REACT-LLM: A Benchmark for Evaluating LLM Integration with Causal Features in Clinical Prognostic Tasks

Linna Wang[1*], Zhixuan You[1*], Qihui Zhang[2*], Jiunan Wen[1], Ji Shi[1], Yimin Chen[3], Yusen Wang[1], Fanqi Ding[1], Ziliang Feng[1], Li Lu[1†]

[1]Sichuan University
[2] Peking University
[3] The Second Affiliated Hospital of Kunming Medical University
lenawang@stu.scu.edu.cn, youzhixuan@stu.scu.edu.cn, zqhui_scu@foxmail.com, jiunanwen@stu.scu.edu.cn, 2024141520215@stu.scu.edu.cn, 20241105@kmmu.edu.cn, 2024141520213@stu.scu.edu.cn, dingfanqi@stu.scu.edu.cn, fengziliang@scu.edu.cn, luli@scu.edu.cn

Abstract

Large Language Models (LLMs) and causal learning each hold strong potential for clinical decision making (CDM). However, their synergy remains poorly understood, largely due to the lack of systematic benchmarks evaluating their integration in clinical risk prediction. In real-world healthcare, identifying features with causal influence on outcomes is crucial for actionable and trustworthy predictions. While recent work highlights LLMs' emerging causal reasoning abilities, there lacks comprehensive benchmarks to assess their causal learning and performance informed by causal features in clinical risk prediction. To address this, we introduce REACT-LLM, a benchmark designed to evaluate whether combining LLMs with causal features can enhance clinical prognostic performance and potentially outperform traditional machine learning (ML) methods. Unlike existing LLM-clinical benchmarks that often focus on a limited set of outcomes, REACT-LLM evaluates 7 clinical outcomes across 2 real-world datasets, comparing 15 prominent LLMs, 6 traditional ML models, and 3 causal discovery (CD) algorithms. Our findings indicate that while LLMs perform reasonably in clinical prognostics, they have not yet outperformed traditional ML models. Integrating causal features derived from CD algorithms into LLMs offers limited performance gains, primarily due to the strict assumptions of many CD methods, which are often violated in complex clinical data. While the direct integration yields limited improvement, our benchmark reveals a more promising synergy: LLMs serve effectively as knowledge-rich collaborators for identifying and optimizing causal features. Additionally, in-context learning improves LLM predictions when prompts are tailored to the task and model. Different LLMs show varying sensitivity to structured data encoding formats, for example, open-source models perform better with JSON, while smaller models benefit from narrative serialization. These findings highlight the need to match prompts and data formats to model architecture and pretraining.

Code —
https://github.com/LinnaWang-Lena/REACT_LLM

Extended version — https://arxiv.org/abs/2511.07127

Introduction

In clinical environments such as the Intensive Care Unit (ICU), timely and accurate risk assessment is critical for enabling early interventions and improving patient outcomes (Fihn et al. 2024; Yeh et al. 2024). The widespread adoption of Electronic Health Records (EHRs) has provided access to rich, structured clinical data, opening new opportunities for data-driven decision support (Khalifa, Albadawy, and Iqbal 2024; Wang et al. 2025). One key application is prognostic modeling (Van Smeden et al. 2021), which estimates the risk of patients developing specific conditions over time. In this context, machine learning (ML) models outperform traditional statistical methods in terms of scalability and predictive performance. However, ML models often depend on high-dimensional features and remain vulnerable to input variability and spurious correlations (Rajpurkar et al. 2022; Wang et al. 2025). This has motivated growing interest in applying causal discovery (CD) to clinical downstream tasks, as advances in CD methods (Zhou and Chen 2022; Zanga, Ozkirimli, and Stella 2022; Feuerriegel et al. 2024) enable the extraction of meaningful causal relationships from observational data.

Recently, Large Language Models (LLMs), known for their impressive capabilities across a wide range of natural language processing tasks, have been increasingly applied to clinical applications (Ferdush, Begum, and Hossain 2024; Singhal et al. 2025; Liu et al. 2023; Sandmann et al. 2024; Kafkas et al. 2025; Kang et al. 2025). Leveraging strategies such as fine-tuning (Ben Shoham and Rappoport 2024; Wang et al. 2024a), retrieval-augmented generation (RAG) (Bedi, Thukral, and Dhiman 2025), and task-specific prompting (Zheng et al. 2025), LLMs have shown promise in clinical risk prediction. For example, (Ben Shoham and Rappoport 2024) fine-tuned a pre-trained LLM to predict diagnoses and hospital readmissions, achieving state-of-the-art performance.

Although LLMs and causal learning each show strong potential for CDM, their potential synergy remains largely

[*]These authors contributed equally.
[†]Corresponding author.

unexplored. This is primarily due to the lack of a systematic benchmark for evaluating their integrated application in clinical risk prediction tasks. Moreover, evidence on LLM performance in this domain is not uniformly positive, with some studies concluding that LLMs are not yet prepared for autonomous CDM (Hager et al. 2024; Brown et al. 2025). A recent benchmark, ClinicalBench (Chen et al. 2024), systematically evaluated general-purpose and medical-specific LLMs against traditional ML models on 3 prognostic tasks. The study revealed that despite variations in model scale and the use of different prompting or fine-tuning strategies, LLMs consistently underperformed ML models. Therefore, the predictive capabilities of LLMs in clinical risk prediction require further investigation. A comprehensive benchmark that integrates LLMs, causal learning, and clinical prediction is urgently needed. Motivated by this, this study aims to rigorously evaluate whether incorporating causal knowledge can improve the performance of LLMs on critical clinical risk prediction tasks. Unlike prior work limited to narrow clinical endpoints, we extend the evaluation to a broader set of prognostic outcomes to address the central question: *Can the integration of LLMs with causal features enhance performance in clinical prognostic tasks and potentially outperform traditional ML models?*

REACT-LLM Design. To answer this question, we present REACT-LLM (Risk Evaluation and Causal features Test with LLMs), a novel benchmark for assessing how effectively LLMs can serve as inference and error-correction experts that augment CD methods in uncovering causal features of clinical risk outcomes (Figure 1). Using structured data from 2 real-world datasets, MIMIC-III(Johnson et al. 2016) and MIMIC-IV(Johnson et al. 2023), we investigate 7 representative prognostic tasks: (1) In-hospital mortality (InHospDeath), (2) 30-day hospital readmission (Readmit30), (3) Multiple ICU stays during a single hospitalization (MultiICU), (4) Sepsis during ICU stay (SepsisICU), (5) Acute kidney injury during ICU stay (AKIICU), (6) Prolonged hospital stay (LOS), and (7) Early ICU admission (EarlyICU). Here are the main tasks:

▷ Baseline evaluation: Benchmark all MLs/LLMs on 7 outcomes across 2 datasets using complete feature sets.

▷ Prompt engineering evaluation: Beyond direct prompting, 4 representative prompt engineering strategies (Chain-of-Thought (CoT), Self-Reflection (SR), Role-Playing (RP), and In-Context Learning (ICL)) are employed to assess LLMs performance across 7 tasks.

▷ Input format sensitivity evaluation: Assess LLMs performance across 5 formats for structured patient data: Row-Column Format (RCF), JSON, LaTeX, Template-Based Natural Language (TBNL) and Narrative Serialization (NS).

▷ Causal feature evaluation: Use 3 representative CD methods to identify features with direct or indirect causal relationships to each outcome. Evaluate LLMs using only these causal features as input.

▷ LLM-assisted causal feature editing evaluation: Prompt LLMs to optimize the causal feature sets derived from 3 CD methods, and separately, to generate causal feature sets for each outcome relying solely on their internal knowledge.

Overall, the contributions can be summarized as follows:

- Benchmark: Evaluate **3** CD algorithms, **6** ML models, and **15** LLMs (spanning diverse model sizes and architectures) for predicting **7** clinical outcomes.
- Two Datasets: We curate the MIMIC-III and MIMIC-IV datasets, encompassing **6** categories of clinical information and **7** prognostic outcome labels, providing a comprehensive foundation for evaluating LLM-based medical risk prediction models.
- Actionable Insights & Nuanced Findings: (1) LLMs have yet to outperform traditional ML models in clinical outcome prediction. (2) Engineering prompts offer slight gains under imbalanced EHR conditions. (3) LLMs refine CD-derived features effectively, enabling a human–AI synergy for causal feature engineering. (4) Different LLMs exhibit varying sensitivity to structured EHR encoding formats. RCF benefits proprietary large models, JSON consistently enhances performance in open-source models, while NS proves effective for smaller models. LLMs can infer causal relations even among variables with ambiguous or coded names, such as ICD codes.

REACT-LLM Construction

Goals

- **Causal Feature Identification.** Recover the causal structure from observational data using 3 CD methods by assigning each outcome to its direct and indirect causes.
- **Clinical Binary Outcome Prediction.** Estimate the probability of binary outcomes.
- **LLM Evaluation Tasks.** Evaluate LLMs across 5 groups of experiments to address the following questions:

  Q1: Can LLM outperform traditional ML in clinical prognosis? → *Based on baseline evaluation.*

  Q2: Can prompt engineering boost LLM performance in clinical risk prediction? → *Based on prompt engineering evaluation.*

  Q3: How does the encoding format of structured EHR data affect LLM performance? → *Based on input format sensitivity evaluation.*

  Q4: Can CD improve LLM predictions by identifying causal features? → *Based on causal feature evaluation.*

  Q5: Can LLM validate or identify causal features in clinical contexts? → *Based on LLM-assisted causal feature editing evaluation.*

Datasets Preprocessing

Study Population. This study uses 2 public datasets[1]: MIMIC-III (v1.4) (Johnson et al. 2016) and MIMIC-IV (v3.1) (Johnson et al. 2023), which contain de-identified EHRs from ICU and emergency department admissions at Beth Israel Deaconess Medical Center in Boston. MIMIC-III contains ICU records from 2001 to 2012 and is widely used in critical care research, while MIMIC-IV extends coverage through 2022 with an updated schema and improved

[1]https://physionet.org

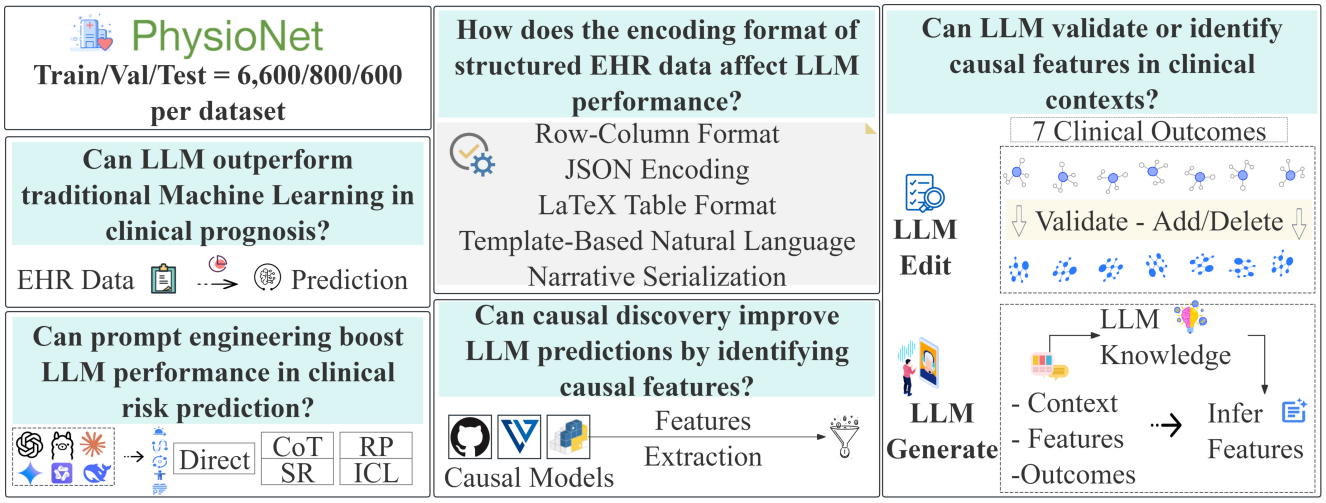

Figure 1: The REACT-LLM framework diagram summarizes the dataset partitions, clinical outcome prediction tasks, and the full evaluation pipeline, incorporating Direct prompting, Chain-of-Thought (CoT), Self-Reflection (SR), Role-Playing (RP), and In-Context Learning (ICL).

data quality. We include only adult patients (age $\geq$ 18), retained the first ICU stay per admission, and excluded ICU stays shorter than one day to ensure clinical relevance.

Data Collection. We extract 5 categories of features to characterize each ICU admission: (1) demographics and admission details (age, gender, admission type, initial ICU unit); (2) 65 high-frequency admission diagnoses from MIMIC-III/IV, encoded as binary indicators (present/absent); (3) 27 high-frequency clinical procedures (binary encoded); (4) 55 medications administered within the first 24 hours (total dosage per drug); and (5) 115 vital signs and laboratory results from the first 24 hours (median summarized). The 24-hour window supports early prediction while capturing key clinical information from the initial phase of ICU care.

Experiment Setup

Prediction Tasks. We define 7 outcome labels (not mutually exclusive): (1) InHospDeath: in-hospital mortality; (2) Readmit30: a binary indicator set to 1 if a subsequent hospital admission occurs within 30 days of discharge; (3) MultiICU: more than one ICU stay associated with the same hospital admission; (4) SepsisICU: identified using the database-provided label based on the Sepsis-3 definition (Singer et al. 2016), which requires a SOFA score $\geq$ 2 and clinical suspicion of infection; (5) AKIICU: defined according to KDIGO criteria (Khwaja 2012), using creatinine or urine output when available, with baseline creatinine taken as the lowest value within the prior 7 days; (6) LOS: length of hospital stay exceeding 14 days; and (7) EarlyICU: ICU admission occurring within 12 hours of hospital admission.

Dataset Division. To manage the computational cost of querying LLMs, we apply stratified random sampling to ensure proportional representation of positive cases for each label, selecting 8,000 samples from each datasets, resulting in a combined cohort of 16,000 patients. For traditional ML models, the combined dataset is split into training (6,600), validation (800), and test (600) sets. The test sets from both datasets are shared with LLM evaluations to ensure fair and consistent comparisons.

Metrics. This paper uses AUROC, AUPRC and F1 score. We report 95% confidence intervals based on 1,000 bootstrap samples for ML models and 5-run results for LLMs.

Method

Causal Discovery Models: We evaluate 3 representative CD approaches: (1) functional-based (DirectLiNGAM (Shimizu et al. 2011)), (2) score-based (GES (Chickering 2002)), and (3) gradient-based (CORL (Wang et al. 2021)). All methods were implemented using the `gCastle` toolkit[2], an open-source causal discovery library developed by Noah's Ark Lab (Zhang et al. 2021). We apply each CD method 5 times on the full dataset to identify direct and indirect causes for each outcome. Features appearing in more than 2 runs are retained to form the causal feature set. The parameter Settings are shown in the Appendix A.2.

Benchmarked ML Models: We include 6 common used baseline models: AdaBoost, Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and XGBoost. All models are run with default hyperparameters to ensure fairness and reproducibility, allowing us to examine whether the LLM can outperform unoptimized traditional ML models.

Prompt Protocols. Following predefined clinical categories (patient demographics, diagnoses, procedures, medications, and laboratory results including vital signs), we evaluate 5 prevalent encoding strategies for transforming structured EHR data into formats suitable for LLM input: (1) RCF: Encodes records in a flat, CSV-like format using

[2]https://github.com/huawei-noah/trustworthyAI/tree/master/gcastle

comma-separated values. (2) JSON Encoding: Represents data as hierarchical key-value pairs organized by clinical categories. (3)LaTeX Table Format: Structures EHR data into formatted LaTeX-style tables by category. (4) TBNL: Converts each clinical category into sentence-level descriptions using predefined templates to enhance textual fluency. (5) NS: Converts tabular EHR data into coherent, human-readable narratives, structured by the predefined clinical categories. We proceed as follows:

- Baseline Evaluation: Applies the Direct Prompting approach using LLMs on input formatted with NS.
- Prompt Engineering Evaluation: Extends beyond Direct Prompting by incorporating strategies CoT, SR, RP and ICL, all under the NS format.
- Input Format Sensitivity Evaluation: Tests the Direct Prompting approach across all 5 input formats (RCF, JSON, LaTeX, TBNL, NS) to assess the impact of encoding on performance.
- Causal Feature Evaluation: Uses only the feature subsets identified by CD methods, test under the Direct Prompting setting with NS input.

The prompt protocols for these 4 groups follow a consistent design and are detailed in the Appendix C and E.

For the LLM-assisted causal feature editing evaluation, we design 2 protocols (details are in the Appendix D):

- LLM edited CD feature sets: For each clinical outcome, feature sets generated by a CD algorithm are provided to the LLM, which was prompted to act as an expert in intensive care medicine and refine the list. The LLM's output was used as the optimized feature set.
- LLM-Generated causal feature sets: The LLM independently generates a causal feature set based solely on its internalized clinical knowledge.

LLM Implement Details: Experiments are conducted on 3 NVIDIA RTX A800 GPUs. We benchmark a diverse set of LLMs spanning different architectures and scales. To ensure reproducibility, all inferences use greedy decoding (temperature = 0, do_sample = False). *Proprietary models* included advanced reasoning models (GPT-o1, GPT-o3 mini, Claude-3.7-Sonnet, Gemini-2.5-Pro, Gemini-2.5-Flash), large models (GPT-4o, Claude-4, Claude-3.5-Haiku), and small models (GPT-4o-mini). *Open-source models* included a top reasoning model (DeepSeek-R1), large models (Llama-3.1-405b, DeepSeek-V3, Qwen3-235b), and small models (Qwen3-8b, Qwen3-14b). Details are in Appendix A.1.

Empirical Results and Analysis

Baseline Evaluation

We evaluate 15 LLMs with direct prompting across 7 clinical prediction tasks on 2 datasets, using 3 metrics (complete results in the Appendix B.1 and B.2). As shown in Table 1, among traditional ML models, XGBoost, RF, and LR consistently achieve top performance across clinical outcomes. In the open-source LLM category, larger models (e.g., DeepSeek-V3, Llama-3.1-405b) outperform smaller and thinking models more frequently. While thinking models rarely lead on individual tasks, they exhibit more stable performance than smaller models overall. A similar trend holds for proprietary LLMs: larger models generally outperform other ones. Notably, Gemini-2-Flash achieves the highest AUPRC on the EarlyICU task (0.8540), exceeding the best ML model (RF, 0.8473). Thinking models in this group also show more consistent performance than smaller variants. Overall, proprietary LLMs outperform open-source ones more often across tasks.

However, with the exception of Gemini-2-Flash on the EarlyICU AUPRC metric, none of the LLMs surpass traditional ML models on any clinical prediction task. In most cases, LLM performance lags behind ML baselines by a substantial margin, typically around 10–20%.

Based on this, we select **6** representative small, large and thinking models from both open-source and proprietary LLMs for subsequent experiments.

Finding for Q1: Current LLMs remain immature and unreliable for clinical prognostic decision support. Among LLMs, proprietary and large-scale models tend to offer relatively better performance.

Prompt Engineering Evaluation

We evaluate 6 LLMs with 5 prompting strategies across 7 clinical prediction tasks(complete results in Appendix B.3). Across all tasks and metrics, ICL surpasses the baseline 41 times and ranks best in 22. SR also performs well, exceeding the baseline in 38 cases and leading in 19. Role-Playing and CoT follow. As shown in Figure 2, ICL outperforms direct prompting in 18 cases and achieves the highest AUROC in 13. Prompting strategies yield clear improvements on the Readmit30 task. ICL enhances performance: Gemini-2-Pro's AUROC increases by 0.077, GPT-o3-mini's AUROC and F1 improve by 0.072 and 0.084, respectively, and Qwen3-235B gains 0.059 in AUROC. DeepSeek-R1 benefits more from CoT prompting, with a 0.0271 increase in AUPRC and a 0.0912 gain in AUROC. However, despite these gains, none of the prompting strategies enable LLMs to surpass traditional ML models across any outcome.

Finding for Q2: ICL can enhance LLM performance

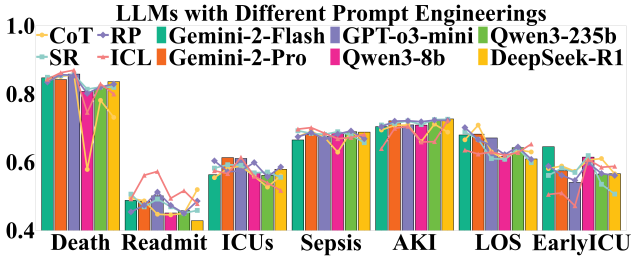

Figure 2: AUROC of LLMs under different prompting strategies on MIMIC-III across 7 clinical labels. Outcomes include in-hospital mortality, 30-day readmission, multiple ICU stays, progression to Sepsis-3, development of AKI, prolonged LOS, and early ICU transfer. Baseline results from direct prompting are shown as bars.

Outcome	InHospDeath		Readmit30		MultiICU		SepsisICU		AKIICU		LOS		EarlyICU	
Positive Ratio	27.0%		13.6%		17.7%		46.9%		55.0%		29.1%		69.5%	
Metric	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
Machine Learning Models														
SVM	0.8949	0.7707	0.6668	0.2393	0.7467	0.4141	0.8059	0.7956	0.8221	0.8558	0.8499	0.6949	**0.7336**	0.8304
LR	0.9079	0.8047	**0.7326**	**0.2912**	**0.7555**	**0.4419**	0.8059	0.7856	0.8285	0.865	**0.8552**	0.7022	0.6884	0.8406
DT	0.7022	0.6354	0.5323	0.2535	0.5617	0.3227	0.659	0.7265	0.6499	0.7815	0.6482	0.5699	0.5885	0.8347
RF	0.8936	0.7806	0.6222	0.2122	0.7109	0.3507	**0.8229**	**0.8059**	0.8323	0.8576	0.8521	0.6917	0.7263	**0.8473**
Adaboost	0.8694	0.7231	0.6837	0.253	0.7418	0.3532	0.8111	0.795	0.8318	**0.8655**	0.8404	0.6736	0.7180	0.8443
XGboost	**0.9088**	**0.8148**	0.6863	0.2469	0.7199	0.3737	0.8183	0.7964	**0.8332**	0.8606	0.8516	**0.7099**	0.7312	0.8319
Open-source Large Language Models														
Qwen3-14b	0.8075	0.6544	0.4502	0.1002	0.5719	0.1472	0.6848	0.6528	0.7092	0.7679	0.6219	0.3143	0.6179	0.739
Qwen3-8b	0.8076	0.6537	0.4511	0.0999	0.5711	0.1476	0.6866	0.6546	0.7089	0.7676	0.6172	0.3129	0.615	0.7377
Llama-3.1-405b	0.7012	0.4658	0.4929	0.1304	0.5545	0.1681	0.5687	**0.7139**	0.6988	**0.7969**	0.5355	**0.6419**	0.5389	0.7477
Qwen3-235b	0.8194	0.6878	0.4580	0.0964	0.5615	0.1695	**0.6894**	0.6604	0.7221	0.7409	0.6425	0.5511	0.5641	0.7197
DeepSeek-V3	0.8370	0.6873	0.4478	0.0796	0.5845	0.2126	0.6756	0.6384	**0.7413**	0.7862	0.6531	0.4192	0.5567	0.7796
DeepSeek-R1	0.8363	0.6649	0.4285	0.1030	0.5793	0.2104	0.6883	0.6487	0.7270	0.7667	0.6096	0.3732	0.5663	0.7660
Proprietary Large Language Models														
GPT-4o-mini	0.8080	0.7036	0.4712	0.091	0.5538	0.1748	0.6673	0.6479	0.7104	0.7487	0.6158	0.4195	0.5450	0.7855
Gemini-2-Flash	0.8475	0.6817	0.4881	0.1209	0.5634	0.1759	0.6653	0.6703	0.7045	0.7507	0.6799	0.4592	**0.6455**	**0.8540**
GPT-4o	0.8223	0.6961	0.4780	**0.2133**	0.5790	0.1878	0.6751	0.6298	0.7069	0.7687	0.6618	0.4375	0.5513	0.7676
Claude-3.5-Haiku	0.8149	0.7020	**0.5227**	0.0992	0.5517	0.1616	0.6586	0.5482	0.7085	0.7501	0.6172	0.4412	0.5970	0.7640
Claude-4	0.8252	0.7039	0.4500	0.0860	0.5689	0.2167	0.6707	0.6765	0.6846	0.7449	0.6354	0.5465	0.5128	0.7559
Gemini-2-Pro	0.8424	0.6904	0.4848	0.1177	0.6135	**0.2300**	0.6806	0.6373	0.7214	0.7625	**0.7122**	0.5089	0.5748	0.7722
GPT-o1	0.8396	0.6695	0.4742	0.1163	**0.6194**	0.2269	0.6605	0.5984	0.7127	0.7558	0.6331	0.3691	0.5931	0.7675
GPT-o3-mini	0.8585	0.6831	0.5015	0.1224	0.6109	0.1955	0.6758	0.6585	0.7095	0.7433	0.6712	0.4735	0.5408	0.8013
Claude-3.7-Sonnet	**0.8699**	**0.7365**	0.4665	0.1146	0.5815	0.1999	0.6530	0.6101	0.7139	0.7588	0.6891	0.4883	0.5867	0.7544

Table 1: Baseline performance of LLMs and traditional ML models across 7 clinical prediction tasks on MIMIC-III. Results on MIMIC-IV are provided in Appendix B.1 and B.2. 'Positive Ratio' refers to the proportion of samples labeled as 1. 'ROC' denotes AUROC, and 'PRC' denotes AUPRC. Bold scores indicate the best performance within ML models and LLMs category.

in clinical prediction tasks. In highly imbalanced settings, prompting strategies show some benefits. However, no single strategy consistently improves results across all tasks and models, suggesting their effectiveness depends on the clinical context and model capacity.

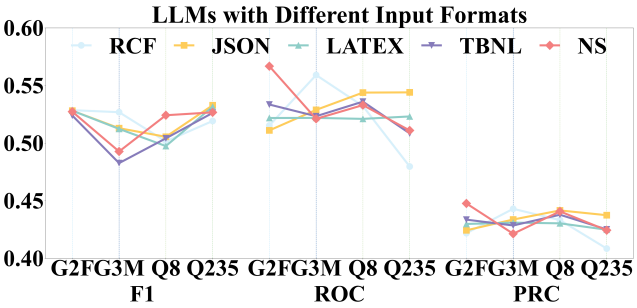

Figure 3: Average performance of LLMs with 5 Input Formats on MIMIC-III. All 3 scores represent the mean values across 7 outcomes. Abbreviations: Gemini-2-Flash (G2F), GPT-o3-mini (G3M), Qwen3-8b (Q8), Qwen3-235b (Q235).

Input Format Sensitivity Evaluation

We analyze the performance of 4 LLMs across 3 metrics using 5 input format strategies (complete results in the Appendix B.4). As shown in Figure 3, the proprietary large model GPT-o3-mini responds best to RCF, with weaker performance on NS and TBNL. The open-source large model Qwen3-235b performs best with JSON, leading across all metrics and outperforming RCF by up to 0.064, but responds least effectively to RCF. Among smaller models, the proprietary Gemini-2-Flash performs best with NS, achieving its highest AUROC and AUPRC, surpassing RCF by 0.051 and 0.026, respectively. The open-source Qwen3-8b shows strong results with both JSON and NS.

Finding for Q3: Different LLMs show varying sensitivity to structured EHR encoding formats. RCF works well for proprietary large models. JSON consistently improves performance for open-source models. LaTeX tables offer limited benefit across models. TBNL produces mixed results, indicating rigid templates require task-specific tuning. NS is particularly effective for smaller models.

Why do different types of LLMs prefer different input formats? This preference likely stems from the interplay between pretraining data, model architecture, and capacity. Proprietary large models tend to favor RCF, possibly due to exposure to structured documents during pretraining. This

aligns with previous study (Li et al. 2024), which suggest LLMs exhibit a bias toward formal data formats. In contrast, open-source large models prefer JSON, as their training data often include public datasets rich in JSON-formatted templates, API references, and structured records, common on platforms like Hugging Face. JSON's key-value structure closely matches the data distribution these models encounter, and a study (Zhu et al. 2024) shows that reinforcing JSON format improves hierarchical parsing. Smaller models perform better with NS, which converts inputs into fluent narrative text. This format reduces reliance on structure parsing and better matches the natural language data these models are trained on, aligning with a previous study showing that smaller models benefit from training-aligned data distributions (Yam and Paek 2024).

Causal Feature Evaluation

In this section, we investigate our core hypothesis (Q4): can causally selected features improve LLM prediction? We evaluate the performance of 6 LLMs using 3 CD feature sets (complete results in the Appendix B.5). As shown in Figure 4, across all LLMs, CD-derived features often lead to performance degradation. Compared to the baseline, DirectLiNGAM leads to the most significant drop. However, CORL yields slight improvements on Gemini-2-Pro, DeepSeek-R1 and Qwen3-235b. GES shows consistent gains on open-source models including DeepSeek-R1, Qwen3-8b and Qwen3-235b.

Contrary to expectations, although some improvements are observed, LLMs leveraging CD features do not consistently outperform the baseline. However, this does not reflect the quality of the CD methods themselves, as ground-truth causal graphs are rarely available in clinical data for validation. In clinical risk prediction, strict assumptions in CD algorithms may result in a limited set of outcome-related causal features, while non-causal but highly correlated features still provide strong predictive signals. Additionally,

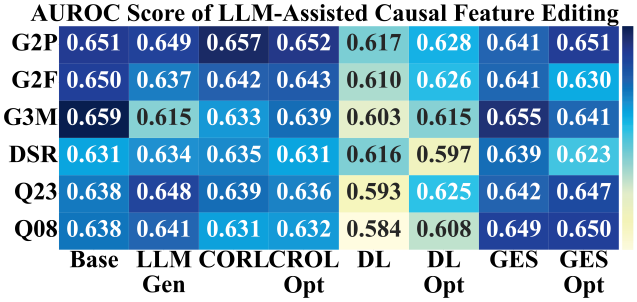

Figure 4: Average AUROC of LLMs across 7 outcomes over MIMIC-III and MIMIC-IV. 'Base' denotes performance using all features. Abbreviations: G2P (Gemini-2-Pro), G2F (Gemini-2-Flash), G3M (GPT-o3-mini), DSR (DeepSeek-R1), Q23 (Qwen3-235b), Q08 (Qwen3-8b). 'LLM Gen' refers to the causal feature set generated by the LLM itself. 'DL' indicates DirectLiNGAM. 'Opt' indicates the optimized feature set refined from the CD outputs by the corresponding LLM.

most CD methods lack prior knowledge and domain-specific constraints, making it difficult to recover complex causal structures from high-dimensional clinical data.

Finding for Q4: Causal feature subsets derived from commonly used CD algorithms did not consistently improve LLMs prediction. This may be due to strict assumptions inherent in many CD methods (e.g., the Causal Faithfulness assumption or the absence of hidden confounders), which are often violated in complex, high-dimensional clinical data. However, this highlights the potential of further exploring whether integrating LLM-derived prior knowledge into CD outcomes can enhance performance (Zhou et al. 2024; Ban et al. 2025).

LLM-Assisted Causal Feature Editing Evaluation

Building on the findings from Q4, we further explore whether LLM-assisted causal features could enhance LLM performance. We evaluate 2 strategies: (1) optimizing causal feature sets derived from CD methods using different LLMs, and (2) allowing LLMs to directly generate causal feature sets based on clinical knowledge. In both cases, the corresponding LLM is used to predict the outcome using its respective feature set (complete results in the Appendix B.5).

Figure 4 shows the average AUROC performance of each model on both MIMIC-III and MIMIC-IV datasets. LLM optimization is key for DirectLiNGAM feature sets. Except for DeepSeek-R1, the optimized DirectLiNGAM features consistently outperform their unoptimized counterparts, though still fall short of the baseline. For CORL and GES, optimization leads to mixed results, with both improvements and declines. Notably, optimized GES achieves further gains on Qwen3-8b and Qwen3-235b, surpassing the baseline. While LLM-optimized CD features do not always exceed baseline performance, they generally improve upon raw CD outputs. LLM-generated causal features often outperform those from CORL and DirectLiNGAM, and show competitive performance with GES. To further analyze this performance, Table 2 provides a breakdown by dataset for representative proprietary and open-source LLMs, showing that LLM-generated causal feature sets and LLM-optimized GES feature sets yield comparable results.

	MIMIC III		MIMIC IV	
Model	AUROC	F1	AUROC	F1
Gemini-2-Pro				
Base	0.6571	0.5471	0.6451	0.5426
GES Opt	0.6403	0.5142	0.6420	0.5479
LLM Gen	0.6559	0.5474	0.6453	0.5673
Qwen3-8b				
Base	0.6368	0.5214	0.6392	0.5456
GES Opt	0.6449	0.4732	0.6537	0.5240
LLM Gen	0.6518	0.4696	0.6481	0.5288

Table 2: Average performance of LLMs across 7 outcomes on MIMIC-III and MIMIC-IV, respectively. 'Base' denotes performance using all features. 'LLM Gen' denotes the causal feature set generated by the LLM itself, while 'Opt' refers to the GES causal features optimized by the LLM.

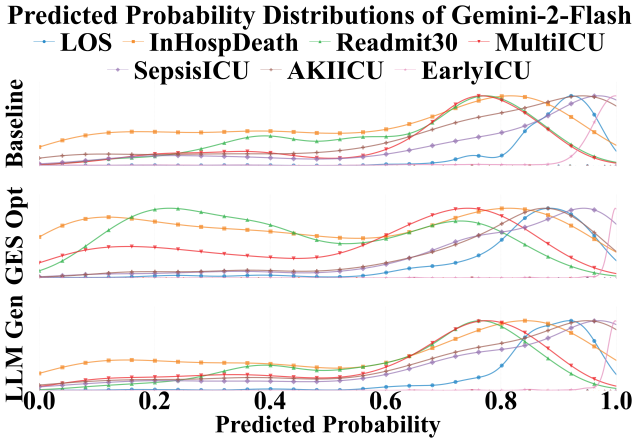

Figure 5: Predicted probability distributions of Gemini-2-Flash across 7 outcomes using 3 different feature sets on MIMIC III: Baseline (all features), Optimized GES (causal features optimized by Gemini-2-Flash), and LLM-Generated (features generated by Gemini-2-Flash).

Figure 5 shows the predicted probability distributions of Gemini-2-Flash using the baseline features, optimized GES features and LLM-generated causal features. For outcomes like Readmit30, MultiICU, and InHospDeath, the LLM-generated features exhibit prediction preferences similar to the baseline, tending toward more severe outcomes. Across all 3 feature sets, the prediction distributions for probabilities above 0.5 are largely consistent.

Finding for Q5: LLMs show potential in optimizing and identifying causal features in clinical settings. This highlights a valuable synergy: CD algorithms provide a structured foundation, while LLMs contribute prior knowledge to refine and enhance feature selection (Darvariu, Hailes, and Musolesi 2024; Zhou et al. 2024). Our results further demonstrate that LLMs can serve as a complementary tool for causal feature selection in clinical prediction tasks, aligning with recent studies showing that LLMs are capable of inferring causal relationships from clinical data (Naik et al. 2024; Kiciman et al. 2023).

Related Works

Clinical decision support is rapidly advancing, driven by the growth of medical data and the shift toward precision medicine. In ICUs, robust and interpretable ML models like LR, XGBoost, and RF are widely used for tasks such as predicting in-hospital mortality (Wang et al. 2025), readmission (Fathy, Emeriaud, and Cheriet 2025), AKI (Lin, Shi, and Kong 2025), and sepsis (Gao et al. 2024), enhancing diagnostic efficiency and enabling early intervention for high-risk patients (Hyland et al. 2020).

Recently, LLMs have shown promise by directly processing unstructured text, preserving critical clinical details often lost in manual feature extraction (Hager et al. 2024; Wang et al. 2024b; Ahmed et al. 2025). Through fine-tuning and RAG, LLMs integrate hospital-specific data with general medical knowledge, improving decision support (Ansari et al. 2025; Jin et al. 2024). Current benchmarks focus on clinical question answering, multimodal integration, and real-world evaluation frameworks (Liu et al. 2024; Budler et al. 2025; Esteitieh, Mandal, and Laliotis 2025).

Meanwhile, causal learning enhances clinical prediction interpretability at both model and data levels. ML may overfit spurious correlations, resulting in unreliable predictions (Huang et al. 2025). In contrast, causal learning aims to uncover the true causal structures, offering more robust and interpretable insights (Feuerriegel et al. 2024; Zhou and Chen 2022; Zanga, Ozkirimli, and Stella 2022). Major causal discovery methods include constraint-based (e.g., PC (Spirtes, Glymour, and Scheines 2000), score-based (e.g., GES (Chickering 2002)), functional causal models (e.g., DirectLiNGAM (Shimizu et al. 2011)), and optimization-based approaches (e.g., CORL (Wang et al. 2021)). Among them, constraint-based methods are limited by reliance on the faithfulness assumption and inability to determine causal directions within Markov equivalence classes(Zhou and Chen 2022).

Integrating LLMs with causality advances clinical prediction by producing reliable causal chains, enhancing counterfactual reasoning, and translating complex causal relationships into clinically meaningful insights (Zeng et al. 2025; Kiciman et al. 2023; Kweon et al. 2024). However, large-scale evaluations of LLM and causal methods specifically for clinical risk prediction remain scarce. Detailed related work can be found in Appendix F.

Conclusion

This study evaluates LLMs with causal features in clinical prognostic tasks. While integrating CD features into LLMs did not yield notable gains in clinical risk prediction, this does not diminish the inherent value of CD. The modest performance is largely due to limitations of current CD algorithms in clinical settings: the strict assumptions that often result in sparse or incomplete feature sets. These challenges highlight the need for caution when applying CD outputs directly to LLM-based tasks without further refinement. Nonetheless, our findings suggest a promising synergy: CD provides a data-driven foundation for causal feature identification, while LLMs contribute rich domain knowledge to enhance feature selection. This aligns with growing evidence that LLMs can assist in uncovering causal relationships in complex biomedical contexts. Moving forward, incorporating LLM-guided priors into the CD process, or using LLMs to refine CD-derived features post hoc, offers a promising path toward more robust clinical prediction models. We also find that ICL improves LLM performance in clinical prediction, especially when prompts are tailored to the task and model. Input format also matters: proprietary models prefer RCF, open-source models perform better with JSON, and smaller models benefit from NS inputs. These results highlight the need to align prompts and data formats with model architecture, capacity, and pretraining.

References

Ahmed, A.; Saleem, M.; Alzeen, M.; Birur, B.; Fargason, R. E.; Burk, B. G.; Harkins, H. R.; Alhassan, A.; and Al-Garadi, M. A. 2025. Leveraging Large Language Models to Enhance Machine Learning Interpretability and Predictive Performance: A Case Study on Emergency Department Returns for Mental Health Patients. *arXiv preprint arXiv:2502.00025*.

Ansari, M. S.; Khan, M. S. A.; Revankar, S.; Varma, A.; and Mokhade, A. S. 2025. Lightweight Clinical Decision Support System using QLoRA-Fine-Tuned LLMs and Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.03406*.

Ban, T.; Chen, L.; Lyu, D.; Wang, X.; Zhu, Q.; and Chen, H. 2025. Llm-driven causal discovery via harmonized prior. *IEEE Transactions on Knowledge and Data Engineering*.

Bedi, P.; Thukral, A.; and Dhiman, S. 2025. XLR-KGDD: leveraging LLM and RAG for knowledge graph-based explainable disease diagnosis using multimodal clinical information. *Knowledge and Information Systems*, 1–21.

Ben Shoham, O.; and Rappoport, N. 2024. Cpllm: Clinical prediction with large language models. *PLOS Digital Health*, 3(12): e0000680.

Brown, K. E.; Yan, C.; Li, Z.; Zhang, X.; Collins, B. X.; Chen, Y.; Clayton, E. W.; Kantarcioglu, M.; Vorobeychik, Y.; and Malin, B. A. 2025. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of the American Medical Informatics Association*, 32(5): 811–822.

Budler, L. C.; Chen, H.; Chen, A.; Topaz, M.; Tam, W.; Bian, J.; and Stiglic, G. 2025. A Brief Review on Benchmarking for Large Language Models Evaluation in Healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2): e70010.

Chen, C.; Yu, J.; Chen, S.; Liu, C.; Wan, Z.; Bitterman, D.; Wang, F.; and Shu, K. 2024. ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction? *arXiv preprint arXiv:2411.06469*.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554.

Darvariu, V.-A.; Hailes, S.; and Musolesi, M. 2024. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*.

Esteitieh, Y.; Mandal, S.; and Laliotis, G. 2025. Towards metacognitive clinical reasoning: Benchmarking mdpie against state-of-the-art llms in medical decision-making. *medRxiv*, 2025–01.

Fathy, W.; Emeriaud, G.; and Cheriet, F. 2025. A comprehensive review of ICU readmission prediction models: From statistical methods to deep learning approaches. *Artificial Intelligence in Medicine*, 103126.

Ferdush, J.; Begum, M.; and Hossain, S. T. 2024. ChatGPT and clinical decision support: scope, application, and limitations. *Annals of Biomedical Engineering*, 52(5): 1119–1124.

Feuerriegel, S.; Frauen, D.; Melnychuk, V.; Schweisthal, J.; Hess, K.; Curth, A.; Bauer, S.; Kilbertus, N.; Kohane, I. S.; and van der Schaar, M. 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4): 958–968.

Fihn, S. D.; Berlin, J. A.; Haneuse, S. J.; and Rivara, F. P. 2024. Prediction Models and Clinical Outcomes—A Call for Papers. *JAMA Network Open*, 7(4): e249640–e249640.

Gao, J.; Lu, Y.; Ashrafi, N.; Domingo, I.; Alaei, K.; and Pishgar, M. 2024. Prediction of sepsis mortality in ICU patients using machine learning methods. *BMC Medical Informatics and Decision Making*, 24(1): 228.

Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9): 2613–2622.

Huang, L.; Dou, Z.; Fang, F.; Zhou, B.; Zhang, P.; and Jiang, R. 2025. Prediction of mortality in intensive care unit with short-term heart rate variability: Machine learning-based analysis of the MIMIC-III database. *Computers in Biology and Medicine*, 186: 109635.

Hyland, S. L.; Faltys, M.; Hüser, M.; Lyu, X.; Gumbsch, T.; Esteban, C.; Bock, C.; Horn, M.; Moor, M.; Rieck, B.; et al. 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature medicine*, 26(3): 364–373.

Jin, M.; Yu, Q.; Shu, D.; Zhang, C.; Fan, L.; Hua, W.; Zhu, S.; Meng, Y.; Wang, Z.; Du, M.; et al. 2024. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kafkas, Ş.; Abdelhakim, M.; Althagafi, A.; Toonsi, S.; Alghamdi, M.; Schofield, P. N.; and Hoehndorf, R. 2025. The application of Large Language Models to the phenotype-based prioritization of causative genes in rare disease patients. *Scientific Reports*, 15(1): 15093.

Kang, Y.; Yang, M.; Peng, Y.; Cai, J.; Zhao, L.; Gao, Z.; Li, N.; and Pu, B. 2025. LLM-DG: Leveraging large language model for enhanced disease prediction via inter-patient and intra-patient modeling. *Information Fusion*, 121: 103145.

Khalifa, M.; Albadawy, M.; and Iqbal, U. 2024. Advancing clinical decision support: The role of artificial intelligence across six domains. *Computer Methods and Programs in Biomedicine Update*, 5: 100142.

Khwaja, A. 2012. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4): c179–c184.

Kiciman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Kweon, S.; Kim, J.; Kwak, H.; Cha, D.; Yoon, H.; Kim, K.; Yang, J.; Won, S.; and Choi, E. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37: 124575–124611.

Li, J.; Cao, Y.; Huang, S.; and Chen, J. 2024. Formality is Favored: Unraveling the Learning Preferences of Large Language Models on Data with Conflicting Knowledge. *arXiv preprint arXiv:2410.04784*.

Lin, Y.; Shi, T.; and Kong, G. 2025. Acute kidney injury prognosis prediction using machine learning methods: a systematic review. *Kidney Medicine*, 7(1): 100936.

Liu, F.; Li, Z.; Zhou, H.; Yin, Q.; Yang, J.; Tang, X.; Luo, C.; Zeng, M.; Jiang, H.; Gao, Y.; et al. 2024. Large language models in the clinic: a comprehensive benchmark. *arXiv preprint arXiv:2405.00716*.

Liu, S.; Wright, A. P.; Patterson, B. L.; Wanderer, J. P.; Turer, R. W.; Nelson, S. D.; McCoy, A. B.; Sittig, D. F.; and Wright, A. 2023. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, 30(7): 1237–1245.

Naik, N.; Khandelwal, A.; Joshi, M.; Atre, M.; Wright, H.; Kannan, K.; Hill, S.; Mamidipudi, G.; Srinivasa, G.; Bifulco, C.; et al. 2024. Applying large language models for causal structure learning in non small cell lung cancer. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 688–693. IEEE.

Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature medicine*, 28(1): 31–38.

Sandmann, S.; Riepenhausen, S.; Plagwitz, L.; and Varghese, J. 2024. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature communications*, 15(1): 2050.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvarinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; Bollen, K.; and Hoyer, P. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr): 1225–1248.

Singer, M.; Deutschman, C. S.; Seymour, C. W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G. R.; Chiche, J.-D.; Coopersmith, C. M.; et al. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, 315(8): 801–810.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3): 943–950.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

Van Smeden, M.; Reitsma, J. B.; Riley, R. D.; Collins, G. S.; and Moons, K. G. 2021. Clinical prediction models: diagnosis versus prognosis. *Journal of clinical epidemiology*, 132: 142–145.

Wang, H.; Gao, C.; Dantona, C.; Hull, B.; and Sun, J. 2024a. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ digital medicine*, 7(1): 16.

Wang, J.; Ahn, S.; Dalal, T.; Zhang, X.; Pan, W.; Zhang, Q.; Chen, B.; Dodge, H. H.; Wang, F.; and Zhou, J. 2024b. Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models. *arXiv preprint arXiv:2405.16413*.

Wang, L.; Guo, X.; Shi, H.; Ma, Y.; Bao, H.; Jiang, L.; Zhao, L.; Feng, Z.; Zhu, T.; and Lu, L. 2025. CRISP: A causal relationships-guided deep learning framework for advanced ICU mortality prediction. *BMC Medical Informatics and Decision Making*, 25(1): 165.

Wang, X.; Du, Y.; Zhu, S.; Ke, L.; Chen, Z.; Hao, J.; and Wang, J. 2021. Ordering-based causal discovery with reinforcement learning. *arXiv preprint arXiv:2105.06631*.

Yam, H. M.; and Paek, N. J. 2024. What should baby models read? Exploring sample-efficient data composition on model performance. *arXiv preprint arXiv:2411.06672*.

Yeh, Y.-C.; Kuo, Y.-T.; Kuo, K.-C.; Cheng, Y.-W.; Liu, D.-S.; Lai, F.; Kuo, L.-C.; Lee, T.-J.; Chan, W.-S.; Chiu, C.-T.; et al. 2024. Early prediction of mortality upon intensive care unit admission. *BMC Medical Informatics and Decision Making*, 24: 394.

Zanga, A.; Ozkirimli, E.; and Stella, F. 2022. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151: 101–129.

Zeng, H.; Yin, C.; Chai, C.; Wang, Y.; Dai, Q.; and Sun, H. 2025. Cancer gene identification through integrating causal prompting large language model with omics data–driven causal inference. *Briefings in Bioinformatics*, 26(2).

Zhang, K.; Zhu, S.; Kalander, M.; Ng, I.; Ye, J.; Chen, Z.; and Pan, L. 2021. gcastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*.

Zheng, X.; Ji, S.; Sun, J.; Chen, R.; Gao, W.; and Srivastava, M. 2025. ProMind-LLM: Proactive Mental Health Care via Causal Reasoning with Sensor Data. *arXiv preprint arXiv:2505.14038*.

Zhou, W.; and Chen, Q. 2022. A survey on causal discovery. In *China Conference on Knowledge Graph and Semantic Computing*, 123–135. Springer.

Zhou, Y.; Wu, X.; Huang, B.; Wu, J.; Feng, L.; and Tan, K. C. 2024. Causalbench: A comprehensive benchmark for causal learning capability of llms. *arXiv preprint arXiv:2404.06349*.

Zhu, T.; Dong, D.; Qu, X.; Ruan, J.; Chen, W.; and Cheng, Y. 2024. Dynamic data mixing maximizes instruction tuning for mixture-of-experts. *arXiv preprint arXiv:2406.11256*.