, Lucia Migliorelli, Daniele Berardini, Abele Donati, Andrea Carsetti, Maria Grazia Bocci, Pedro David Wendel Garcia, et al. 2021. Machine learning using the extreme gradient boosting (xgboost) algorithm predicts 5-day delta of sofa score at icu admission in covid-19 patients. *Journal of Intensive Medicine*, 1(02):110–116.

Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor xai: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 629–639.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2014. Optimizing f-measures by cost-sensitive classification. *Advances in neural information processing systems*, 27.

Yuchao Qin, Mihaela van der Schaar, and Changhee Lee. 2024. Risk-averse active sensing for timely outcome prediction under cost pressure. *Advances in Neural Information Processing Systems*, 36.

Weijieying Ren, Xiaoting Li, Huiyuan Chen, Vineeth Rakesh, Zhuoyi Wang, Mahashweta Das, and Vasant G Honavar. 2024. Tablog: Test-time adaptation for tabular data using logic rules. In *Forty-first International Conference on Machine Learning*.

Savyasachi V Shah. 2024. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953–e2425953.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. 2021. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics*, 115:103671.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Kai-Fu Tang, Hao-Cheng Kao, Chun-Nan Chou, and Edward Y Chang. 2016. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *In Proceedings of NIPS Workshop on Deep Reinforcement Learning*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. 2023b. Biobridge: Bridging biomedical foundation models via knowledge graph. *arXiv preprint arXiv:2310.03320*.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Hao-Ren Yao, Nairen Cao, Katina Russell, Der-Chen Chang, Ophir Frieder, and Jeremy T Fineman. 2024. Self-supervised representation learning on electronic health records with graph kernel infomax. *ACM Transactions on Computing for Healthcare*, 5(2):1–28.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2414–2423.

Jinsung Yoon, James Jordon, and Mihaela Schaar. 2019. Asac: Active sensing using actor-critic models. In *Machine Learning for Healthcare Conference*, pages 451–473. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Appendix

Model	NHANES		
	Recall@1	F1	MRR
Mistralv0.3-7B	28.60	40.48	37.41
Llama3.1-8B	29.70	45.69	47.64
Qwen2.5-7B	28.97	44.83	44.38
BioMistral	36.19	48.25	43.13
ChatDoctor	39.83	54.29	46.72
Meditron-7B	36.82	52.47	47.77
Meditron3-8B	37.69	55.34	49.78
Mistralv0.3-7B$_{mlp}$	91.98	92.47	97.99
Llama3.1 8B$_{mlp}$	92.15	92.66	98.06
Qwen2.5-7B$_{mlp}$	92.07	92.45	98.02
DiaLLMs	**95.94**	**96.07**	**98.97**

Table 4: Performance Comparison of DiaLLM and Baselines on NHANES Dataset.

A.1 Methodology

A.1.1 Illustration of EHR data.

We summarize main elements and corresponding examples in EHR data in Table 5. EHR data is heterogeneous and can be represented as a sequence of tables.

A.1.2 Dialogue Data Construction from EHR

The EHR transformation comprises two components (1) Clinical Code Translation and (2) Clinical Test Interpretation. Examples of our constructed *Clinical Code Translation* data is shown in Table 6 and examples of *Clinical Test Interpretation* data is shown in 7, respectively.

A.2 Experiments

A.2.1 Statistical Analysis of Datasets.

We exclude clinical tests with a frequency of less than 10 across the entire dataset. Additionally, we filter out tests for which no test results are available for each patient. Statistical analysis of datasets is shown in Table 8.

	TriNetX-Metabolic	TriNetX-Respiratory	TriNetX-Circulatory	NHANES
#Samples	15844	10068	6747	8897
#Diagnosis	49	47	49	4
#Clinical Test	272	301	217	45

Table 8: Statistical Analysis of Datasets.

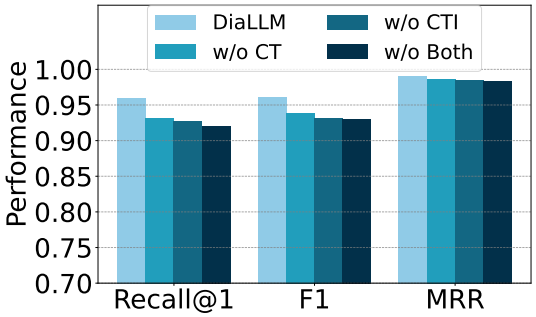


Figure 5: Ablation Study on NHANES Dataset.

A.2.2 Evaluation Metrics

The Mean Reciprocal Rank (MRR) is a metric used to evaluate systems that return a ranked list of answers to queries, focusing on the position of the first relevant answer. A higher MRR indicates that relevant items tend to appear higher in the ranked list of results. It is defined as the average of the reciprocal ranks of the first relevant answer for a set of queries. Formally, MRR is expressed as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{4}$$

A.2.3 Main Results on NHANES Dataset

Experimental results on the NHANES dataset is shown in Table 4. It can be observed that: (1) Medical-specialized models consistently outperform general foundation models; (2) When adapting foundation models with an MLP classifier trained on extracted embeddings, performance improves drastically; (3) DiaLLM significantly outperforms all baselines, demonstrating its superior ability in diagnosis prediction. These results highlight the effectiveness of DiaLLM's integration of structured EHR data and reinforcement learning framework, enabling more accurate and clinically relevant predictions compared to both zero-shot LLMs and embedding-based MLP classifiers.

A.2.4 Ablation Study on NHANES Dataset

Experimental results on the NHANES dataset is shown in Figure 5. The NHANES dataset presents a simpler task compared to TriNetX data. The ablation study highlights the critical role of clinically grounded data transformation and rewarding modeling for improving diagnosis performance.

A.2.5 Case Study

The case study example is presented in Table 9.

EHR Data	Description	Event	Visit Time
Demographics	General characteristics of patients	Age, gender, ethnicity/race	
Symptom	Patient symptoms are cataloged in the ICD system under codes R00-R99.	ICD-10-CM R10.12: Left upper quadrant pain	2019-02-15
Vital Signs	Medical signs indicating the status of the body's vital functions	LOINC 8462-4 82 mm[Hg] LOINC 29463-7, 84 [lb_av] LOINC 8480-6, 124 mm[Hg]	2019-02-15 2019-02-19 2019-02-22
Laboratory Results	Medical examination results, generate are organized in a code format	LOINC 73578-7,8 mmol/L, LOINC 6768-6 48 U/L, LOINC 19261-7 Negative	2019-02-15 2019-02-19 2019-02-22
Diagnostic	Codes representing diseases and related health problems (e.g., ICD-9 and ICD-10)	ICD-10-CM Q65: Congenital deformities of hip, ICD-10-CM M25.552: Pain in left hip, ICD-10-CM R06.02: Shortness of breath	2019-02-22

Table 5: Description of key components and corresponding examples in EHR data.

LOINC Code	Description
100091-8	Trypanosoma cruzi Ab [Units/volume] in Serum by Immunoassay
100092-6	Trypanosoma cruzi Ab bands panel - Serum by Immunoblot
100093-4	Trypanosoma cruzi 15-16kD IgG Ab [Presence] in Serum by Immunoblot
100094-2	Trypanosoma cruzi 21-22kD IgG Ab [Presence] in Serum by Immunoblot
100095-9	Trypanosoma cruzi 27-28kD IgG Ab [Presence] in Serum by Immunoblot
100096-7	Trypanosoma cruzi 42kD IgG Ab [Presence] in Serum by Immunoblot
100097-5	Trypanosoma cruzi 45-47kD IgG Ab [Presence] in Serum by Immunoblot
100098-3	Trypanosoma cruzi 120-200kD IgG Ab [Presence] in Serum by Immunoblot
100099-1	Trypanosoma cruzi 160kD IgG Ab [Presence] in Serum by Immunoblot
1001-7	DBG Ab [Presence] in Serum or Plasma from Donor
10010-7	R' wave amplitude in lead AVF
100100-7	Fasciola sp IgG Ab [Presence] in Serum by Immunoassay
100101-5	Fasciola sp 8-9kD IgG Ab [Presence] in Serum by Immunoblot
100102-3	Fasciola sp 27-28kD IgG Ab [Presence] in Serum by Immunoblot

Table 6: The Mapping between LOINC Codes and their Corresponding Clinical Descriptions.

Lab Code	Test Name	Age Range	Gender	Reference Range	Unit	Critical Value
2823-3	Potassium Serum	1–18 years	Any	3.4–4.7	mEq/L	-
2823-3	Potassium Serum	Any ≥18 years	Any	3.5–5.2	mEq/L	-
17861-6	Total Calcium	<1 year	Any	8.7–11.0	mg/dL	-
17861-6	Total Calcium	1–17 years	Any	9.3–10.6	mg/dL	-
17861-6	Total Calcium	18–59 years	Any	8.6–10.0	mg/dL	-
17861-6	Total Calcium	≥60 years	Any	8.8–10.2	mg/dL	-
33914-3	eGFR (Estimated Glomerular Filtration Rate)	Any	Any	>90	mL/min/1.73m²	-
33914-3	eGFR (Estimated Glomerular Filtration Rate)	Any	Any	None	mL/min/1.73m²	Mildly low: 60–89
33914-3	eGFR (Estimated Glomerular Filtration Rate)	Any	Any	None	mL/min/1.73m²	Moderate to severe kidney disease: 30–59
33914-3	eGFR (Estimated Glomerular Filtration Rate)	Any	Any	None	mL/min/1.73m²	Kidney failure: <15

Table 7: Examples of the Manually Constructed Clinical Test Interpretation Data.

A Sub-piece of Initial EHR Data: Demographic Information: age: 30, gender: female, race: white Symptom Code : R10.2, R10.32 Lab Test Code : 2028-9, 2075-0, 2951-2, 3040-3, 33037-3, 48642-3, 777-3, 1742-6, 2028-9, 2075-0 Lab Results Values : 31.0, 105.0, 83.0, 139.0,7.0, 99.0, 283.0, 27.0, Vital Sign Code : 8302-2, 8310-5, 8462-4, 8480-6, 9279-1 Vital Sign Values : 62.0, 98.0, 70.7, 119.7, 98.7 Diagnosis Codes: E03.9, K80.62, K81.0, N80.9, K80.20, K80.10
Prompts: Given the patient demographic information : (age: 30, gender: female, race: white). Patient shows symptom of : (1) Pelvic and perineal pain; (2) Right lower quadrant pain. These lab tests show normal results : (1) Carbon dioxide, totalin Serum or Plasma, (2) Chloridein Serum or Plasma, (3) Glucosein Serum or Plasma, (4) Sodiumin Serum or Plasma, (5) Lipasein Serum or Plasma, (6) Glomerular filtration rate/1.73 sq M predicted among non-blacksin Serum or Plasma by Creatinine-based formula (MDRD). These lab tests show abnormal results : (1) Plateletsin Blood by Automated count. These vital sign show normal Results : (1) Temperature, (2) Blood Pressure (Systolic), (3) Blood Pressure (Diastolic), (4) Respiratory Rate, These vital sign show abnormal values : (1) Heart Rate What are your recommended diagnosis? If you need further information, what are your recommended clinical test for further decision making?
DiaLLM I recommend you to take (1) Alanine aminotransferase in Serum or Plasma; (2) Carbon dioxide in Serum or Plasma, (3) Chloride in Serum or Plasma. **Prompts (User)**: My Alanine aminotransferase in Serum or Plasma results is 27.0 U/L , Carbon dioxide in Serum or Plasma is 83.0 mEq/L, Chloride in Serum or Plasma is 139.0 mEq/L. **DiaLLM**: These lab test shows normal results: (1) Alanine aminotransferase in Serum or Plasma results; These lab test shows abnormal results: (1) Carbon dioxide in Serum or Plasma, (2) Chloride in Serum or Plasma. I further recommend you to take (1) Glucose in Serum or Plasma; (2) Sodium in Serum or Plasma; **Prompts (User)**: My Glucose in Serum or Plasma is 7.0 $mmol/L$, (2) Sodium in Serum or Plasma 99.0 $mmol/L$. **DiaLLM**: These lab test shows abnormal results: (1) Glucose in Serum or Plasma; (2) Sodium in Serum or Plasma. I recommend the following possible diagnosis: (1) disorders of thyroid gland, (2) disorders of gallbladder, biliary tract and pancreas, (3) Noninflammatory disorders of female genital tract.
BioMistral I recommend the following possible diagnosis: (1) Chlamydia trachomatis infection, (2) Pelvic inflammatory disease, (3) Pelvic pain
Meditron-7B I recommend the following possible diagnosis: (1) Pelvic inflammatory disease, (2) Acute appendicitis, (3) Ovarian torsion, (4) Ectop
Meditron-8B I recommend the following possible diagnosis: (1) Pelvic inflammatory disease, (2) Chlamydia trachomatis infection, (3) Pelvic inflammatory disease with chlamydia trach.
Chatdoctor I recommend the following possible diagnosis: (1) pelvic inflammatory disease, (2) tuberculosis, (4) acute pancreat.

Table 9: Sample Responses for Case Study.