ClinicalBench: Can LLMs Beat Traditional ML Models in Clinical Prediction?

Canyu Chen*[1], Jian Yu*‡, Shan Chen[3], Che Liu[5], Zhongwei Wan[6], Danielle S. Bitterman[3], Fei Wang[4], Kai Shu†[2]
[1]Illinois Institute of Technology [2]Emory University
[3]Mass General Brigham & Boston Children's Hospital, Harvard Medical School
[4]Weill Cornell Medicine, Cornell University [5]Imperial College London [6]Ohio State University
cchen151@hawk.iit.edu,jian.yu0422@gmail.com,{schen73,dbitterman}@bwh.harvard.edu
che.liu21@imperial.ac.uk,wan.512@osu.edu,few2001@med.cornell.edu,kai.shu@emory.edu

Project website: https://clinicalbench.github.io

Abstract

Large Language Models (LLMs) hold great promise to revolutionize current clinical systems for their superior capacities on medical text processing tasks and medical licensing exams. Meanwhile, traditional ML models such as SVM and XGBoost have still been mainly adopted in clinical prediction tasks. An emerging question is *Can LLMs beat traditional ML models in clinical prediction?* Thus, we build a new benchmark ClinicalBench to comprehensively study the clinical predictive modeling capacities of both general-purpose and medical LLMs, and compare them with traditional ML models. ClinicalBench embraces three common clinical prediction tasks, two databases, 14 general-purpose LLMs, 8 medical LLMs, and 11 traditional ML models. Through extensive empirical investigation, we discover that **both general-purpose and medical LLMs, even with different model scales, diverse prompting or fine-tuning strategies, still cannot beat traditional ML models in clinical prediction yet**, shedding light on their potential deficiency in clinical reasoning and decision-making. We call for caution when practitioners adopt LLMs in clinical applications. ClinicalBench can be utilized to bridge the gap between LLMs' development for healthcare and real-world clinical practice.

1 Introduction

Large Language Models (LLMs) have shown great potential to revolutionize existing clinical systems for their superior capacities on a variety of medical text processing tasks including document classification, report generation and summarization, and information extraction (Jahan et al., 2024; Chen et al., 2023a; Zhou et al., 2023). Also, LLMs could even rival human clinician performance in medical licensing exams (Singhal et al., 2023a,b; Gilson et al., 2023) and clinical case challenges (Kanjee et al., 2023; McDuff et al., 2023). However, traditional machine learning (ML) models such as Logistic Regression, SVM, XGBoost, MLP and Transformer have still been predominantly adopted for clinical prediction tasks in real-world practice (van de Water et al., 2024; Moazemi et al., 2023), which play an essential role in modern health systems to improve patient outcomes and support clinical decision-making. Thus, considering LLMs' profound world knowledge and strong reasoning capacities, an emerging question is: ***Can LLMs beat traditional ML models in clinical prediction?***

In this paper, we propose to build a new benchmark ClinicalBench to comprehensively and systematically investigate the capacities of LLMs in clinical prediction tasks. More importantly, we compare the performance of LLMs with diverse traditional ML models in a head-to-head way to explore the feasibility of adopting LLMs in real-world clinical prediction. Specifically, following previous works (Wang et al., 2020; van de Water et al., 2024), we adopt three common clinical prediction tasks

*Equal Contribution. ‡Independent researcher. †Corresponding author.

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

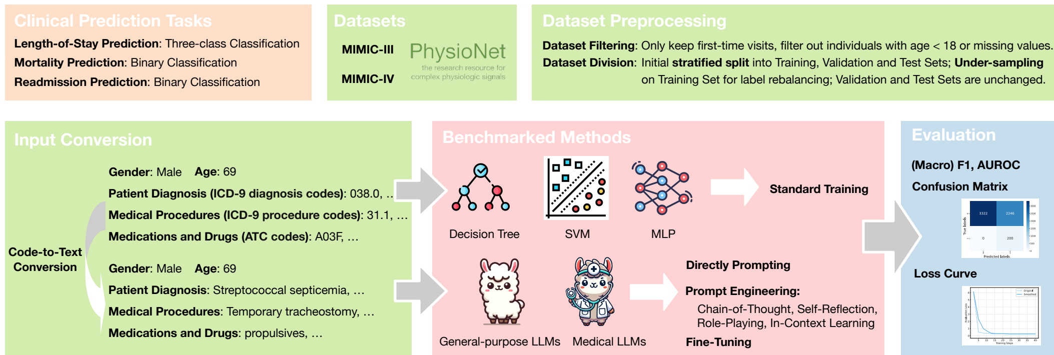


Figure 1: **Overview of** ClinicalBench.

including Length-of-Stay Prediction, Mortality Prediction and Readmission Prediction and two real clinical databases MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2020). As for the benchmarked methods, ClinicalBench embraces 11 traditional ML models and 22 LLMs with different scales, which include 14 *general-purpose LLMs* and 8 *medical LLMs*.

In ClinicalBench, we aim to answer three research questions through extensive empirical studies. The first question is *Can directly prompting LLMs beat traditional ML models?* As shown in Figure 1, to effectively leverage the language understanding capacities of LLMs, we first convert the original clinical codes in datasets to texts and then let LLMs output the predicted label directly. Comparing the performance of both 9 general-purpose LLMs and 6 medical LLMs with around 7B parameters to traditional ML models, we discover that **traditional ML models can mostly outperform LLMs with a moderate scale by a large margin across three tasks and two datasets**. Comparing general-purpose and medical LLMs, we also find that the **medical LLMs do not noticeably surpass general-purpose LLMs with a similar scale in clinical prediction**. We further studied the impact of *decoding temperature* and *model size* on LLMs' predictive modeling capacities. We find that these factors could potentially influence the clinical prediction performance in a negative or positive way for different LLMs. However, **adjusting temperatures or scaling parameters of LLMs still may not reach the performance of typical traditional ML models such as XGBoost, SVM, and RNN**.

The second question is *Can LLMs with prompt engineering beat traditional ML models?* We studied the efficacy of four common prompting strategies including Zero-shot Chain-of-Thought (Kojima et al., 2022), Self-Reflection (Pan et al., 2023), Role-Playing (Chen et al., 2024b), and In-Context Learning (Dong et al., 2023) on both general-purpose LLMs and medical LLMs. In general, we find that **the effectiveness of different prompting strategies is very limited**. Except the improvement brought by In-Context Learning on Llama3-8B, Gemma2-9B, and Internist-7B for Length-of-Stay Prediction, the aforementioned prompting strategies do not enhance the performance of LLMs in clinical prediction across three tasks and two datasets in a noticeable way.

The third question is *Can fine-tuned LLMs beat traditional ML models?* We compare the performance of traditional ML models and LLMs with the same dataset split on training, validation and test sets. With thorough investigation over 2 fine-tuning strategies, 4 LLMs, 3 clinical prediction tasks and 2 datasets, we find that **fine-tuning is clearly effective for certain tasks though the effectiveness varies across different models**. As for Length-of-Stay Prediction and Mortality Prediction, we observe that fine-tuning can evidently enhance the clinical prediction performance of LLMs though the extent of improvement depends on models. As for Readmission Prediction, we do not notice fine-tuning can bring any enhancement. However, **most fine-tuned LLMs still cannot surpass typical traditional ML models such as XGBoost, SVM, Transformer and RNN**.

Overall, the contributions of this paper can be summarized as follows:

- We built a new benchmark ClinicalBench, embracing 14 general-purpose LLMs, 8 medical LLMs, 11 traditional ML models, three tasks and two databases, which made the first attempt to compare the clinical prediction capacities of LLMs and traditional ML models in a head-to-head way.
- We discover that both general-purpose and medical LLMs, even with different model sizes, prompting or fine-tuning strategies, still cannot beat traditional ML models in clinical prediction yet.
- Our findings demonstrate **the potential deficiency of both general-purpose and medical LLMs in real-world clinical reasoning and decision-making**, which could have almost clinician-level performance in medical licensing exams and clinical case challenges. We call for caution when

adopting LLMs in practical clinical applications. ClinicalBench could be leveraged to bridge the gap between the development of LLMs for healthcare and real-world clinical practice.

2 ClinicalBench: Benchmarking LLMs and Traditional ML Models in Clinical Prediction

Clinical Prediction Tasks Following previous works (Wang et al., 2020; van de Water et al., 2024), we focus on three common tasks including *Length-of-Stay Prediction*, *Mortality Prediction*, and *Readmission Prediction*. Specifically, *Length-of-Stay Prediction* is simplified as a three-class classification task aiming to predict the length of the current hospital visit as less than one week, 1 ~ 2 weeks, or more than two weeks. *Mortality Prediction* is a binary classification task intending to estimate whether the patient will decease in the current visit. *Readmission Prediction* is a binary classification task designed to identify patients who are at high risk of being readmitted to the hospital within a specific time frame. Following PyHealth (Yang et al., 2023), each task requires models to make predictions based on patients' demographic features (*e.g.*, gender and age) and clinical information including diagnosis, medical procedures, medications and drugs for the current visit.

Dataset Processing We adopt MIMIC-III-v1.4 (Johnson et al., 2016) and MIMIC-IV-v2.2 (Johnson et al., 2020) in ClinicalBench, which are two large and freely accessible databases and widely adopted in the fields of healthcare data science. They contain de-identified data associated with ICU admissions at the Beth Israel Deaconess Medical Center during 2001-2012 and 2008-2019 respectively. More details of dataset processing are as follows: **(1) Dataset Filtering** We generally follow the filtering criteria in previous literature (Yang et al., 2023; Purushotham et al., 2018; Wang et al., 2020; Tang et al., 2020). First, we only keep the samples of the first-time visits for each patient since other visits require the consideration of patients' history, which desire different prediction strategies compared with first-time visits. Second, we filter out patients who are younger than 18 due to the substantial differences between pediatric and adult physiology. Third, we also did not consider patients with missing values following the literature. **(2) Dataset Division** For traditional ML models, we first adopt **stratified split** to divide the original dataset into *training*, *validation* and *test* sets. Due to the label imbalance of *training* set for three tasks, we apply **under-sampling** to rebalance the training set but leave the *validation* and *test* sets unchanged, which can reflect the real-world clinical prediction performance of different models. For prompting-based methods, we directly assess the performance of LLMs on the same *test* set. For fine-tuning based methods, the division is the same as traditional ML models for fair comparison. **(3) Code-to-Text Conversion** It is worth noting that the information of patient diagnosis, medical procedures, medications and drugs in original MIMIC-III (MIMIC-IV) dataset is stored in the form of ICD-9 (ICD-10) diagnosis codes, ICD-9 (ICD-10) procedure codes and ATC codes respectively. To effectively leverage the natural language understanding and reasoning capacities of LLMs, as shown in Figure 1, we first convert the codes into texts and then design the specific prompts for LLMs.

Benchmarked Methods We compare mainstream traditional ML models and LLMs in ClinicalBench. The traditional ML models embrace XGBoost, Logistic Regression, Decision Tree, Random Forest, AdaBoost, SVM, Naive Bayes and neural network models such as MLP, Transformer and RNN. For LLMs, we first investigated directly prompting both general-purpose and medical LLMs with different decoding temperatures and model scales. Then, we explored the effectiveness of diverse prompting and fine-tuning strategies. The examples of specific prompts for the different strategies are in Appendix D. It is worth noting that we only adopt open-source LLMs since closed-source models such as GPT-4 are prohibited for MIMIC-III and MIMIC-IV according to their data use policy [2].

3 Can Directly Prompting LLMs Beat Traditional ML Models?

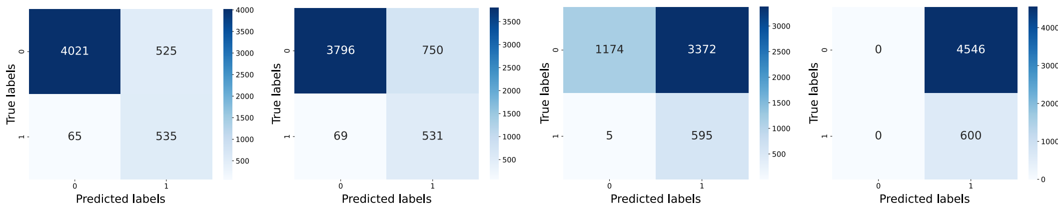
In this section, we aim to investigate the question *Can directly prompting LLMs beat traditional ML models?* We first compare the performance of 9 general-purpose and 6 medical LLMs with around 7B parameter scale to a variety of traditional ML models. Then, we further explore the impact of decoding temperatures and parameter scaling on LLMs' clinical prediction performance.

Main Results As shown in Table 1, red and orange represent the performance regarding (Macro) F1% and AUROC% respectively. We can clearly observe that the shades of both red and orange for traditional ML models are darker than those for LLMs. It shows that **traditional ML models generally outperform both general-purpose and medical LLMs with a moderate scale in clinical prediction**. Furthermore, we have calculated the ranges of performance with 95% Confidence Interval

[2] https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	**AUROC (%)** 95% CI	**F1 (%)** 95% CI	**AUROC (%)** 95% CI	**F1 (%)** 95% CI	**AUROC (%)** 95% CI
Majority	23.37 (23.37, 23.37)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)
Minority	10.72 (10.72, 10.72)	50.00 (50.00, 50.00)	20.88 (20.88, 20.88)	50.00 (50.00, 50.00)	13.64 (13.64, 13.64)	50.00 (50.00, 50.00)
			Traditional ML Models			
XGBoost	67.94 (67.87, 68.01)	93.83 (93.78, 93.88)	65.75 (65.56, 65.94)	95.97 (95.93, 96.01)	19.92 (19.75, 20.09)	69.24 (68.75, 69.73)
LR	66.52 (66.43, 66.61)	93.09 (92.99, 93.19)	63.09 (62.96, 63.22)	94.59 (94.53, 94.65)	19.88 (19.73, 20.03)	69.19 (68.85, 69.53)
DecisionTree	59.14 (59.04, 59.24)	85.22 (85.14, 85.30)	52.17 (52.05, 52.29)	84.18 (83.93, 84.43)	16.1 (15.94, 16.26)	57.57 (57.05, 58.09)
RandomForest	61.14 (61.09, 61.19)	91.23 (91.15, 91.31)	54.23 (54.10, 54.36)	91.65 (91.56, 91.74)	18.78 (18.64, 18.92)	69.3 (69.04, 69.56)
AdaBoost	66.6 (66.48, 66.72)	92.91 (92.85, 92.97)	61.19 (60.99, 61.39)	93.62 (93.54, 93.70)	17.92 (17.75, 18.09)	64.51 (64.20, 64.82)
SVM	70.28 (70.20, 70.36)	94.77 (94.71, 94.83)	63.97 (63.81, 64.13)	95.69 (95.65, 95.73)	20.17 (20.03, 20.31)	71.74 (71.36, 72.12)
NaiveBayes	42.93 (42.84, 43.02)	61.55 (61.43, 61.67)	23.43 (23.41, 23.45)	56.85 (56.80, 56.90)	16.86 (16.77, 16.95)	58.76 (58.56, 58.96)
KNN	44.27 (44.20, 44.34)	73.72 (73.59, 73.85)	44.52 (44.37, 44.67)	85.39 (85.34, 85.44)	15.09 (14.91, 15.27)	59.45 (58.73, 60.17)
MLP	63.54 (63.47, 63.61)	90.61 (90.52, 90.70)	59.32 (59.16, 59.48)	93.01 (92.92, 93.10)	18.86 (18.69, 19.03)	67.47 (67.22, 67.72)
Transformer	63.56 (63.45, 63.67)	89.58 (89.50, 89.66)	57.38 (57.06, 57.70)	93.35 (93.26, 93.44)	18.97 (18.81, 19.13)	67.27 (67.01, 67.53)
RNN	64.23 (64.12, 64.34)	90.08 (90.01, 90.15)	59.03 (58.74, 59.32)	91.97 (91.89, 92.05)	19.40 (19.26, 19.54)	69.32 (69.12, 69.52)
			General-purpose Large Language Models			
Llama3-8B	25.78 (25.72, 25.84)	83.82 (83.74, 83.90)	25.81 (25.78, 25.84)	85.40 (85.33, 85.47)	13.64 (13.64, 13.64)	43.00 (42.50, 43.50)
Mistral-v0.3-7B	11.59 (11.57, 11.61)	69.73 (69.64, 69.82)	38.11 (38.06, 38.16)	86.84 (86.80, 86.88)	13.44 (13.38, 13.50)	47.16 (46.69, 47.63)
Gemma2-9B	17.30 (17.26, 17.34)	79.19 (79.08, 79.30)	43.03 (42.96, 43.10)	86.46 (86.41, 86.51)	13.66 (13.66, 13.66)	47.14 (46.87, 47.41)
Qwen2-7B	12.17 (12.14, 12.20)	73.97 (73.82, 74.12)	34.98 (34.81, 35.15)	85.97 (85.91, 86.03)	2.14 (1.88, 2.40)	44.04 (43.61, 44.47)
Yi-v1.5-9B	10.89 (10.88, 10.90)	74.51 (74.41, 74.61)	28.75 (28.72, 28.78)	76.05 (75.96, 76.14)	11.35 (11.23, 11.47)	45.02 (44.76, 45.28)
Vicuna-v1.5-7B	22.05 (22.01, 22.09)	73.20 (73.11, 73.29)	20.91 (20.91, 20.91)	77.90 (77.80, 78.00)	13.64 (13.64, 13.64)	39.51 (39.18, 39.84)
Phi3.5-mini-3.8B	13.17 (13.14, 13.20)	73.67 (73.63, 73.71)	26.45 (26.43, 26.47)	80.37 (80.30, 80.44)	12.40 (12.30, 12.50)	46.64 (46.37, 46.91)
InternLM2.5-7B	11.53 (11.52, 11.54)	82.38 (82.28, 82.48)	33.89 (33.57, 34.21)	83.89 (83.81, 83.97)	0.00 (0.00, 0.00)	45.84 (45.51, 46.17)
MiniCPM3-4B	22.25 (22.21, 22.29)	75.95 (75.86, 76.04)	24.01 (24.00, 24.02)	86.28 (86.21, 86.35)	13.36 (13.31, 13.41)	42.23 (41.95, 42.51)
			Medical Large Language Models			
Meditron-7B	33.43 (33.39, 33.47)	76.63 (76.51, 76.75)	20.88 (20.88, 20.88)	77.26 (77.20, 77.32)	13.64 (13.64, 13.64)	44.93 (44.43, 45.43)
Medllama3-8B	14.88 (14.85, 14.91)	78.27 (78.22, 78.32)	24.12 (24.11, 24.13)	78.74 (78.66, 78.82)	13.64 (13.64, 13.64)	49.80 (49.24, 50.36)
BioMistral-7B	21.96 (21.89, 22.03)	40.94 (40.73, 41.15)	30.53 (30.45, 30.61)	77.88 (77.81, 77.95)	13.37 (13.28, 13.46)	40.93 (40.54, 41.32)
Med42-8B	12.82 (12.80, 12.84)	84.76 (84.70, 84.82)	28.02 (27.98, 28.06)	82.55 (82.46, 82.64)	13.62 (13.60, 13.64)	42.95 (42.56, 43.34)
BioMedGPT-7B	14.11 (14.11, 14.11)	61.54 (61.45, 61.63)	22.55 (22.54, 22.56)	78.93 (78.82, 79.04)	13.72 (13.65, 13.79)	42.89 (42.67, 43.11)
Internist-7B	11.71 (11.68, 11.74)	81.10 (81.04, 81.16)	37.97 (37.92, 38.02)	87.25 (87.20, 87.30)	12.07 (11.90, 12.24)	47.67 (47.36, 47.98)

Table 1: **Performance of LLMs and Traditional ML Models on Typical Clinical Prediction Tasks**, which include Length-of-Stay Prediction, Mortality Prediction, and Readmission Prediction. Experiments are conducted in **MIMIC-III** dataset. More experiment results on **MIMIC-IV** dataset are in **Appendix B.1**. More experiment results of traditional ML models on different scales of training set are in **Appendix B.2**. (Macro) F1% and AUROC% are adopted as metrics considering label imbalance. Average performance over five runs is reported. The darker shades of red and orange indicate that the scores are closer to 100%. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.



(a) XGBoost (b) Transformer (c) Llama3-8B (d) Meditron-7B

Figure 2: **Examples of Confusion Matrix of LLMs and Traditional ML Models on Mortality Prediction**. Experiments are conducted in MIMIC-III dataset. The complete confusion matrices across different methods, tasks and datasets are in **Appendix B.4**.

through the five-run results, which are shown as the numbers in bracket in Table 1. The lower bounds of traditional ML models' ranges are still mostly larger than the higher bounds of LLMs' ranges, indicating the **high statistical significance of the advantage of traditional ML models over LLMs**. In addition, we have evaluated the performance of traditional ML models with 5%, 10%, 20%, 40% of the original training set in Appendix B.2 and their performance is relatively stable and still surpasses LLMs, which shows **the advantage of traditional ML models in low-resource scenarios**.

Also, we notice that there is generally no notable difference between the performances of general-purpose and medical LLMs, and some medical LLMs can even underperform their general-purpose counterparts. For example, Med42-8B (Christophe et al., 2024) is adapted from Llama3-8B through fine-tuning on medical corpus and can outperform Llama3-8B in various medical knowledge benchmarks such as MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) (The model download links are in Appendix A). However, there is a clear drop for Length-of-Stay Prediction and no substantial difference for the other two tasks when comparing Med42-8B to Llama3-8B regarding their clinical prediction performance, which illustrates that **adapting general-purpose LLMs to medical domain may not improve and could even hurt their clinical reasoning capacities**.

It is noteworthy that the performance of some LLMs (*e.g.*, Llama3-8B and Meditron-7B for Mortality Prediction) is even comparable to that of "Majority" or "Minority" method, which refers to selecting the "Majority" or "Minority" class as the predicted label directly, which further shows **LLMs' deficiency in some clinical prediction tasks**. In more detail, we also explore the confusion matrices of the predictions of both traditional ML models and LLMs. As shown in Figure 2, we can see that **the prediction patterns of LLMs could be distinct from those of traditional ML models**. With under-sampling on the training set, traditional ML models can generally balance precision and recall. However, the predictions of LLMs could be overly biased. For example, Llama3-8B and Meditron-7B have a high rate of false positives, which results in the poor performance on precision and F1.

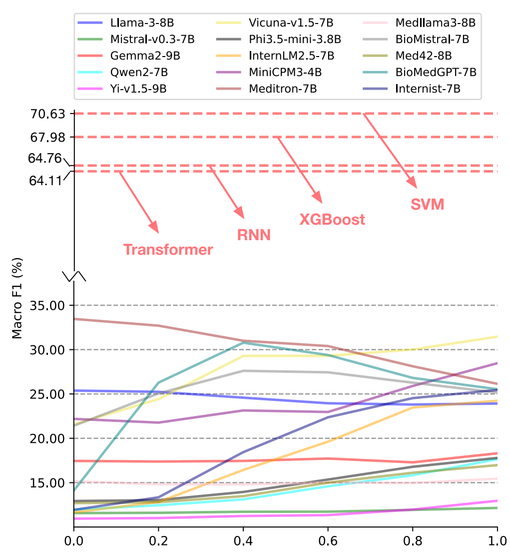


Figure 3: **Performance Comparison Between LLMs with Different Temperatures and Traditional ML Models on Length-of-Stay Prediction.** Experiments are conducted in MIMIC-III dataset. More results on Mortality and Readmission Prediction are in **Appenidix B.3**.

The Impact of Decoding Temperature Since decoding temperature is a key hyperparameter and could have a considerable impact on LLMs' reasoning capacities (Qiu et al., 2024; Renze & Guven, 2024), we explore whether adjusting temperatures can enhance LLMs' clinical prediction performance. As shown in Figure 3, when the temperature increases, **the impact on LLMs' clinical prediction performance could be positive or negative for different LLMs**. For example, when the temperature is higher, the performance of Internist-7B constantly increases but that of Meditron-7B keeps decreasing. However, **only adjusting the decoding temperature of LLMs cannot reach the performance of typical traditional ML models such as XGBoost, SVM, RNN and Transformer**.

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	MIMIC-III	MIMIC-IV	MIMIC-III	MIMIC-IV	MIMIC-III	MIMIC-IV
Traditional ML Models						
XGBoost	67.98	64.19	64.46	48.15	19.33	27.73
SVM	70.63	66.43	63.10	46.27	18.85	28.27
Transformer	64.11	54.97	56.46	42.81	17.80	26.40
RNN	64.76	59.89	56.32	48.9	19.96	25.58
General-purpose Large Language Models						
Llama3-8B	25.38	17.73	26.06	15.12	13.64	25.05
Llama3-70B	18.66	23.01	30.33	18.43	12.40	22.81
Qwen2-0.5B	5.11	5.78	1.94	3.11	0.00	0.60
Qwen2-1.5B	23.37	29.55	20.91	6.88	13.72	24.82
Qwen2-7B	11.96	6.48	35.65	43.53	4.26	0.29
Yi-v1.5-6B	25.81	30.18	28.35	18.26	13.69	24.66
Yi-v1.5-9B	10.94	6.96	28.82	16.55	10.74	20.73
Yi-v1.5-34B	18.92	18.42	41.78	33.63	13.46	24.12
Medical Large Language Models						
Meditron-7B	33.46	26.90	20.88	6.70	13.64	24.92
Meditron-70B	30.15	35.67	43.16	47.93	8.67	9.12
Med42-8B	12.69	9.97	28.59	16.79	13.59	25.06
Med42-70B	15.43	22.95	42.96	31.21	12.99	23.87

Table 2: **Performance Comparison Between LLMs with Different Scales and Traditional ML Models on Clinical Prediction Tasks**. Experiments are conducted in **MIMIC-III** and **MIMIC-IV** datasets. Considering label imbalance, Length-of-Stay Prediction adopts **Macro F1**% and the other two tasks use **F1**% as the metric. The darker shades of red indicate that the scores are closer to 100%.

The Impact of Parameter Scaling in LLMs LLMs with more parameters in the same model series generally perform better across different tasks owing to deeper world knowledge and stronger reasoning capacities. For example, the performance on diverse datasets such as MMLU (Hendrycks et al., 2021) and AGIEval (Zhong et al., 2023) shows a steady growth as scales expand for Yi series models (Young et al., 2024). Thus, after evaluating the performance of LLMs with a around 7B parameter size in Table 1, we explored the impact of parameter scaling on LLMs' clinical prediction performance and investigated whether LLMs with a larger scale could beat traditional ML models.

As shown in Table 2, we have assessed three general-purpose model series including Llama3 (Dubey et al., 2024), Qwen2 (Yang et al., 2024), and Yi-v1.5 (Young et al., 2024) and two medical model series including Meditron (Chen et al., 2023b) and Med42 (Christophe et al., 2024). Comparing LLMs with different scales in the same model series, we can observe that **the parameter scaling does not necessarily lead to better clinical prediction performance**. In Mortality Prediction, LLMs with a larger parameter scale in the same model series tend to perform better. For example, the performance of Qwen2 series models consistently increases on MIMIC-III and MIMIC-IV datasets as the parameter scales grow. However, this tendency does not appear in Length-of-Stay Prediction and Readmission Prediction. Qwen2-1.5B outperforms Qwen2-0.5B and Qwen2-7B for Length-of-Stay Prediction and Readmission Prediction. We also notice that **parameter scaling could even hurt the clinical prediction performance**. For example, Yi-v1.5-6B performs much better than Yi-v1.5-9B and Yi-v1.5-34B for Length-of-Stay Prediction on both MIMIC-III and MIMIC-IV datasets.

However, through the shades of red color, we can clearly see that even though the performance of some LLMs is improved for certain tasks **as the parameter scales expand, they still underperform typical traditional ML models such as XGBoost, SVM, RNN and Transformer**. For example, although the performance is substantially enhanced for Meditron-70B compared to Meditron-7B in Mortality Prediction, there is still a large gap from traditional ML models.

Finding 1: Directly prompting general-purpose and medical LLMs, even with different decoding temperatures or parameter scales, cannot beat traditional ML models in clinical prediction yet.

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	MIMIC-III	MIMIC-IV	MIMIC-III	MIMIC-IV	MIMIC-III	MIMIC-IV
Traditional ML Models						
XGBoost	63.88	57.25	48.04	40.00	18.80	26.95
SVM	66.16	59.08	56.47	36.59	18.48	26.52
Transformer	61.49	53.39	51.85	29.13	17.32	23.21
RNN	64.26	60.09	58.96	49.80	20.80	26.24
General-purpose Large Language Models						
Llama3-8B	26.36	16.14	25.84	14.35	13.78	25.40
+ CoT	14.76	9.22	23.18	10.81	12.05	22.90
+ Self-Reflection	25.88	17.53	24.10	14.29	15.70	22.80
+ Role-Playing	27.44	17.30	25.44	14.11	13.78	24.87
+ ICL	51.77	54.72	28.02	14.85	11.97	24.46
Gemma2-9B	16.22	21.66	44.78	32.99	13.78	23.32
+ CoT	19.34	17.65	22.35	9.94	13.78	24.38
+ Self-Reflection	16.67	17.89	0.00	0.00	13.96	24.07
+ Role-Playing	16.29	20.24	44.19	41.03	13.83	23.74
+ ICL	44.13	46.29	23.92	12.45	14.68	23.95
Medical Large Language Models						
Meditron-7B	35.88	27.53	20.79	6.58	13.78	24.87
+ CoT	16.04	12.58	13.90	5.95	12.83	24.34
+ Self-Reflection	18.36	10.40	20.47	6.60	13.78	24.91
+ Role-Playing	27.73	17.54	20.79	6.58	13.78	24.87
+ ICL	14.09	14.62	23.72	11.37	11.64	23.27
Internist-7B	11.00	14.14	38.57	29.06	11.54	17.80
+ CoT	22.86	17.03	21.67	8.86	14.26	25.48
+ Self-Reflection	23.57	15.09	21.23	10.49	12.16	21.36
+ Role-Playing	10.87	12.61	25.44	15.38	14.42	23.08
+ ICL	47.08	34.33	31.15	16.59	11.59	22.86

Table 3: **Performance Comparison Between LLMs with Prompt Engineering and Traditional ML Models on Clinical Prediction Tasks**. Experiments are conducted in sampled **MIMIC-III** and **MIMIC-IV** datasets. Length-of-Stay Prediction adopts **Macro F1**% and the other two tasks use **F1**% as the metric for label imbalance. The darker shades of red indicate the scores are closer to 100%.

4 Can LLMs with Prompting Engineering Beat Traditional ML Models?

It has been shown that different prompting engineering techniques can exert a notable influence on LLMs' performance across various tasks (Schulhoff et al., 2024; Sahoo et al., 2024). Thus, we investigate the effectiveness of different prompting methods on both general-purpose and medical LLMs for clinical prediction, and also compare them with traditional ML models in this section.

We studied four typical prompting engineering techniques including Zero-shot Chain-of-Thought (Kojima et al., 2022), Self-Reflection (Pan et al., 2023), Role-Playing (Chen et al., 2024b), and In-Context Learning (Dong et al., 2023). The examples of the specific prompt design for different strategies across three tasks are shown in Appendix D. As shown in Table 3, we can observe that **the effectiveness of different prompting strategies is generally very limited though it may vary across models and tasks**. Except that In-Context Learning can explicitly enhance the performance of Llama3-8B, Gemma2-9B, and Internist-7B on Length-of-Stay Prediction, other strategies do not bring a notable improvement and could even degrade the original performance. More importantly, through the shades of red color, we can clearly see that **traditional ML models such as XGBoost, SVM, RNN and Transformer still outperform LLMs with different prompting strategies**.

Finding 2: The effectiveness of typical prompting engineering techniques is generally limited and they still cannot make general-purpose and medical LLMs beat traditional ML models yet.

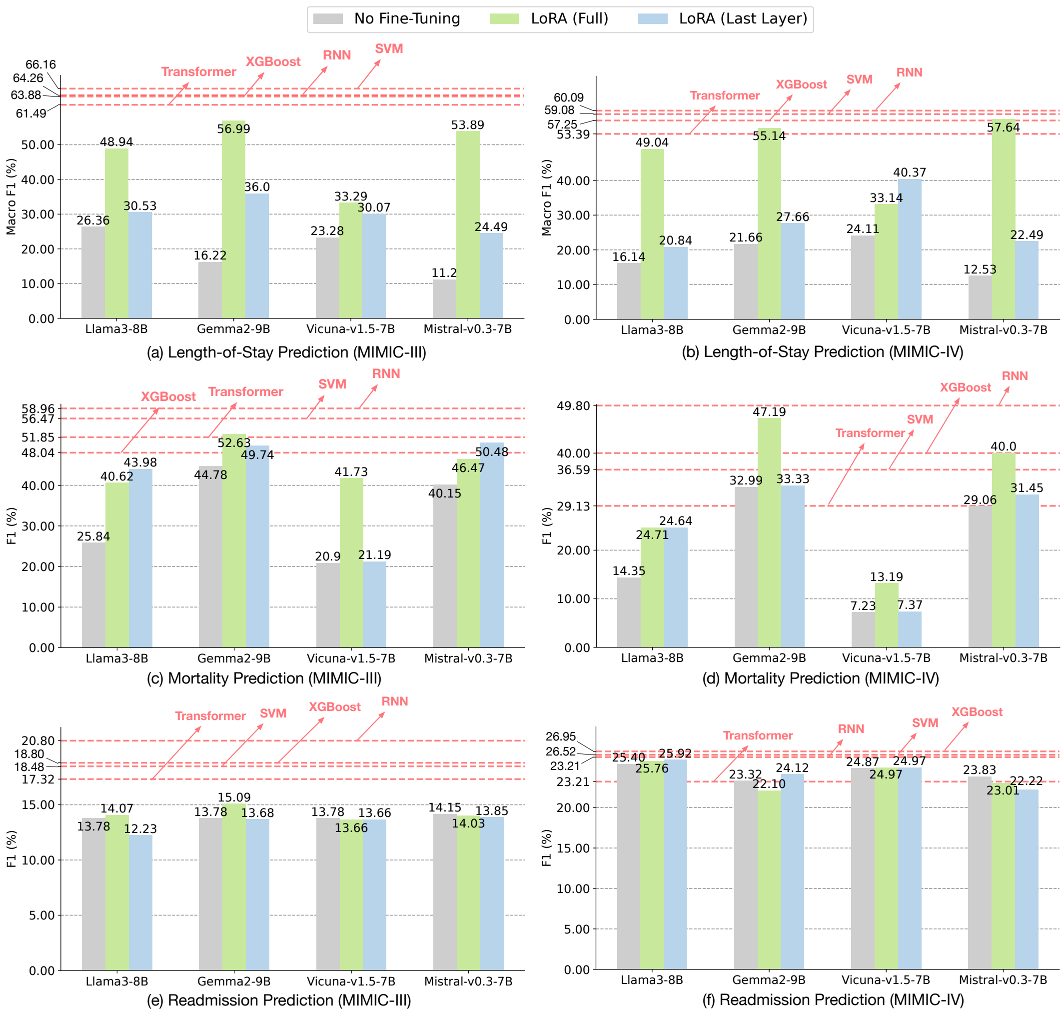


Figure 4: **Performance Comparison Between Fine-tuned LLMs and Traditional ML Models on Clinical Prediction Tasks**. Experiments are conducted on sampled MIMIC-III and MIMIC-IV datasets. Length-of-Stay Prediction adopts **Macro F1**% and the others use **F1**% as the metric. **LoRA (Full)** and **LoRA (Last Layer)** refer to applying LoRA to full layers and only last layer respectively.

5 Can Fine-Tuned LLMs Beat Traditional ML Models?

Fine-tuning is a common paradigm to adapt LLMs to specific tasks or domains (Zhang et al., 2023; Lou et al., 2024). Our results on medical LLMs in Table 1 have shown that only adapting general-purpose LLMs to medical domain may not improve and could even degrade the clinical reasoning capacities. In this section, we explored whether directly adapting general-purpose LLMs to clinical prediction tasks can enhance their performance and also compare them with traditional ML models.

For a fair comparison, we adopted the same dataset division of *training*, *validation* and *test* sets for fine-tuning LLMs and training traditional ML models. The details of fine-tuning data construction for the three clinical prediction tasks are in Appendix C. Considering the relatively small scale of the *training* set, we adjusted the original LoRA (Hu et al., 2022) into two fine-tuning strategies named LoRA (Full) and LoRA (Last Layer). We set the training epochs for both LLMs and traditional ML models as 20. The loss curves of fine-tuning in Appendix B.5 show that LLMs can mostly converge within 20 epochs and we leveraged the *validation* set to select the optimal fine-tuned checkpoint.

As shown in Figure 4, we can observe that **fine-tuning can clearly enhance the clinical prediction performance of LLMs on Length-of-Stay Prediction and Mortality Prediction**, though it may not improve the performance on Readmission Prediction. More specifically, LoRA (Full) brings a more evident improvement than LoRA (Last Layer) across different LLMs on Length-of-Stay

Prediction. The enhancement from fine-tuning on Mortality Prediction is generally less notable than that on Length-of-Stay Prediction. However, **most fine-tuned LLMs still cannot surpass typical traditional ML models such as XGBoost, SVM, RNN and Transformer**. In particular, even with the substantial improvement brought by fine-tuning on Length-of-Stay Prediction, there is still an explicit gap between the performance of most LLMs and typical traditional ML models. Although the fine-tuned Gemma2-9B with LoRA (Full) has been demonstrated to surpass three traditional ML models on MIMIC-IV for Mortality Prediction, the same fine-tuning strategy cannot bring a similar improvement to Gemma2-9B on MIMIC-III for Mortality Prediction.

Finding 3: Fine-tuning can clearly enhance LLMs' performance on certain clinical prediction tasks but most fine-tuned LLMs still cannot beat traditional ML models yet.

6 Implications and Hypothesis

It has attracted increasing attention to transform the existing healthcare systems with LLMs (Zhou et al., 2023, 2024; Liu et al., 2024b; Wang et al., 2024a). While the high performance in medical licensing exams and question-answering benchmarks such as MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) has shown that LLMs may have clinician-level general medical knowledge, it is still highly under-explored whether they could perform clinical reasoning and decision-making in real-world complex scenarios. ClinicalBench made the first attempt to systematically and comprehensively benchmark the clinical predictive modeling capacities of both general-purpose and medical LLMs based on three real-world clinical prediction tasks. By comparing their performance with traditional ML models, we have shed light on LLMs' potential deficiency in performing real-world clinical reasoning and decision-making, implying their **safety and reliability risks in practical adoption** and **the challenges of integrating LLMs into clinical workflows**. Our discoveries also echo the recent discussions on LLMs' potential limitations (Hager et al., 2024; Han et al., 2024; Gallifant et al., 2024; Jin et al., 2024; Wang et al., 2024d). We hypothesize that LLMs' limited clinical predictive modeling abilities could be attributed to **the lack of realistic and relevant data in both of the pre-training and post-training stages**, considering the sensitivity nature of patients' information. The recent emerging clinical digital twin (Katsoulakis et al., 2024; Das et al., 2023; Sun et al., 2023) and data synthesis (Liu et al., 2024d; Bauer et al., 2024; Tan et al., 2024) techniques could be explored in the future to address this challenge. ClinicalBench could be adopted to facilitate the progress in enhancing LLMs' clinical reasoning and decision-making, and minimize the gap between the development of LLMs for healthcare and clinical practice in the real world.

7 Related Work

Clinical Prediction Clinical prediction tasks play a critical role in current healthcare systems with multifaceted significance including improving patient outcomes, optimizing hospital resources, and supporting clinical decision-making (Rajkomar et al., 2019). Length-of-Stay Prediction (Stone et al., 2022), Mortality Prediction (Jentzer et al., 2021), and Readmission Prediction (Artetxe et al., 2018) are among the most common ones. While traditional ML models such as XGBoost, SVM, and RNN have still been widely adopted in these tasks (Moazemi et al., 2023; van de Water et al., 2024), many advanced models have also been developed for clinical predictive modeling (Wang et al., 2024c). For example, Xu et al. (2023) proposed to leverage hypergraph transformers with patients as hyperedges and medical codes as nodes for predictive tasks. Jiang et al. (2024) leveraged personalized knowledge graphs and attention-augmented graph neural networks for enhancing the prediction performance. Cui et al. (2024) conducted diffusion-based data augmentation to further improve health risk prediction. Although the emerging LLMs may have been demonstrated to perform well on medical question-answering and hold great promise to transform healthcare systems, their capacities on clinical prediction tasks are largely under-explored. ClinicalBench shows that LLMs cannot beat traditional ML models yet, suggesting their critical limitations in clinical applications.

Clinical Benchmarks for LLMs The majority of existing clinical benchmarks for LLMs can generally be categorized from two perspectives. The first one aims to assess LLMs' capacities in ***clinical text processing tasks*** (Harris et al., 2024; Jahan et al., 2024; Feng et al., 2024; Chen et al., 2023a; Wang et al., 2023a; Luo et al., 2024). For example, Jahan et al. (2024) has comprehensively evaluated four LLMs in different typical biomedical text tasks (*e.g.*, named entity recognition, relation extraction, entity linking, text classification and text summarization with biomedical texts). The second one intends to evaluate LLMs performance in ***clinical question-answering tasks***. Besides MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022), many recent benchmarks have been built to test the medical knowledge of LLMs in different aspects (Korgul et al., 2023; Chen et al., 2024a; Vladika et al., 2024; Shoham & Rappoport, 2024). For example, Chen et al. (2024c) and Wang et al. (2024b) have developed QA benchmarks to assess the diagnostic performance of LLMs in rare diseases. Kweon et al. (2024), Kasai et al. (2023), Rosoł et al. (2023), Alonso et al. (2024), Cai et al. (2024a), Wang et al. (2023b), and Liu et al. (2024a,c) designed QA benchmarks with languages beyond English such as Korean, Japanese, Polish and Chinese. However, benchmarks on LLMs' clinical reasoning capacities with real-world complex scenarios are relatively lacking. ClinicalBench filled the gap through a comprehensive investigation involving three common clinical prediction tasks, two databases, 14 general-purpose LLMs, 8 medical LLMs and has provided valuable insights.

8 Conclusion

In this paper, we have built a new benchmark ClinicalBench to comprehensively and systematically compare the effectiveness of traditional ML models and LLMs in typical clinical prediction tasks including Length-of-Stay Prediction, Mortality Prediction, and Readmission Prediction. With extensive empirical evidence, we find that general-purpose and medical LLMs, even with different scales of parameters, diverse prompting or fine-tuning strategies, still cannot beat traditional ML models in clinical prediction yet. Our findings have demonstrated the potential limitations of LLMs in performing real-world clinical reasoning and decision-making in complex scenarios. We urge practitioners to exercise caution when adopting LLMs in real-world clinical applications.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219, 2024. URL https://arxiv.org/abs/2404.14219.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *ArXiv preprint*, abs/2404.05590, 2024. URL https://arxiv.org/abs/2404.05590.

Arkaitz Artetxe, Andoni Beristain, and Manuel Grana. Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, 164: 49–64, 2018.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *ArXiv preprint*, abs/2401.02524, 2024. URL https://arxiv.org/abs/2401.02524.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17709–17717, 2024a.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *ArXiv preprint*, abs/2403.17297, 2024b. URL https://arxiv.org/abs/2403.17297.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *ArXiv preprint*, abs/2402.18060, 2024a. URL https://arxiv.org/abs/2402.18060.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *ArXiv preprint*, abs/2404.18231, 2024b. URL https://arxiv.org/abs/2404.18231.

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *ArXiv preprint*, abs/2305.16326, 2023a. URL https://arxiv.org/abs/2305.16326.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4850–4861, 2024c.

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

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. Med42-v2: A suite of clinical llms. *ArXiv preprint*, abs/2408.06142, 2024. URL https://arxiv.org/abs/2408.06142.

Suhan Cui, Jiaqi Wang, Yuan Zhong, Han Liu, Ting Wang, and Fenglong Ma. Automated fusion of multimodal electronic health records for better medical predictions. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pp. 361–369. SIAM, 2024.

Trisha Das, Zifeng Wang, and Jimeng Sun. Twin: Personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 402–413, 2023.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey on in-context learning. *ArXiv preprint*, abs/2301.00234, 2023. URL https://arxiv.org/abs/2301.00234.

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

Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark: Comparative study. *medRxiv*, pp. 2024–05, 2024.

Jack Gallifant, Shan Chen, Pedro Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. Language models are surprisingly fragile to drug names in biomedical benchmarks. *ArXiv preprint*, abs/2406.12066, 2024. URL https://arxiv.org/abs/2406.12066.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312, 2023.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Impact of high-quality, mixed-domain data on the performance of medical language models. *Journal of the American Medical Informatics Association*, 31(9):1875–1883, 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae120. URL https://doi.org/10.1093/jamia/ocae120.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. Towards safe and aligned large language models for medicine. *ArXiv preprint*, abs/2403.03744, 2024. URL https://arxiv.org/abs/2403.03744.

Joshua Harris, Timothy Laurence, Leo Loman, Fan Grayson, Toby Nonnenmacher, Harry Long, Loes WalsGriffith, Amy Douglas, Holly Fountain, Stelios Georgiou, et al. Evaluating large language models for public health classification and extraction tasks. *ArXiv preprint*, abs/2405.14766, 2024. URL https://arxiv.org/abs/2405.14766.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *ArXiv preprint*, abs/2404.06395, 2024. URL https://arxiv.org/abs/2404.06395.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in biology and medicine*, 171:108189, 2024.

Jacob C Jentzer, Anthony H Kashou, Francisco Lopez-Jimenez, Zachi I Attia, Suraj Kapa, Paul A Friedman, and Peter A Noseworthy. Mortality risk stratification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. *European Heart Journal Acute Cardiovascular Care*, 10(5):532–541, 2021.

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. Mistral 7b. *ArXiv preprint*, abs/2310.06825, 2023. URL https://arxiv.org/abs/2310.06825.

Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tVTN7Zs0ml.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, et al. Hidden flaws behind expert-level accuracy of gpt-4 vision in medicine. *ArXiv preprint*, abs/2401.08396, 2024. URL https://arxiv.org/abs/2401.08396.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80, 2023.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *ArXiv preprint*, abs/2303.18027, 2023. URL https://arxiv.org/abs/2303.18027.

Evangelia Katsoulakis, Qi Wang, Huanmei Wu, Leili Shahriyari, Richard Fletcher, Jinwei Liu, Luke Achenie, Hongfang Liu, Pamela Jackson, Ying Xiao, et al. Digital twins for health: a scoping review. *NPJ Digital Medicine*, 7(1):77, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Karolina Korgul, Andrew M Bean, Felix Krones, Robert McCraith, and Adam Mahdi. Exploring the landscape of large language models in medical question answering: Observations and open questions. *ArXiv preprint*, abs/2310.07225, 2023. URL https://arxiv.org/abs/2310.07225.

Sunjun Kweon, Byungjin Choi, Minkyu Kim, Rae Woong Park, and Edward Choi. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations. *ArXiv preprint*, abs/2403.01469, 2024. URL https://arxiv.org/abs/2403.01469.

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

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36, 2024a.

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

Mianxin Liu, Jinru Ding, Jie Xu, Weiguo Hu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, et al. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *ArXiv preprint*, abs/2407.10990, 2024c. URL https://arxiv.org/abs/2407.10990.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024d. URL https://openreview.net/forum?id=OJaWBhh61C.

Renze Lou, Kai Zhang, and Wenpeng Yin. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, pp. 1–10, 2024.

Xudong Luo, Zhiqi Deng, Binxia Yang, and Michael Y Luo. Pre-trained language models in medicine: A survey. *Artificial Intelligence in Medicine*, pp. 102904, 2024.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *ArXiv preprint*, abs/2308.09442, 2023. URL https://arxiv.org/abs/2308.09442.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *ArXiv preprint*, abs/2312.00164, 2023. URL https://arxiv.org/abs/2312.00164.

Sobhan Moazemi, Sahar Vahdati, Jason Li, Sebastian Kalkhoff, Luis JV Castano, Bastian Dewitz, Roman Bibo, Parisa Sabouniaghdam, Mohammad S Tootooni, Ralph A Bundschuh, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular icus: a systematic review. *Frontiers in Medicine*, 10:1109411, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *ArXiv preprint*, abs/2308.03188, 2023. URL https://arxiv.org/abs/2308.03188.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

Zi-Hao Qiu, Siqi Guo, Mao Xu, Tuo Zhao, Lijun Zhang, and Tianbao Yang. To cool or not to cool? temperature network meets large foundation models via dro. *ArXiv preprint*, abs/2404.04575, 2024. URL https://arxiv.org/abs/2404.04575.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models. *ArXiv preprint*, abs/2402.05201, 2024. URL https://arxiv.org/abs/2402.05201.

Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports*, 13(1):20512, 2023.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *ArXiv preprint*, abs/2402.07927, 2024. URL https://arxiv.org/abs/2402.07927.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *ArXiv preprint*, abs/2406.06608, 2024. URL https://arxiv.org/abs/2406.06608.

Ofir Ben Shoham and Nadav Rappoport. Medconceptsqa–open source medical concepts qa benchmark. *ArXiv preprint*, abs/2405.07348, 2024. URL https://arxiv.org/abs/2405.07348.

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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *ArXiv preprint*, abs/2305.09617, 2023b. URL https://arxiv.org/abs/2305.09617.

Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4): e0000017, 2022.

Tianze Sun, Xiwang He, and Zhonghai Li. Digital twin in healthcare: Recent updates and challenges. *Digital Health*, 9:20552076221149651, 2023.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey, 2024.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12): 1921–1934, 2020.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118, 2024. URL https://arxiv.org/abs/2408.00118.

Robin van de Water, Hendrik Nils Aurel Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

Juraj Vladika, Phillip Schneider, and Florian Matthes. Medreqal: Examining medical knowledge recall of large language models via question answering. *ArXiv preprint*, abs/2406.05845, 2024. URL https://arxiv.org/abs/2406.05845.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56 (3):1–52, 2023a.

Chong Wang, Mengyao Li, Junjun He, Zhongruo Wang, Erfan Darzi, Zan Chen, Jin Ye, Tianbin Li, Yanzhou Su, Jing Ke, et al. A survey for large language models in biomedicine. *ArXiv preprint*, abs/2409.00133, 2024a. URL https://arxiv.org/abs/2409.00133.

Guanchu Wang, Junhao Ran, Ruixiang Tang, Chia-Yuan Chang, Yu-Neng Chuang, Zirui Liu, Vladimir Braverman, Zhandong Liu, and Xia Hu. Assessing and enhancing large language models in rare disease question-answering. *ArXiv preprint*, abs/2408.08422, 2024b. URL https://arxiv.org/abs/2408.08422.

Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, et al. Recent advances in predictive modeling with electronic health records. *ArXiv preprint*, abs/2402.01077, 2024c. URL https://arxiv.org/abs/2402.01077.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 222–235, 2020.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. *ArXiv preprint*, abs/2308.08833, 2023b. URL https://arxiv.org/abs/2308.08833.

Zifeng Wang, Hanyin Wang, Benjamin Danek, Ying Li, Christina Mack, Hoifung Poon, Yajun Wang, Pranav Rajpurkar, and Jimeng Sun. A perspective for adapting generalist ai to specialized medical ai applications and their challenges. *arXiv preprint arXiv: 2411.00024*, 2024d.

Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. Hypergraph transformers for ehr-based clinical predictions. *AMIA Summits on Translational Science Proceedings*, 2023:582, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *ArXiv preprint*, abs/2407.10671, 2024. URL https://arxiv.org/abs/2407.10671.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin Danek, and Jimeng Sun. PyHealth: A deep learning toolkit for healthcare predictive modeling. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2023*, 2023. URL https://github.com/sunlabuiuc/PyHealth.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *ArXiv preprint*, abs/2403.04652, 2024. URL https://arxiv.org/abs/2403.04652.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *ArXiv preprint*, abs/2308.10792, 2023. URL https://arxiv.org/abs/2308.10792.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv preprint*, abs/2304.06364, 2023. URL https://arxiv.org/abs/2304.06364.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *ArXiv preprint*, abs/2311.05112, 2023. URL https://arxiv.org/abs/2311.05112.

Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, et al. Large language models for disease diagnosis: A scoping review. *ArXiv preprint*, abs/2409.00097, 2024. URL https://arxiv.org/abs/2409.00097.

Content of Appendix

A Reproducibility Statement

We conducted the experiments on eight NVIDIA RTX A6000 GPUs. Our code is based on PyHealth (https://github.com/sunlabuiuc/PyHealth) (Yang et al., 2023) and HuggingFace Transformers framework (https://huggingface.co/docs/transformers/en/index). In all experiments except the study on "The Impact of Decoding Temperature", the inference of LLMs is set as Greedy Decoding (*i.e.*, temperature = 0, do_sample = False) to ensure the reproducibility of our results. We also release the code and results for verification and reproduction in the project website https://clinicalbench.github.io.

We have benchmarked 14 general-prupose LLMs including Llama3-8B (Dubey et al., 2024), Llama3-70B (Dubey et al., 2024), Mistral-v0.3-7B (Jiang et al., 2023), Gemma2-9B (Team et al., 2024), Qwen2-0.5B (Yang et al., 2024), Qwen2-1.5B (Yang et al., 2024), Qwen2-7B (Yang et al., 2024), Yi-v1.5-6B (Young et al., 2024), Yi-v1.5-9B (Young et al., 2024), Yi-v1.5-34B (Young et al., 2024), Vicuna-v1.5-7B (Zheng et al., 2023), Phi3.5-mini-3.8B (Abdin et al., 2024), InternLM2.5-7B (Cai et al., 2024b), MiniCPM3-4B (Hu et al., 2024) and 8 medical LLMs including Meditron-7B (Chen et al., 2023b), Meditron-70B (Chen et al., 2023b), Medllama3-8B, BioMistral-7B (Labrak et al., 2024), Med42-8B (Christophe et al., 2024), Med42-70B (Christophe et al., 2024), BioMedGPT-7B (Luo et al., 2023) and Internist-7B (Griot et al., 2024). The model checkpoints are downloaded from https://huggingface.co/. The specific download links are as follows:

- Llama3-8B: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
- Llama3-70B: https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
- Mistral-v0.3-7B: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
- Gemma2-9B: https://huggingface.co/google/gemma-2-9b-it
- Qwen2-0.5B: https://huggingface.co/Qwen/Qwen2-0.5B-Instruct
- Qwen2-1.5B: https://huggingface.co/Qwen/Qwen2-1.5B-Instruct
- Qwen2-7B: https://huggingface.co/Qwen/Qwen2-7B-Instruct
- Yi-v1.5-6B: https://huggingface.co/01-ai/Yi-1.5-6B-Chat
- Yi-v1.5-9B: https://huggingface.co/01-ai/Yi-1.5-9B-Chat
- Yi-v1.5-34B: https://huggingface.co/01-ai/Yi-1.5-34B-Chat
- Vicuna-v1.5-7B: https://huggingface.co/lmsys/vicuna-7b-v1.5
- Phi3.5-mini-3.8B: https://huggingface.co/microsoft/Phi-3.5-mini-instruct
- InternLM2.5-7B: https://huggingface.co/internlm/internlm2_5-7b-chat
- MiniCPM3-4B: https://huggingface.co/openbmb/MiniCPM3-4B
- Meditron-7B: https://huggingface.co/epfl-llm/meditron-7b
- Meditron-70B: https://huggingface.co/epfl-llm/meditron-70b
- Medllama3-8B: https://huggingface.co/ProbeMedicalYonseiMAILab/medllama3-v20
- BioMistral-7B: :https://huggingface.co/BioMistral/BioMistral-7B
- Med42-8B: https://huggingface.co/m42-health/Llama3-Med42-8B
- Med42-70B: https://huggingface.co/m42-health/Llama3-Med42-70B
- BioMedGPT-7B: https://huggingface.co/PharMolix/BioMedGPT-LM-7B
- Internist-7B: https://huggingface.co/internistai/base-7b-v0.2

B More Experiment Results

B.1 Results of Traditional ML Models and Directly Prompting LLMs on MIMIC-IV

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI
Majority	29.56 (29.56, 29.56)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)
Minority	4.56 (4.56, 4.56)	50.00 (50.00, 50.00)	6.70 (6.70, 6.70)	50.00 (50.00, 50.00)	24.92 (24.92, 24.92)	50.00 (50.00, 50.00)
Traditional ML Models						
XGBoost	63.30 (63.18, 63.42)	95.53 (95.49, 95.57)	48.95 (48.77, 49.13)	98.27 (98.24, 98.30)	27.54 (27.42, 27.66)	60.58 (60.42, 60.74)
LR	61.81 (61.77, 61.85)	94.50 (94.47, 94.53)	51.10 (51.02, 51.18)	98.22 (98.19, 98.25)	26.55 (26.38, 26.72)	58.69 (58.45, 58.93)
DecisionTree	55.48 (55.18, 55.78)	89.10 (89.00, 89.20)	32.98 (32.69, 33.27)	88.19 (87.85, 88.53)	26.15 (26.07, 26.23)	56.43 (56.29, 56.57)
RandomForest	57.54 (57.43, 57.65)	92.21 (92.14, 92.28)	33.35 (33.16, 33.54)	95.34 (95.27, 95.41)	28.07 (28.00, 28.14)	61.21 (61.09, 61.33)
AdaBoost	62.17 (62.05, 62.29)	93.23 (93.16, 93.30)	47.66 (47.32, 48.00)	96.98 (96.95, 97.01)	26.86 (26.74, 26.98)	58.86 (58.66, 59.06)
SVM	65.94 (65.86, 66.02)	96.18 (96.12, 96.24)	46.63 (46.43, 46.83)	98.18 (98.17, 98.19)	28.62 (28.49, 28.75)	62.49 (62.32, 62.66)
NaiveBayes	48.03 (47.98, 48.08)	66.52 (66.32, 66.72)	18.83 (18.62, 19.04)	82.05 (81.82, 82.28)	25.51 (25.47, 25.55)	52.95 (52.88, 53.02)
KNN	46.72 (46.58, 46.86)	74.96 (74.80, 75.12)	45.01 (44.78, 45.24)	90.70 (90.58, 90.82)	25.07 (24.93, 25.21)	55.30 (55.07, 55.53)
MLP	59.51 (59.46, 59.56)	92.43 (92.36, 92.50)	47.07 (46.87, 47.27)	97.78 (97.75, 97.81)	25.45 (25.30, 25.60)	56.80 (56.56, 57.04)
Transformer	56.51 (56.34, 56.68)	91.00 (90.91, 91.09)	45.87 (45.34, 46.40)	98.07 (98.05, 98.09)	26.04 (25.90, 26.18)	57.25 (57.03, 57.47)
RNN	59.93 (59.87, 59.99)	92.44 (92.37, 92.51)	49.75 (49.56, 49.94)	97.48 (97.44, 97.52)	25.08 (24.96, 25.20)	55.94 (55.68, 56.20)
General-purpose Large Language Models						
LLama3-8B	18.08 (18.02, 18.14)	91.72 (91.64, 91.8)	14.97 (14.95, 14.99)	95.45 (95.4, 95.5)	25.1 (25.09, 25.11)	49.6 (49.48, 49.72)
Mistral-v0.3-7B	12.29 (12.22, 12.36)	86.1 (86.07, 86.13)	28.97 (28.89, 29.05)	95.64 (95.6, 95.68)	24.7 (24.6, 24.8)	51.22 (51.06, 51.38)
Gemma2-9B	19.83 (19.79, 19.87)	89.68 (89.6, 89.76)	32.91 (32.86, 32.96)	95.82 (95.78, 95.86)	24.83 (24.81, 24.85)	49.92 (49.79, 50.05)
Qwen2-7B	6.57 (6.54, 6.6)	83.46 (83.28, 83.64)	48.05 (47.63, 48.47)	95.46 (95.42, 95.5)	0.23 (0.21, 0.25)	48.83 (48.72, 48.94)
Yi-v1.5-9B	6.97 (6.96, 6.98)	84.35 (84.3, 84.4)	16.87 (16.82, 16.92)	88.89 (88.79, 88.99)	19.14 (19.0, 19.28)	50.33 (50.23, 50.43)
Vicuna-v1.5-7B	23.25 (23.17, 23.33)	81.12 (80.99, 81.25)	7.22 (7.22, 7.22)	88.39 (88.33, 88.45)	24.92 (24.92, 24.92)	51.92 (51.87, 51.97)
Phi3.5-mini-3.8B	11.56 (11.51, 11.61)	82.84 (82.70, 82.98)	16.72 (16.68, 16.76)	93.55 (93.50, 93.60)	21.52 (21.40, 21.64)	50.87 (50.73, 51.01)
InternLM2.5-7B	11.40 (11.35, 11.45)	90.42 (90.31, 90.53)	40.68 (40.39, 40.97)	95.29 (95.25, 95.33)	0.00 (0.00, 0.00)	49.05 (48.89, 49.21)
MiniCPM3-4B	17.80 (17.71, 17.89)	86.41 (86.29, 86.53)	12.28 (12.26, 12.30)	94.54 (94.47, 94.61)	22.93 (22.89, 22.97)	49.90 (49.80, 50.00)
Medical Large Language Models						
Meditron-7B	28.08 (27.96, 28.2)	86.17 (86.03, 86.31)	6.7 (6.7, 6.7)	91.95 (91.9, 92.0)	24.92 (24.92, 24.92)	48.92 (48.75, 49.09)
Medllama3-8B	8.31 (8.27, 8.35)	80.73 (80.57, 80.89)	12.86 (12.84, 12.88)	91.48 (91.46, 91.5)	24.92 (24.92, 24.92)	48.22 (48.04, 48.4)
BioMistral-7B	13.53 (13.46, 13.6)	39.12 (38.99, 39.25)	22.81 (22.75, 22.87)	91.89 (91.81, 91.97)	22.03 (21.94, 22.12)	48.54 (48.45, 48.63)
Med42-8B	9.95 (9.9, 10.0)	90.86 (90.76, 90.96)	16.7 (16.67, 16.73)	93.62 (93.56, 93.68)	25.61 (25.55, 25.67)	50.78 (50.67, 50.89)
BioMedGPT-7B	7.65 (7.65, 7.65)	70.48 (70.22, 70.74)	12.01 (12.00, 12.02)	93.51 (93.49, 93.53)	23.97 (23.92, 24.02)	50.35 (50.30, 50.40)
Internist-7B	14.24 (14.18, 14.30)	89.39 (89.32, 89.46)	28.81 (28.75, 28.87)	96.57 (96.54, 96.60)	15.72 (15.59, 15.85)	50.59 (50.43, 50.75)

Table 4: **Performance of LLMs and Traditional ML Models on Typical Clinical Prediction Tasks**. Experiments are conducted in **MIMIC-IV** dataset. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.

B.2 Results of Traditional ML Models on Different Scales of Training Set

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI
Majority	23.37 (23.37, 23.37)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)
Minority	10.72 (10.72, 10.72)	50.00 (50.00, 50.00)	20.88 (20.88, 20.88)	50.00 (50.00, 50.00)	13.64 (13.64, 13.64)	50.00 (50.00, 50.00)
Traditional ML Models with 40% of Original Training Set from MIMIC-III						
XGBoost	63.99 (63.84, 64.14)	91.53 (91.45, 91.61)	62.16 (62.03, 62.29)	94.67 (94.63, 94.71)	17.69 (17.59, 17.79)	64.23 (64.04, 64.42)
LR	62.44 (62.28, 62.60)	90.61 (90.47, 90.75)	60.24 (60.14, 60.34)	92.91 (92.88, 92.94)	18.32 (18.10, 18.54)	66.06 (65.68, 66.44)
DecisionTree	56.14 (55.99, 56.29)	83.16 (82.98, 83.34)	48.41 (48.20, 48.62)	81.17 (80.89, 81.45)	14.95 (14.86, 15.04)	53.61 (53.31, 53.91)
RandomForest	61.06 (60.96, 61.16)	90.88 (90.78, 90.98)	53.11 (52.96, 53.26)	90.78 (90.70, 90.86)	18.34 (18.09, 18.59)	66.40 (65.95, 66.85)
AdaBoost	62.54 (62.34, 62.74)	89.38 (89.09, 89.67)	57.04 (56.76, 57.32)	90.94 (90.78, 91.10)	16.26 (16.09, 16.43)	59.77 (59.27, 60.27)
SVM	65.05 (64.90, 65.20)	91.71 (91.63, 91.79)	60.51 (60.30, 60.72)	94.04 (93.99, 94.09)	17.97 (17.70, 18.24)	66.84 (66.48, 67.20)
NaiveBayes	45.57 (45.48, 45.66)	62.48 (62.28, 62.68)	24.86 (24.81, 24.91)	59.10 (59.02, 59.18)	16.63 (16.51, 16.75)	58.05 (57.81, 58.29)
KNN	45.34 (45.24, 45.44)	74.43 (74.28, 74.58)	42.17 (41.90, 42.44)	83.75 (83.59, 83.91)	15.25 (15.14, 15.36)	59.80 (59.38, 60.22)
MLP	59.92 (59.75, 60.09)	87.99 (87.81, 88.17)	57.00 (56.85, 57.15)	90.77 (90.70, 90.84)	18.25 (18.07, 18.43)	65.71 (65.38, 66.04)
Transformer	58.73 (58.45, 59.01)	87.22 (87.02, 87.42)	56.05 (55.81, 56.29)	91.74 (91.71, 91.77)	18.21 (18.04, 18.38)	65.49 (65.19, 65.79)
RNN	61.07 (60.88, 61.26)	88.08 (87.94, 88.22)	58.77 (58.60, 58.94)	90.08 (89.96, 90.20)	19.02 (18.86, 19.18)	67.38 (67.11, 67.65)
Traditional ML Models with 20% of Original Training Set from MIMIC-III						
XGBoost	61.95 (61.81, 62.09)	90.89 (90.78, 91.00)	57.37 (57.07, 57.67)	92.86 (92.78, 92.94)	16.55 (16.35, 16.75)	61.33 (60.79, 61.87)
LR	60.78 (60.55, 61.01)	89.71 (89.52, 89.90)	56.55 (56.24, 56.86)	91.57 (91.47, 91.67)	16.69 (16.45, 16.93)	63.59 (63.19, 63.99)
DecisionTree	54.8 (54.64, 54.96)	79.85 (79.65, 80.05)	44.23 (43.66, 44.80)	77.20 (76.99, 77.41)	15.01 (14.86, 15.16)	56.25 (55.94, 56.56)
RandomForest	60.48 (60.28, 60.68)	90.59 (90.48, 90.70)	52.00 (51.85, 52.15)	90.15 (90.05, 90.25)	16.90 (16.77, 17.03)	63.09 (62.93, 63.25)
AdaBoost	60.51 (60.32, 60.70)	87.97 (87.76, 88.18)	53.76 (53.60, 53.92)	88.92 (88.77, 89.07)	15.75 (15.50, 16.00)	58.43 (58.01, 58.85)
SVM	63.05 (62.92, 63.18)	91.10 (91.02, 91.18)	57.35 (57.01, 57.69)	92.72 (92.64, 92.80)	17.41 (17.20, 17.62)	64.88 (64.58, 65.18)
NaiveBayes	43.82 (43.77, 43.87)	62.88 (62.70, 63.06)	25.63 (25.57, 25.69)	60.44 (60.34, 60.54)	16.71 (16.54, 16.88)	58.02 (57.67, 58.37)
KNN	44.49 (44.33, 44.65)	74.31 (74.14, 74.48)	42.39 (42.21, 42.57)	82.91 (82.77, 83.05)	15.04 (14.94, 15.14)	58.93 (58.47, 59.39)
MLP	58.66 (58.41, 58.91)	87.28 (87.05, 87.51)	53.54 (53.24, 53.84)	89.98 (89.84, 90.12)	17.08 (16.80, 17.36)	62.81 (62.39, 63.23)
Transformer	58.69 (58.44, 58.94)	88.18 (87.96, 88.40)	53.02 (52.84, 53.20)	90.74 (90.56, 90.92)	17.04 (16.88, 17.20)	61.73 (61.37, 62.09)
RNN	61.16 (60.89, 61.43)	88.74 (88.51, 88.97)	55.77 (55.36, 56.18)	89.13 (88.89, 89.37)	17.85 (17.66, 18.04)	64.03 (63.78, 64.28)

Table 5: **Performance of Traditional ML Models on Typical Clinical Prediction Tasks with Different Scales of Training Set**. Experiments are conducted in **MIMIC-III** dataset. The training set is obtained through stratified sampling from the original training set in Table 1. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	**AUROC (%)** 95% CI	**F1 (%)** 95% CI	**AUROC (%)** 95% CI	**F1 (%)** 95% CI	**AUROC (%)** 95% CI
Majority	23.37 (23.37, 23.37)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)	0.00 (0.00, 0.00)	50.00 (50.00, 50.00)
Minority	10.72 (10.72, 10.72)	50.00 (50.00, 50.00)	20.88 (20.88, 20.88)	50.00 (50.00, 50.00)	13.64 (13.64, 13.64)	50.00 (50.00, 50.00)
Traditional ML Models with 10% of Original Training Set from MIMIC-III						
XGBoost	61.35 (61.18, 61.52)	90.16 (90.05, 90.27)	51.62 (51.37, 51.87)	89.93 (89.82, 90.04)	17.05 (16.81, 17.29)	61.43 (60.79, 62.07)
LR	60.79 (60.64, 60.94)	90.25 (90.10, 90.40)	53.92 (53.64, 54.20)	90.49 (90.39, 90.59)	17.71 (17.54, 17.88)	62.21 (61.85, 62.57)
DecisionTree	53.07 (52.98, 53.16)	77.11 (76.77, 77.45)	41.43 (41.16, 41.70)	74.65 (74.43, 74.87)	15.32 (15.17, 15.47)	55.75 (55.36, 56.14)
RandomForest	60.39 (60.25, 60.53)	90.61 (90.54, 90.68)	50.00 (49.75, 50.25)	88.72 (88.63, 88.81)	17.47 (17.35, 17.59)	63.51 (63.32, 63.70)
AdaBoost	60.3 (60.23, 60.37)	87.55 (87.40, 87.70)	50.07 (49.63, 50.51)	85.78 (85.61, 85.95)	16.03 (15.80, 16.26)	57.48 (57.07, 57.89)
SVM	62.33 (62.2, 62.46)	90.73 (90.66, 90.8)	52.51 (52.17, 52.85)	90.97 (90.86, 91.08)	17.01 (16.81, 17.21)	52.28 (50.47, 54.09)
NaiveBayes	41.54 (41.39, 41.69)	63.82 (63.65, 63.99)	26.48 (26.32, 26.64)	62.18 (61.81, 62.55)	14.87 (14.79, 14.95)	54.25 (54.11, 54.39)
KNN	42.37 (42.23, 42.51)	73.74 (73.49, 73.99)	41.26 (41.13, 41.39)	82.37 (82.31, 82.43)	14.30 (14.23, 14.37)	57.96 (57.27, 58.65)
MLP	58.65 (58.39, 58.91)	87.51 (87.32, 87.70)	51.55 (51.17, 51.93)	89.57 (89.48, 89.66)	15.95 (15.83, 16.07)	59.23 (59.00, 59.46)
Transformer	58.94 (58.63, 59.25)	89.43 (89.25, 89.61)	50.19 (49.81, 50.57)	89.43 (89.31, 89.55)	15.76 (15.52, 16.00)	58.34 (57.91, 58.77)
RNN	61.81 (61.54, 62.08)	89.20 (89.04, 89.36)	54.62 (54.32, 54.92)	89.00 (88.88, 89.12)	13.97 (13.55, 14.39)	61.82 (61.65, 61.99)
Traditional ML Models with 5% of Original Training Set from MIMIC-III						
XGBoost	59.21 (59.00, 59.42)	89.16 (89.00, 89.32)	48.60 (48.39, 48.81)	87.67 (87.58, 87.76)	16.40 (16.18, 16.62)	58.90 (58.49, 59.31)
LR	59.90 (59.74, 60.06)	89.92 (89.74, 90.10)	52.87 (52.63, 53.11)	89.50 (89.36, 89.64)	15.51 (15.37, 15.65)	58.96 (58.50, 59.42)
DecisionTree	52.35 (52.17, 52.53)	76.17 (75.98, 76.36)	39.02 (38.44, 39.60)	70.48 (69.96, 71.00)	15.90 (15.77, 16.03)	56.24 (56.01, 56.47)
RandomForest	60.89 (60.74, 61.04)	90.34 (90.26, 90.42)	52.10 (51.83, 52.37)	89.15 (88.99, 89.31)	15.75 (15.58, 15.92)	60.29 (59.84, 60.74)
AdaBoost	56.63 (56.35, 56.91)	86.61 (86.43, 86.79)	45.37 (45.06, 45.68)	83.13 (82.77, 83.49)	15.41 (15.33, 15.49)	59.09 (58.89, 59.29)
SVM	61.33 (61.20, 61.46)	90.10 (90.00, 90.20)	52.58 (52.24, 52.92)	90.04 (89.93, 90.15)	15.95 (15.73, 16.17)	49.55 (47.98, 51.12)
NaiveBayes	42.91 (42.71, 43.11)	66.12 (65.90, 66.34)	29.89 (29.72, 30.06)	66.44 (66.25, 66.63)	14.06 (13.82, 14.30)	53.30 (53.02, 53.58)
KNN	39.98 (39.91, 40.05)	71.55 (71.39, 71.71)	40.70 (40.43, 40.97)	83.04 (82.93, 83.15)	14.26 (14.23, 14.29)	55.74 (55.13, 56.35)
MLP	57.97 (57.79, 58.15)	87.22 (87.03, 87.41)	50.27 (50.04, 50.50)	87.92 (87.77, 88.07)	14.91 (14.82, 15.00)	57.36 (57.01, 57.71)
Transformer	58.23 (57.95, 58.51)	88.73 (88.54, 88.92)	48.67 (48.51, 48.83)	87.88 (87.68, 88.08)	14.04 (13.72, 14.36)	54.06 (53.55, 54.57)
RNN	60.94 (60.74, 61.14)	88.19 (88.03, 88.35)	52.84 (52.61, 53.07)	88.04 (87.85, 88.23)	11.92 (11.21, 12.63)	59.50 (59.16, 59.84)

Table 6: **Performance of Traditional ML Models on Typical Clinical Prediction Tasks with Different Scales of Training Set**. Experiments are conducted in **MIMIC-III** dataset. The training set is obtained through stratified sampling from the original training set in Table 1. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI
Majority	29.56 (29.56, 29.56)	50.0 (50.0, 50.0)	0.0 (0.0, 0.0)	50.0 (50.0, 50.0)	0.0 (0.0, 0.0)	50.0 (50.0, 50.0)
Minority	4.56 (4.56, 4.56)	50.0 (50.0, 50.0)	6.7 (6.7, 6.7)	50.0 (50.0, 50.0)	24.92 (24.92, 24.92)	50.0 (50.0, 50.0)
Traditional ML Models with 40% of Original Training Set from MIMIC-IV						
XGBoost	61.81 (61.70, 61.92)	94.51 (94.46, 94.56)	44.06 (43.66, 44.46)	97.36 (97.31, 97.41)	26.87 (26.77, 26.97)	58.86 (58.69, 59.03)
LR	60.60 (60.41, 60.79)	93.64 (93.58, 93.70)	47.10 (46.86, 47.34)	97.63 (97.59, 97.67)	26.24 (26.15, 26.33)	58.01 (57.84, 58.18)
DecisionTree	53.07 (52.76, 53.38)	86.57 (86.50, 86.64)	30.97 (30.42, 31.52)	85.31 (85.00, 85.62)	25.33 (25.05, 25.61)	54.97 (54.84, 55.10)
RandomForest	57.86 (57.78, 57.94)	92.06 (91.97, 92.15)	34.24 (33.77, 34.71)	95.18 (95.04, 95.32)	27.08 (27.00, 27.16)	59.52 (59.35, 59.69)
AdaBoost	60.64 (60.53, 60.75)	92.50 (92.45, 92.55)	43.14 (42.75, 43.53)	95.29 (95.11, 95.47)	26.17 (26.11, 26.23)	56.45 (56.29, 56.61)
SVM	64.50 (64.39, 64.61)	95.48 (95.42, 95.54)	42.79 (42.41, 43.17)	97.51 (97.47, 97.55)	28.32 (28.22, 28.42)	60.85 (60.70, 61.00)
NaiveBayes	42.47 (42.37, 42.57)	70.00 (69.84, 70.16)	16.13 (15.99, 16.27)	77.47 (77.36, 77.58)	25.25 (25.14, 25.36)	52.79 (52.58, 53.00)
KNN	43.54 (43.31, 43.77)	71.62 (71.35, 71.89)	38.66 (38.15, 39.17)	90.02 (89.83, 90.21)	23.11 (22.78, 23.44)	53.63 (53.35, 53.91)
MLP	58.70 (58.62, 58.78)	90.83 (90.75, 90.91)	42.41 (42.14, 42.68)	97.17 (97.14, 97.20)	25.81 (25.65, 25.97)	56.94 (56.76, 57.12)
Transformer	57.66 (57.44, 57.88)	91.56 (91.47, 91.65)	40.94 (40.48, 41.40)	97.38 (97.33, 97.43)	25.33 (25.27, 25.39)	56.36 (56.27, 56.45)
RNN	61.63 (61.53, 61.73)	91.74 (91.69, 91.79)	46.56 (46.37, 46.75)	96.89 (96.85, 96.93)	25.56 (25.47, 25.65)	56.74 (56.60, 56.88)
Traditional ML Models with 20% of Original Training Set from MIMIC-IV						
XGBoost	60.05 (59.97, 60.13)	93.49 (93.41, 93.57)	38.80 (38.51, 39.09)	96.21 (96.15, 96.27)	25.16 (25.05, 25.27)	56.43 (56.21, 56.65)
LR	59.51 (59.31, 59.71)	92.87 (92.71, 93.03)	42.81 (42.55, 43.07)	96.91 (96.84, 96.98)	25.27 (25.16, 25.38)	56.00 (55.89, 56.11)
DecisionTree	52.15 (51.83, 52.47)	84.08 (83.91, 84.25)	27.27 (26.49, 28.05)	83.42 (83.04, 83.80)	21.39 (20.78, 22.00)	52.87 (52.66, 53.08)
RandomForest	57.70 (57.55, 57.85)	91.85 (91.75, 91.95)	34.77 (34.30, 35.24)	95.14 (95.01, 95.27)	26.55 (26.46, 26.64)	58.08 (57.85, 58.31)
AdaBoost	57.76 (57.58, 57.94)	91.66 (91.58, 91.74)	38.77 (38.60, 38.94)	94.49 (94.27, 94.71)	24.88 (24.80, 24.96)	54.71 (54.62, 54.80)
SVM	63.14 (63.06, 63.22)	94.88 (94.82, 94.94)	39.96 (39.57, 40.35)	96.53 (96.46, 96.60)	26.62 (26.45, 26.79)	58.14 (57.93, 58.35)
NaiveBayes	39.13 (39.00, 39.26)	70.63 (70.41, 70.85)	14.53 (14.40, 14.66)	75.68 (75.44, 75.92)	24.66 (24.46, 24.86)	53.69 (53.48, 53.90)
KNN	44.05 (43.75, 44.35)	72.61 (72.43, 72.79)	38.87 (38.39, 39.35)	87.61 (87.30, 87.92)	22.96 (22.78, 23.14)	53.41 (53.23, 53.59)
MLP	57.34 (57.16, 57.52)	89.89 (89.76, 90.02)	39.84 (39.51, 40.17)	96.59 (96.53, 96.65)	24.37 (24.29, 24.45)	54.71 (54.58, 54.84)
Transformer	57.39 (57.19, 57.59)	91.72 (91.57, 91.87)	35.91 (35.18, 36.64)	96.46 (96.38, 96.54)	24.35 (24.24, 24.46)	54.57 (54.43, 54.71)
RNN	60.27 (60.07, 60.47)	90.79 (90.62, 90.96)	42.13 (41.87, 42.39)	96.20 (96.13, 96.27)	24.77 (24.67, 24.87)	55.31 (55.23, 55.39)

Table 7: **Performance of Traditional ML Models on Typical Clinical Prediction Tasks with Different Scales of Training Set**. Experiments are conducted in **MIMIC-IV** dataset. The training set is obtained through stratified sampling from the original training set in Table 1. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.

Method	Length-of-Stay Prediction		Mortality Prediction		Readmission Prediction	
	Macro F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI	F1 (%) 95% CI	AUROC (%) 95% CI
Majority	29.56 (29.56, 29.56)	50.0 (50.0, 50.0)	0.0 (0.0, 0.0)	50.0 (50.0, 50.0)	0.0 (0.0, 0.0)	50.0 (50.0, 50.0)
Minority	4.56 (4.56, 4.56)	50.0 (50.0, 50.0)	6.7 (6.7, 6.7)	50.0 (50.0, 50.0)	24.92 (24.92, 24.92)	50.0 (50.0, 50.0)
Traditional ML Models with 10% of Original Training Set from MIMIC-IV						
XGBoost	56.22 (56.14, 56.30)	91.79 (91.69, 91.89)	33.59 (33.07, 34.11)	94.67 (94.53, 94.81)	24.84 (24.69, 24.99)	54.70 (54.42, 54.98)
LR	57.96 (57.80, 58.12)	92.13 (92.00, 92.26)	40.20 (39.65, 40.75)	96.25 (96.18, 96.32)	24.49 (24.36, 24.62)	54.82 (54.58, 55.06)
DecisionTree	49.48 (49.12, 49.84)	80.08 (79.75, 80.41)	26.82 (26.23, 27.41)	80.25 (79.76, 80.74)	24.23 (23.78, 24.68)	53.16 (52.91, 53.41)
RandomForest	57.15 (56.97, 57.33)	91.66 (91.54, 91.78)	35.29 (34.67, 35.91)	95.16 (94.97, 95.35)	25.42 (25.20, 25.64)	56.09 (55.81, 56.37)
AdaBoost	54.47 (54.31, 54.63)	90.96 (90.80, 91.12)	39.87 (39.48, 40.26)	95.36 (95.29, 95.43)	23.13 (22.95, 23.31)	52.50 (52.33, 52.67)
SVM	62.02 (61.87, 62.17)	94.09 (94.01, 94.17)	36.53 (36.12, 36.94)	95.80 (95.66, 95.94)	24.72 (24.51, 24.93)	50.87 (50.10, 51.64)
NaiveBayes	36.28 (36.20, 36.36)	71.22 (71.14, 71.30)	13.66 (13.58, 13.74)	75.26 (75.19, 75.33)	24.06 (23.90, 24.22)	53.51 (53.39, 53.63)
KNN	40.62 (40.34, 40.90)	70.57 (70.17, 70.97)	34.72 (34.03, 35.41)	85.06 (84.81, 85.31)	22.31 (22.20, 22.42)	52.84 (52.62, 53.06)
MLP	56.07 (55.96, 56.18)	88.97 (88.87, 89.07)	36.12 (35.78, 36.46)	95.75 (95.71, 95.79)	23.86 (23.74, 23.98)	53.71 (53.54, 53.88)
Transformer	55.37 (55.24, 55.50)	89.94 (89.82, 90.06)	30.94 (30.54, 31.34)	95.21 (95.12, 95.30)	24.18 (24.05, 24.31)	53.91 (53.75, 54.07)
RNN	59.04 (58.94, 59.14)	89.39 (89.27, 89.51)	39.13 (38.77, 39.49)	95.64 (95.58, 95.70)	24.37 (24.26, 24.48)	54.44 (54.26, 54.62)
Traditional ML Models with 5% of Original Training Set from MIMIC-IV						
XGBoost	54.47 (54.38, 54.56)	89.99 (89.76, 90.22)	25.04 (24.52, 25.56)	90.50 (90.33, 90.67)	24.39 (24.25, 24.53)	53.80 (53.53, 54.07)
LR	57.20 (57.03, 57.37)	91.51 (91.35, 91.67)	34.45 (33.87, 35.03)	94.75 (94.62, 94.88)	23.96 (23.78, 24.14)	53.57 (53.22, 53.92)
DecisionTree	49.15 (48.75, 49.55)	75.08 (73.98, 76.18)	22.11 (21.63, 22.59)	78.31 (77.88, 78.74)	22.53 (21.91, 23.15)	52.28 (52.01, 52.55)
RandomForest	57.40 (57.28, 57.52)	91.53 (91.35, 91.71)	34.35 (33.68, 35.02)	94.07 (93.88, 94.26)	24.94 (24.76, 25.12)	54.40 (53.97, 54.83)
AdaBoost	50.14 (49.93, 50.35)	90.59 (90.47, 90.71)	37.37 (36.54, 38.20)	92.59 (92.34, 92.84)	23.44 (23.28, 23.60)	52.27 (51.99, 52.55)
SVM	60.63 (60.36, 60.90)	93.00 (92.86, 93.14)	32.92 (32.52, 33.32)	93.75 (93.61, 93.89)	24.75 (24.52, 24.98)	50.00 (49.23, 50.77)
NaiveBayes	34.42 (34.32, 34.52)	70.08 (69.74, 70.42)	12.74 (12.58, 12.90)	73.99 (73.69, 74.29)	23.14 (23.04, 23.24)	52.19 (52.07, 52.31)
KNN	38.72 (38.33, 39.11)	67.40 (67.07, 67.73)	24.09 (23.40, 24.78)	79.08 (78.66, 79.50)	22.13 (21.89, 22.37)	52.32 (52.13, 52.51)
MLP	54.25 (54.10, 54.40)	86.36 (86.17, 86.55)	32.24 (31.75, 32.73)	94.16 (94.05, 94.27)	23.68 (23.61, 23.75)	53.13 (52.86, 53.40)
Transformer	54.50 (54.35, 54.65)	88.37 (88.09, 88.65)	22.59 (22.04, 23.14)	91.43 (91.19, 91.67)	23.99 (23.87, 24.11)	52.83 (52.58, 53.08)
RNN	57.88 (57.70, 58.06)	87.51 (87.14, 87.88)	17.32 (16.98, 17.66)	92.38 (92.25, 92.51)	24.09 (23.96, 24.22)	53.77 (53.48, 54.06)

Table 8: **Performance of Traditional ML Models on Typical Clinical Prediction Tasks with Different Scales of Training Set**. Experiments are conducted in **MIMIC-IV** dataset. The training set is obtained through stratified sampling from the original training set in Table 1. The numbers in bracket are the ranges of performance with **95% Confidence Interval (95% CI)**, which are calculated with the 5-run scores.

B.3 Results of LLMs with Different Temperatures of Decoding

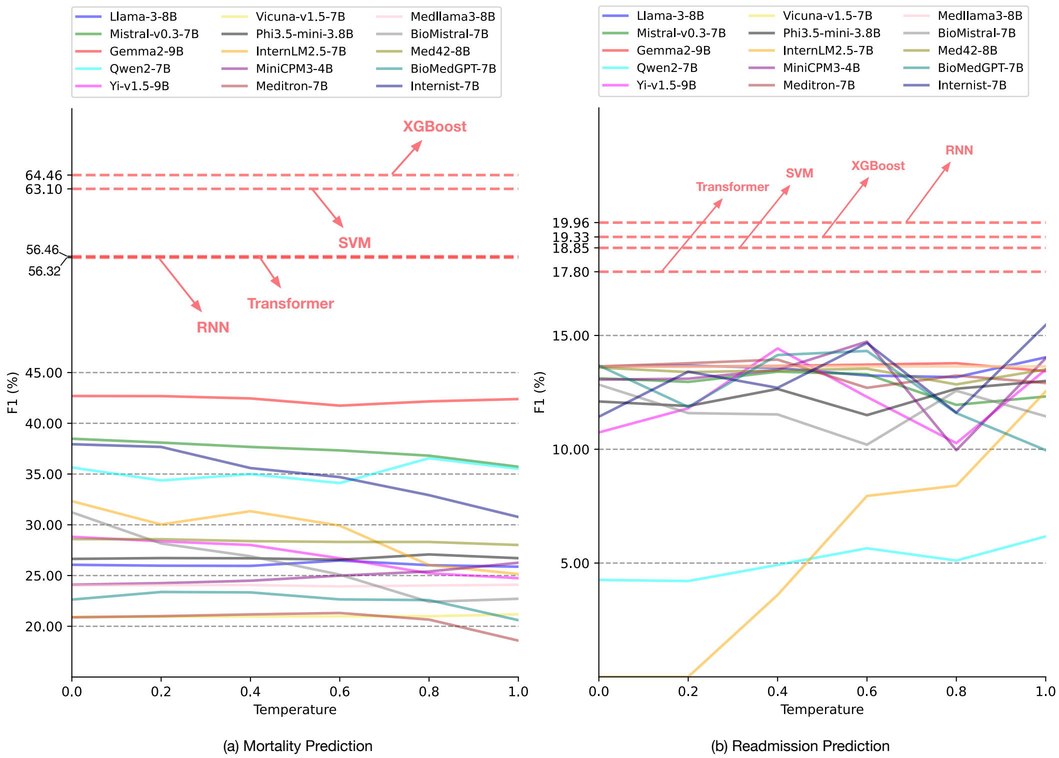
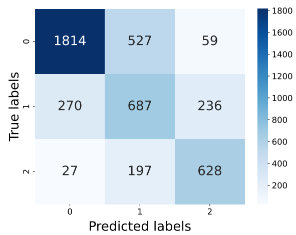


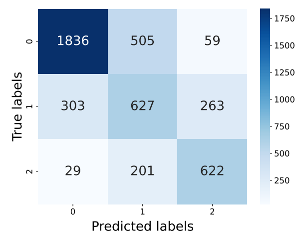
Figure 5: **Performance Comparison Between LLMs with Different Temperatures and Traditional ML Models on Mortality Prediction and Readmission Prediction**. Experiments are conducted in MIMIC-III dataset.

B.4 Confusion Matrix of Traditional ML Models and Directly Prompting LLMs

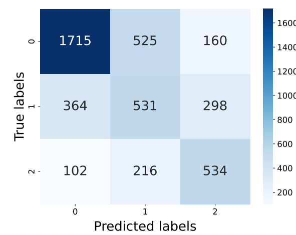
Figure 6: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Length-of-Stay Prediction on MIMIC-III Dataset**.



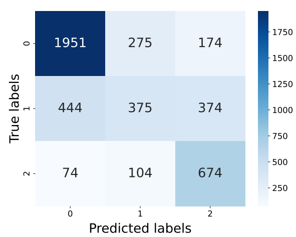
(a) XGBoost



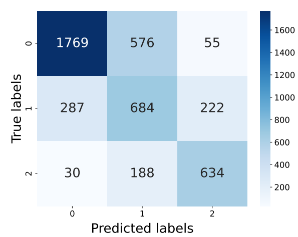
(b) LR



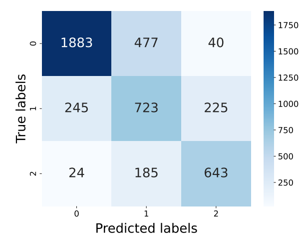
(c) DecisionTree



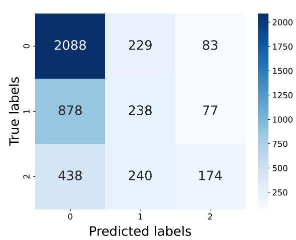
(d) RandomForest



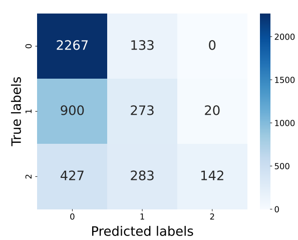
(e) AdaBoost



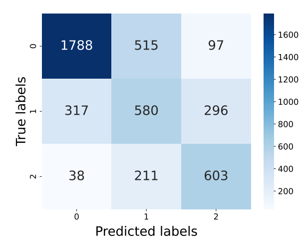
(f) SVM



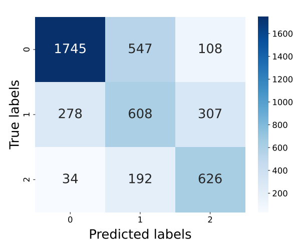
(g) NaiveBayes



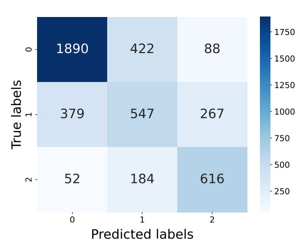
(h) KNN



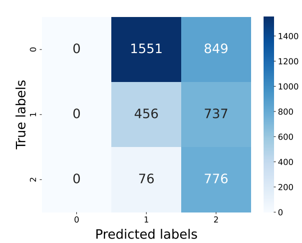
(i) MLP



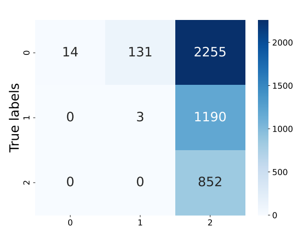
(j) Transformer



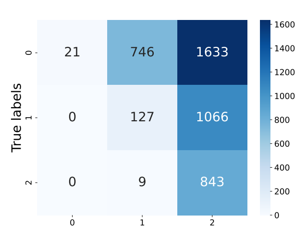
(k) RNN



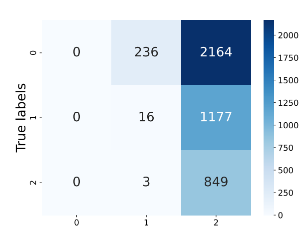
(l) Llama3-8B



(m) Mistral-v0.3-7B

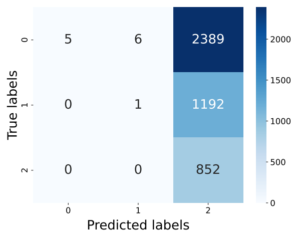


(n) Gemma2-9b

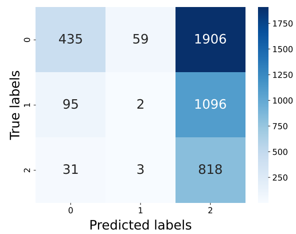


(o) Qwen2-7B

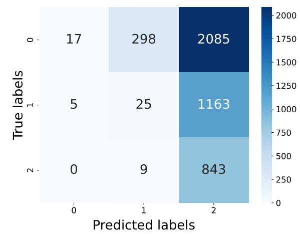
Figure 7: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Length-of-Stay Prediction on MIMIC-III Dataset**.



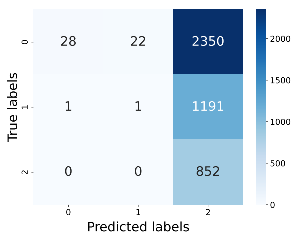
(a) Yi-v1.5-9B



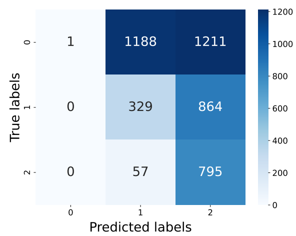
(b) Vicuna-v1.5-7b



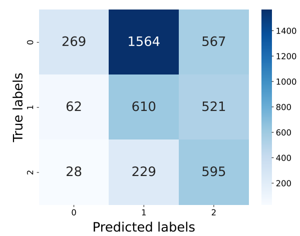
(c) Phi3.5-mini-3.8B



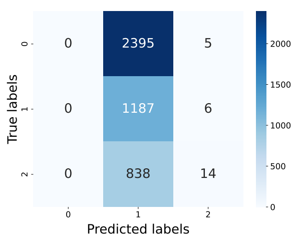
(d) InternLM2.5-7b



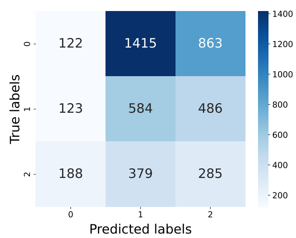
(e) MiniCPM3-4B



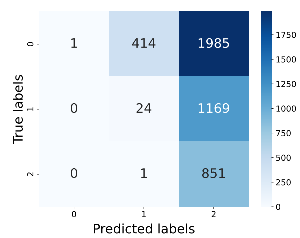
(f) Meditron-7B



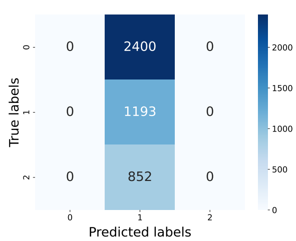
(g) Medllama3-8B



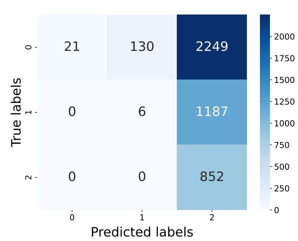
(h) BioMistral-7B



(i) Med42-8B

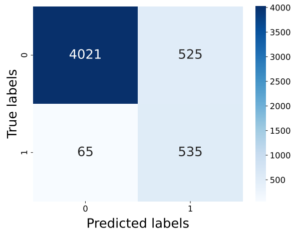


(j) BioMedGPT-7B

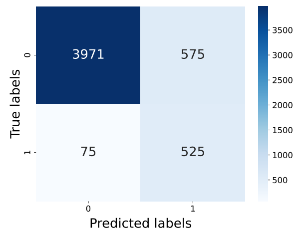


(k) Internist-7B

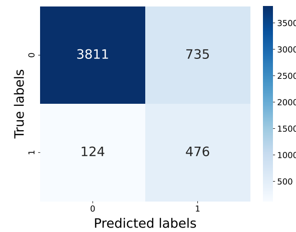
Figure 8: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Mortality Prediction on MIMIC-III Dataset**.



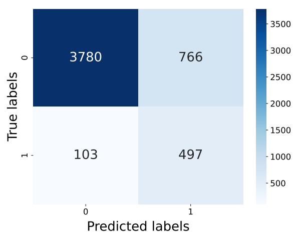
(a) XGBoost



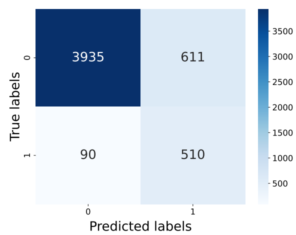
(b) LR



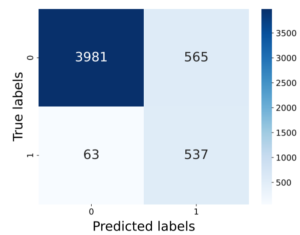
(c) DecisionTree



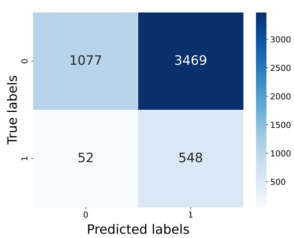
(d) RandomForest



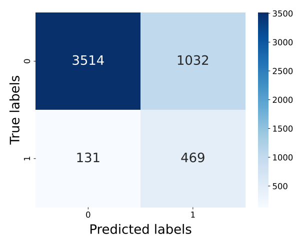
(e) AdaBoost



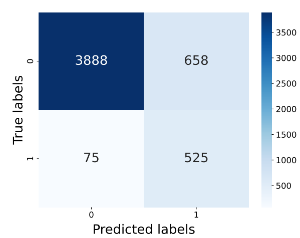
(f) SVM



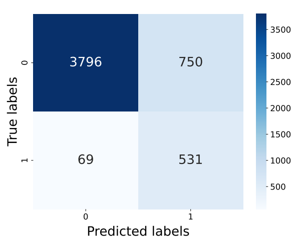
(g) NaiveBayes



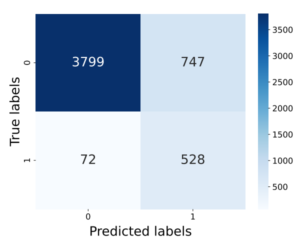
(h) KNN



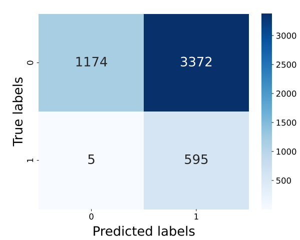
(i) MLP



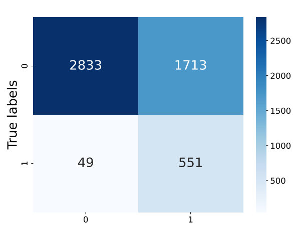
(j) Transformer



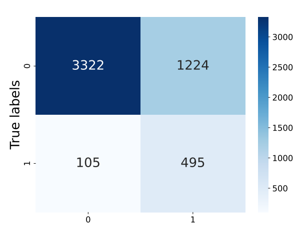
(k) RNN



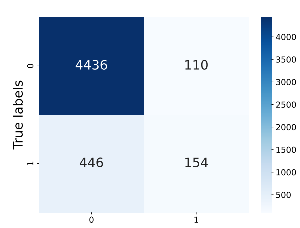
(l) Llama3-8B



(m) Mistral-v0.3-7B

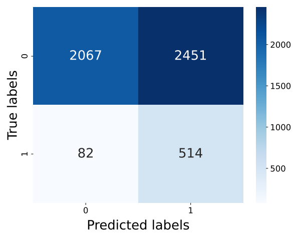


(n) Gemma2-9b

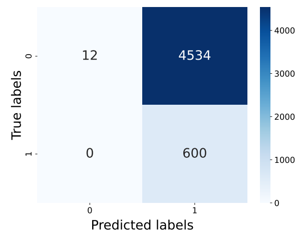


(o) Qwen2-7B

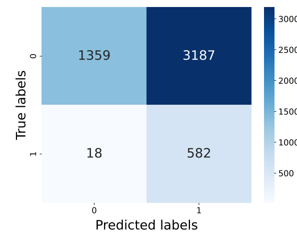
Figure 9: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Mortality Prediction on MIMIC-III Dataset**.



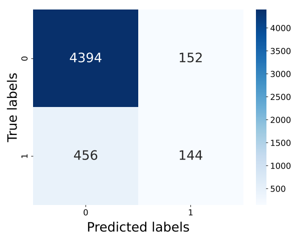
(a) Yi-v1.5-9B



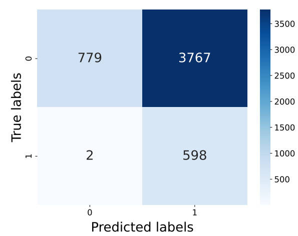
(b) Vicuna-v1.5-7B



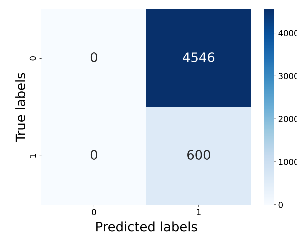
(c) Phi3.5-mini-3.8B



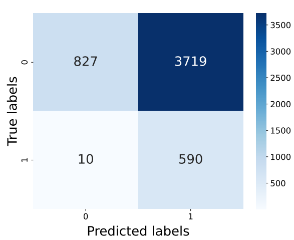
(d) InternLM2.5-7B



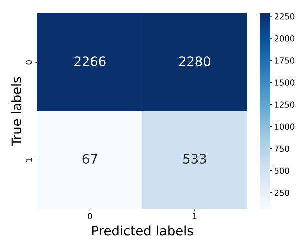
(e) MiniCPM3-4B



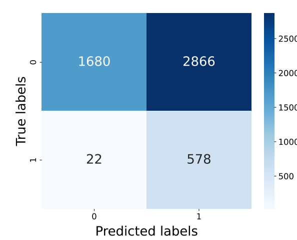
(f) Meditron-7B



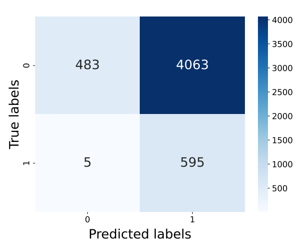
(g) Medllama3-8B



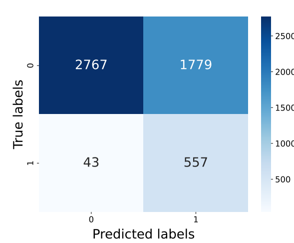
(h) BioMistral-7B



(i) Med42-8B

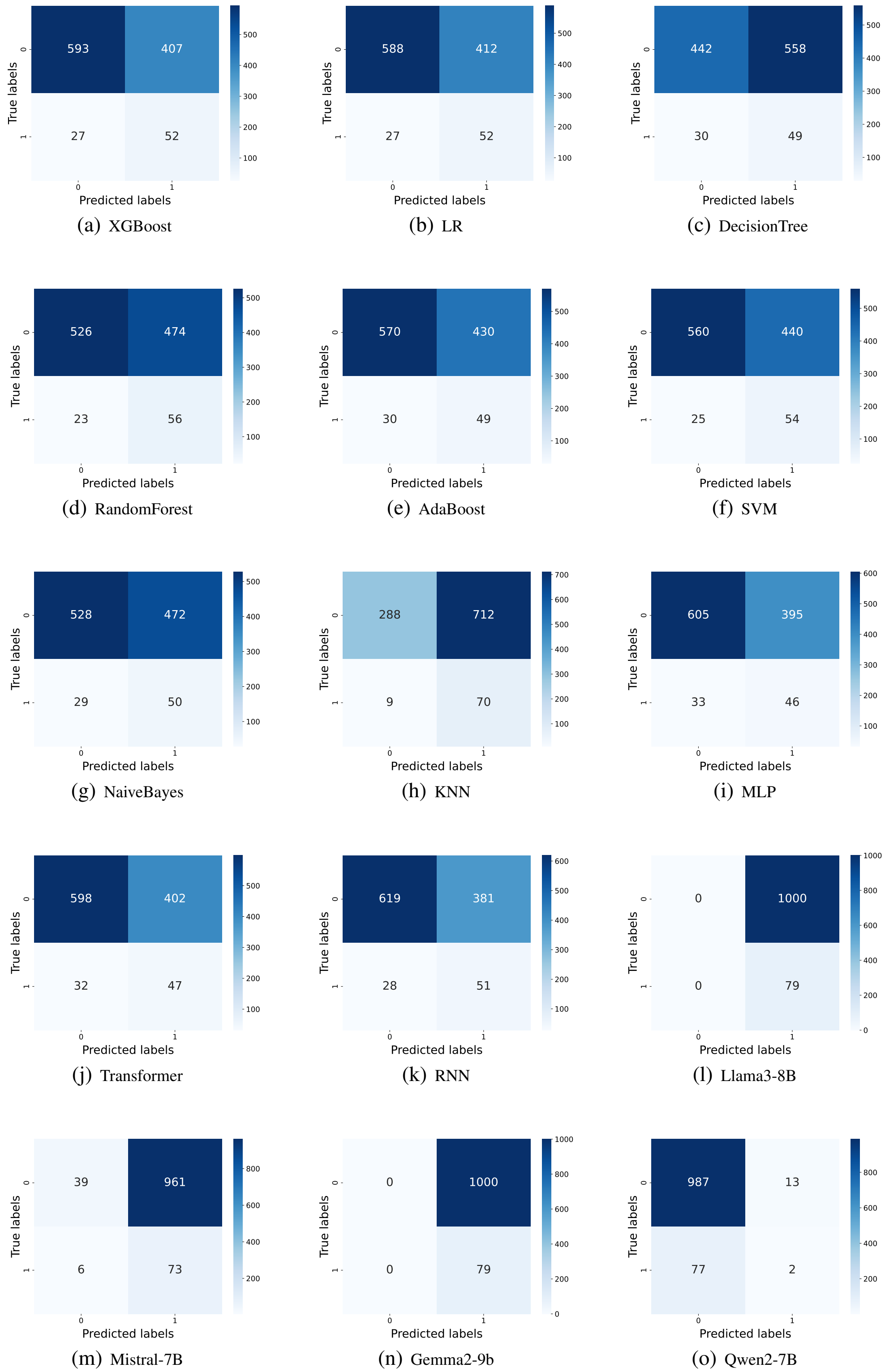


(j) BioMedGPT-7B



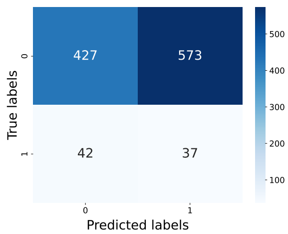
(k) Internist-7B

Figure 10: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Readmission Prediction on MIMIC-III Dataset**.

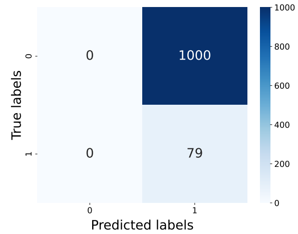


(a) XGBoost (b) LR (c) DecisionTree (d) RandomForest (e) AdaBoost (f) SVM (g) NaiveBayes (h) KNN (i) MLP (j) Transformer (k) RNN (l) Llama3-8B (m) Mistral-7B (n) Gemma2-9b (o) Qwen2-7B

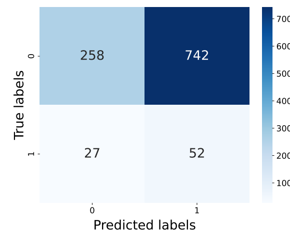
Figure 11: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Readmission Prediction on MIMIC-III Dataset**.



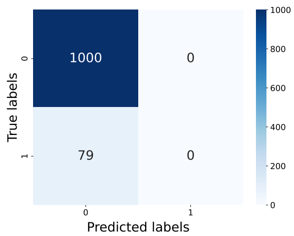
(a) Yi-1.5-9B



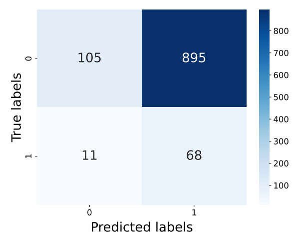
(b) Vicuna-7b-v1.5



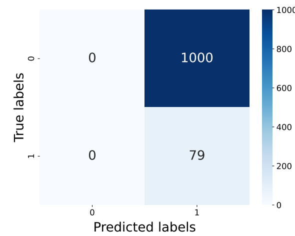
(c) Phi-3.5



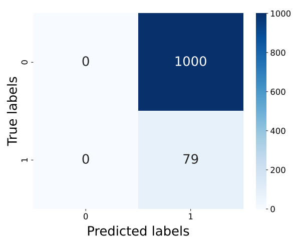
(d) internlm-7b



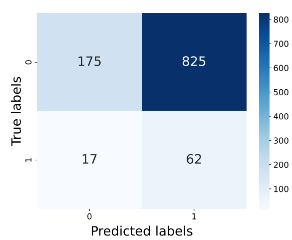
(e) MiniCPM3



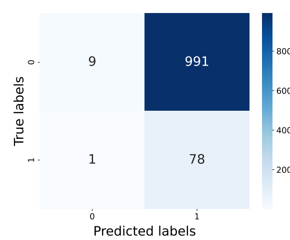
(f) meditron-7b



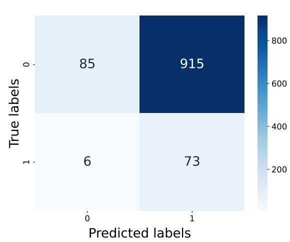
(g) Medllama3-8B



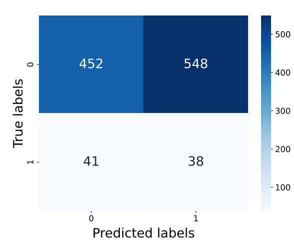
(h) BioMistral-7B



(i) Med42-8B

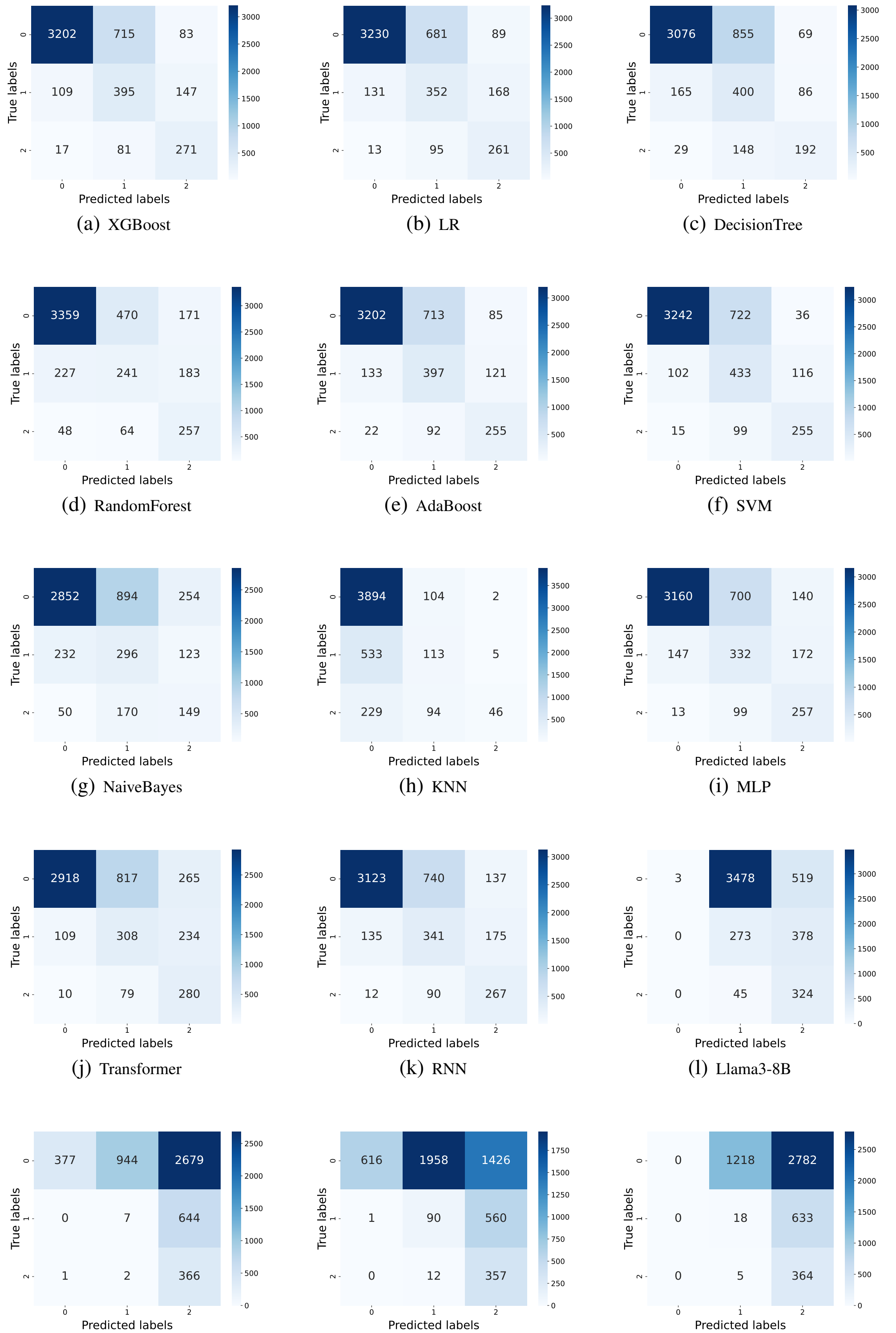


(j) BioMedGPT-7B



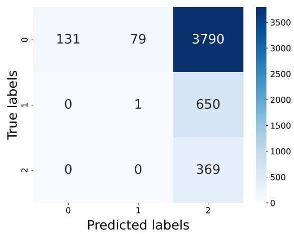
(k) Internist-7B

Figure 12: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Length-of-Stay Prediction on MIMIC-IV Dataset**.

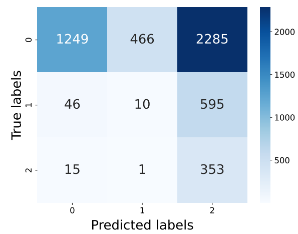


(a) XGBoost (b) LR (c) DecisionTree (d) RandomForest (e) AdaBoost (f) SVM (g) NaiveBayes (h) KNN (i) MLP (j) Transformer (k) RNN (l) Llama3-8B (m) Mistral-v0.3-7B (n) Gemma2-9B (o) Qwen2-7B

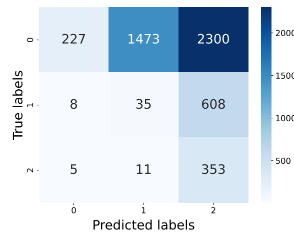
Figure 13: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Length-of-Stay Prediction on MIMIC-IV Dataset**.



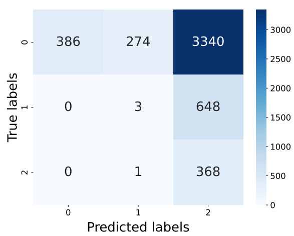
(a) Yi-v1.5-9B



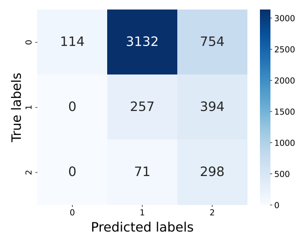
(b) Vicuna-v1.5-7B



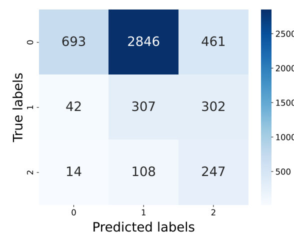
(c) Phi3.5-mini-3.8B



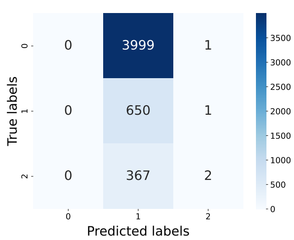
(d) InternLM2.5-7B



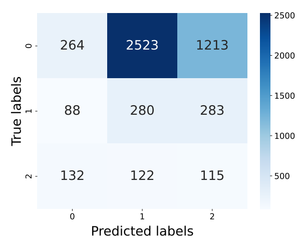
(e) MiniCPM3



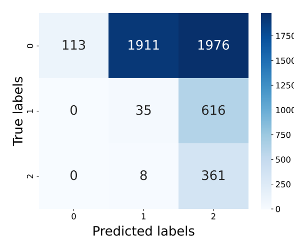
(f) Meditron-7B



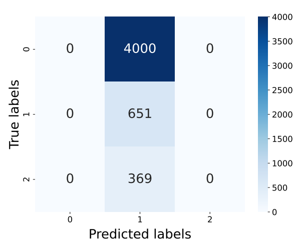
(g) Medllama3-8B



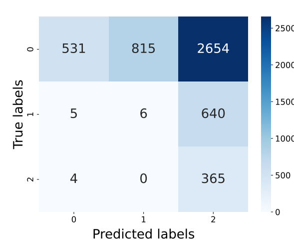
(h) BioMistral-7B



(i) Med42-8B

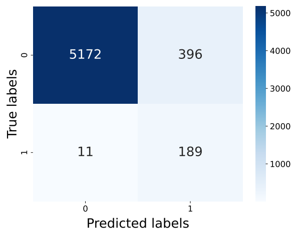


(j) BioMedGPT-7B

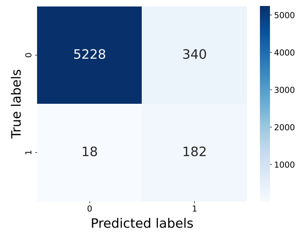


(k) Internist-7B

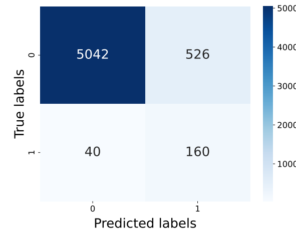
Figure 14: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Mortality Prediction on MIMIC-IV Dataset**.



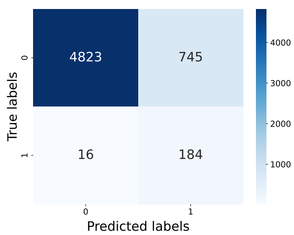
(a) XGBoost



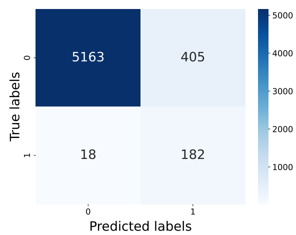
(b) LR



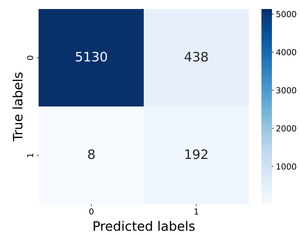
(c) DecisionTree



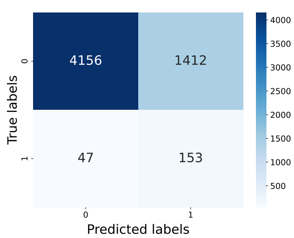
(d) RandomForest



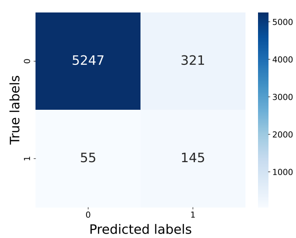
(e) AdaBoost



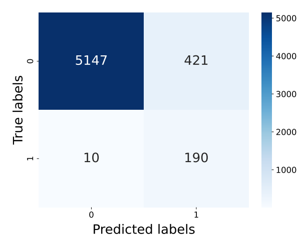
(f) SVM



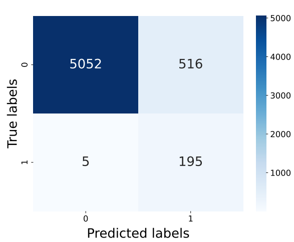
(g) NaiveBayes



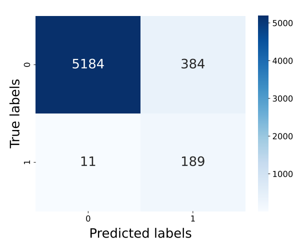
(h) KNN



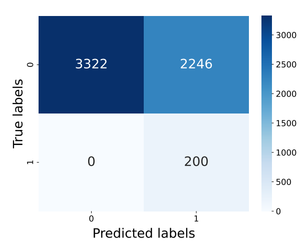
(i) MLP



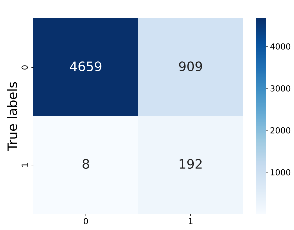
(j) Transformer



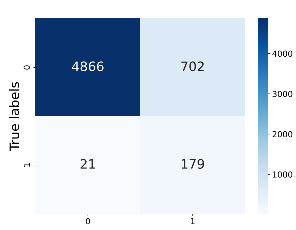
(k) RNN



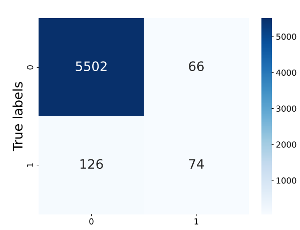
(l) Llama3-8B



(m) Mistral-v0.3-7B

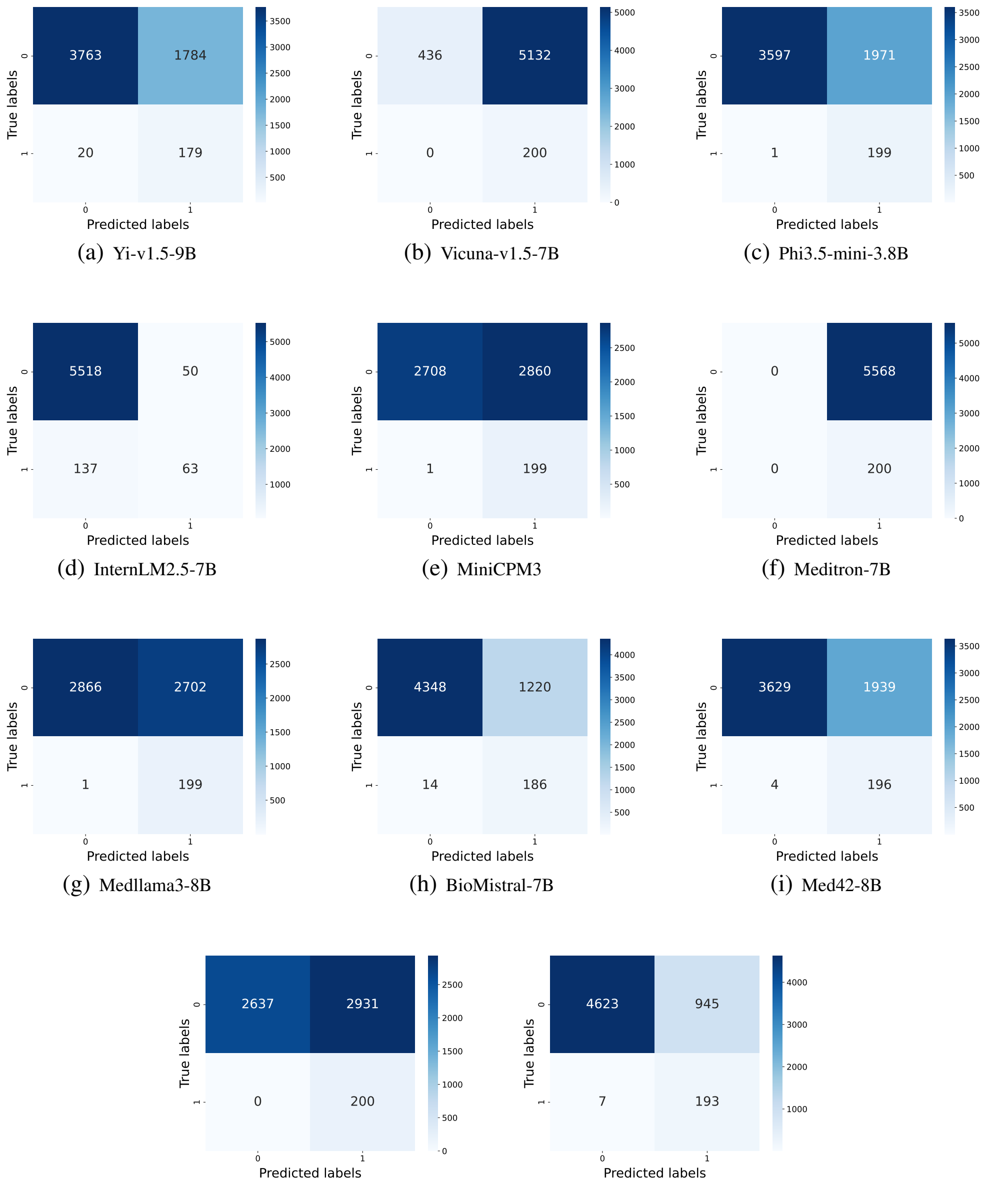


(n) Gemma2-9B



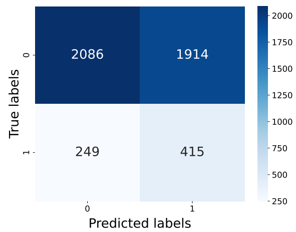
(o) Qwen2-7B

Figure 15: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Mortality Prediction on MIMIC-IV Dataset**.

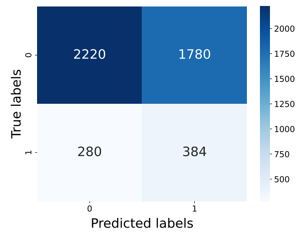


(a) Yi-v1.5-9B
(b) Vicuna-v1.5-7B
(c) Phi3.5-mini-3.8B
(d) InternLM2.5-7B
(e) MiniCPM3
(f) Meditron-7B
(g) Medllama3-8B
(h) BioMistral-7B
(i) Med42-8B
(j) BioMedGPT-7B
(k) Internist-7B

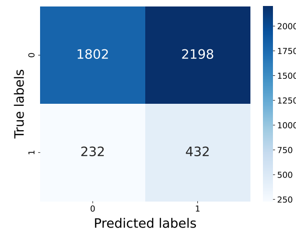
Figure 16: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Readmission Prediction on MIMIC-IV Dataset**.



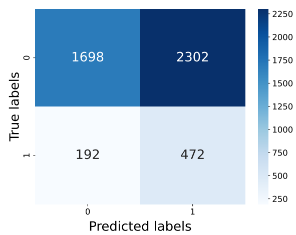
(a) XGBoost



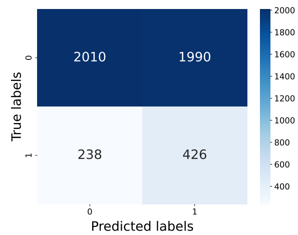
(b) LR



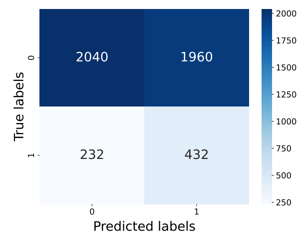
(c) DecisionTree



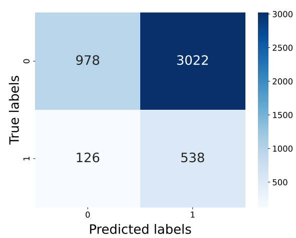
(d) RandomForest



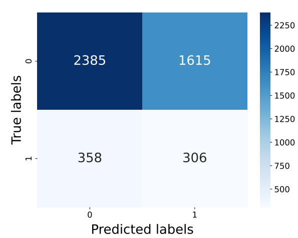
(e) AdaBoost



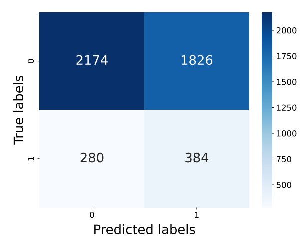
(f) SVM



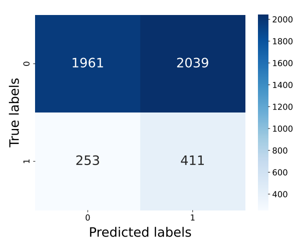
(g) NaiveBayes



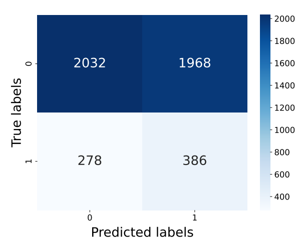
(h) KNN



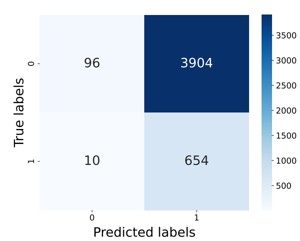
(i) MLP



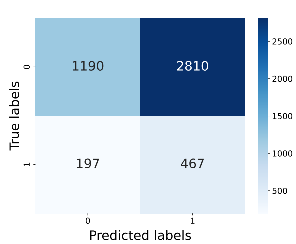
(j) Transformer



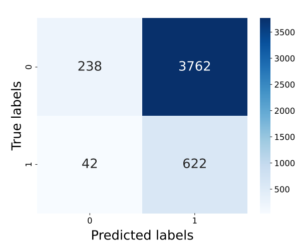
(k) RNN



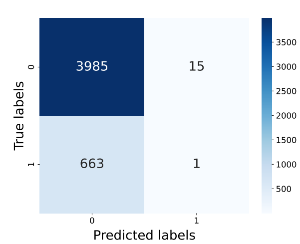
(l) Llama3-8B



(m) Mistral-v0.3-7B

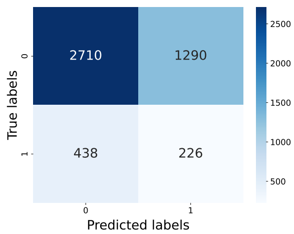


(n) Gemma2-9B

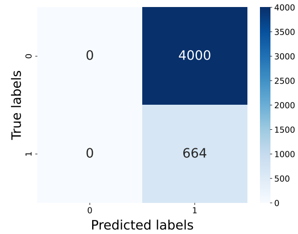


(o) Qwen2-7B

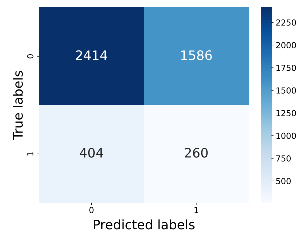
Figure 17: **Confusion Matrix of Traditional ML Models and Directly Prompting LLMs for Readmission Prediction on MIMIC-IV Dataset**.



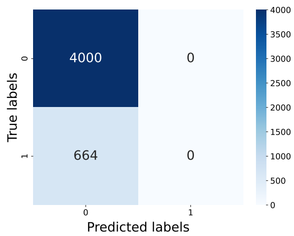
(a) Yi-v1.5-9B



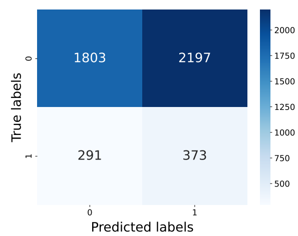
(b) Vicuna-v1.5-7B



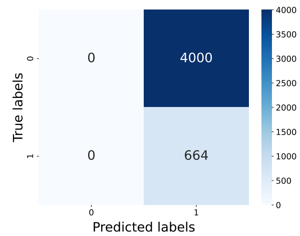
(c) Phi3.5-mini-3.8B



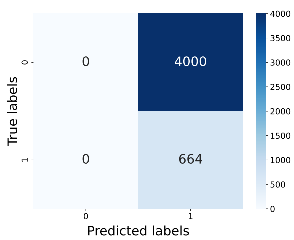
(d) InternLM2.5-7B



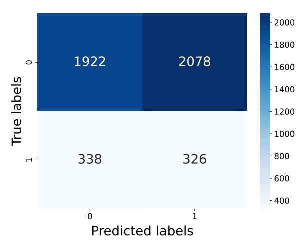
(e) MiniCPM3



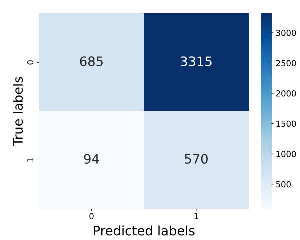
(f) Meditron-7B



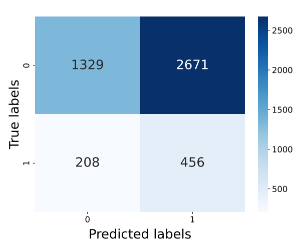
(g) Medllama3-8B



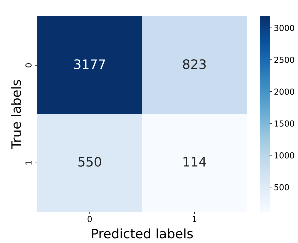
(h) BioMistral-7B



(i) Med42-8B



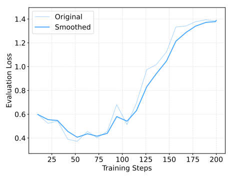
(j) BioMedGPT-7B



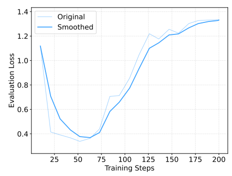
(k) Internist-7B

B.5 Loss Curves of Fine-tuning LLMs

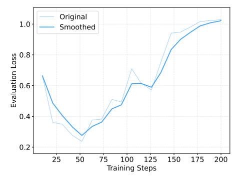
Figure 18: **Loss Curves of LoRA (full) for Length-of-Stay Prediction on MIMIC-III**.



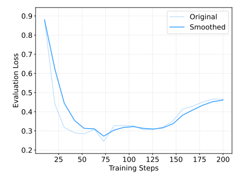
(a) Llama3-8B



(b) Gemma2-9B

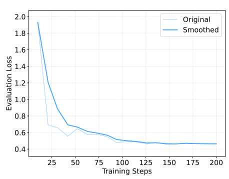


(c) Mistral-v0.3-7B

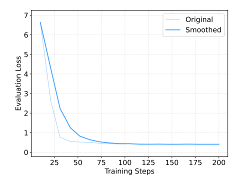


(d) Vicuna-v1.5-7B

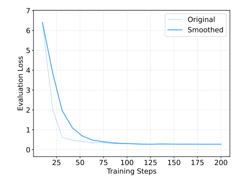
Figure 19: **Loss Curves of LoRA (last layer) for Length-of-Stay Prediction on MIMIC-III**.



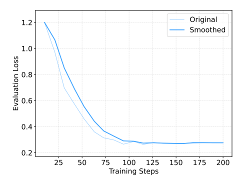
(a) Llama3-8B



(b) Gemma2-9B

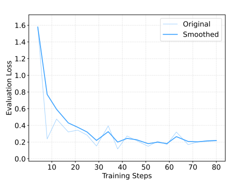


(c) Mistral-v0.3-7B

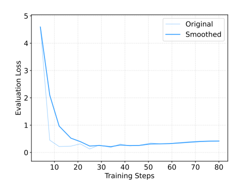


(d) Vicuna-v1.5-7B

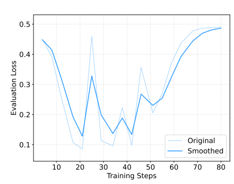
Figure 20: **Loss Curves of LoRA (full) for Mortality Prediction on MIMIC-III**.



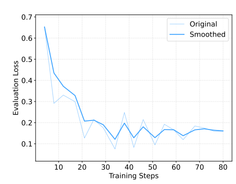
(a) Llama3-8B



(b) Gemma2-9B

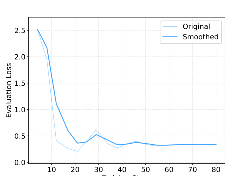


(c) Mistral-v0.3-7B

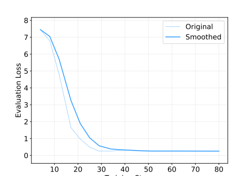


(d) Vicuna-v1.5-7B

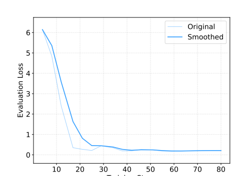
Figure 21: **Loss Curves of LoRA (last layer) for Mortality Prediction on MIMIC-III**.



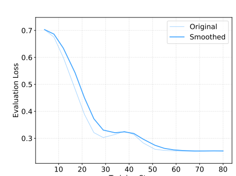
(a) Llama3-8B



(b) Gemma2-9B

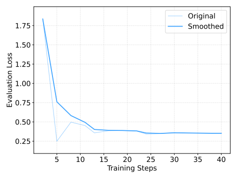


(c) Mistral-v0.3-7B

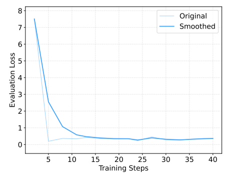


(d) Vicuna-v1.5-7B

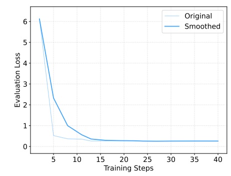
Figure 22: **Loss Curves of LoRA (full) for Readmission Prediction on MIMIC-III**.



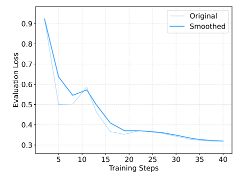
(a) Llama3-8B



(b) Gemma2-9B

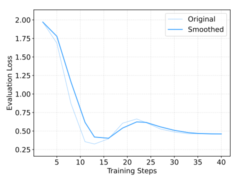


(c) Mistral-v0.3-7B

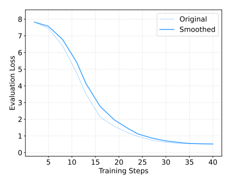


(d) Vicuna-v1.5-7B

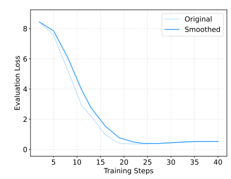
Figure 23: **Loss Curves of LoRA (last layer) for Readmission Prediction on MIMIC-III**.



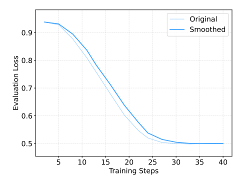
(a) Llama3-8B



(b) Gemma2-9B

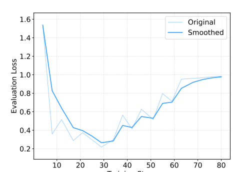


(c) Mistral-v0.3-7B

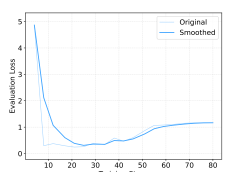


(d) Vicuna-v1.5-7B

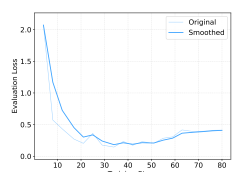
Figure 24: **Loss Curves of LoRA (full) for Length-of-Stay Prediction on MIMIC-IV**.



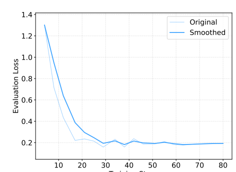
(a) Llama3-8B



(b) Gemma2-9B

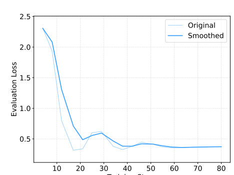


(c) Mistral-v0.3-7B

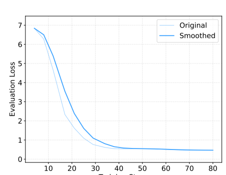


(d) Vicuna-v1.5-7B

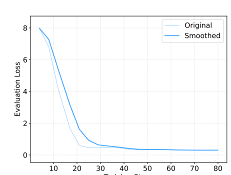
Figure 25: **Loss Curves of LoRA (last layer) for Length-of-Stay Prediction on MIMIC-IV**.



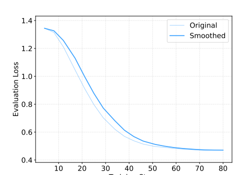
(a) Llama3-8B



(b) Gemma2-9B

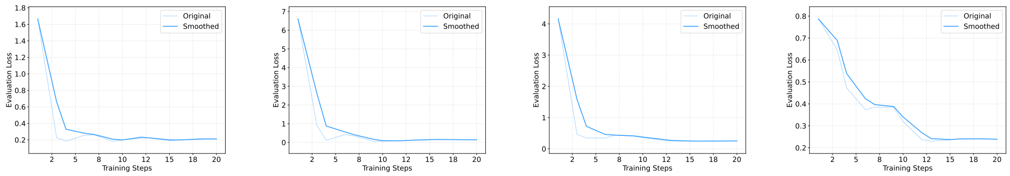


(c) Mistral-v0.3-7B



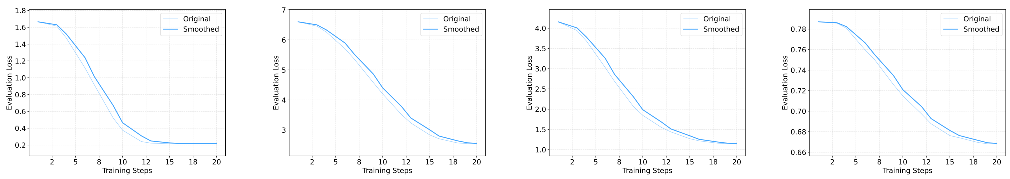
(d) Vicuna-v1.5-7B

Figure 26: **Loss Curves of LoRA (full) for Mortality Prediction on MIMIC-IV**.



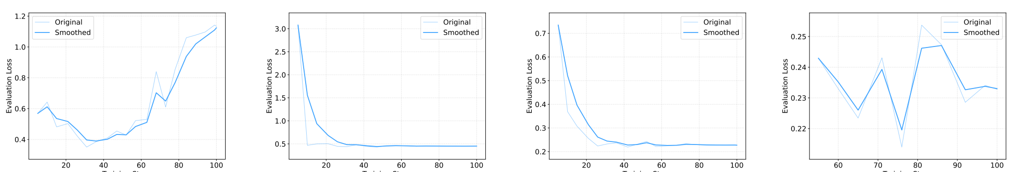
(a) Llama3-8B (b) Gemma2-9B (c) Mistral-v0.3-7B (d) Vicuna-v1.5-7B

Figure 27: **Loss Curves of LoRA (last layer) for Mortality Prediction on MIMIC-IV**.



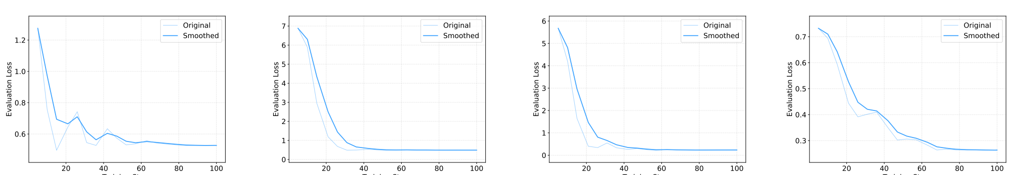
(a) Llama3-8B (b) Gemma2-9B (c) Mistral-v0.3-7B (d) Vicuna-v1.5-7B

Figure 28: **Loss Curves of LoRA (full) for Readmission Prediction on MIMIC-IV**.



(a) Llama3-8B (b) Gemma2-9B (c) Mistral-v0.3-7B (d) Vicuna-v1.5-7B

Figure 29: **Loss Curves of LoRA (last layer) for Readmission Prediction on MIMIC-IV**.



(a) Llama3-8B (b) Gemma2-9B (c) Mistral-v0.3-7B (d) Vicuna-v1.5-7B

C More Details of Fine-Tuning Data Construction for LLMs

Table 9: Example of Fine-Tuning Data Construction for Length-of-Stay Prediction

Instruction: Given the patient information, predict the number of weeks of stay in hospital. Answer 1 if no more than one week, Answer 2 if more than one week but not more than two weeks, Answer 3 if more than two weeks. Answer with only the number
Input: Patient information: Age: 78 Gender: male Conditions: Coronary atherosclerosis of native coronary artery, Chronic total occlusion of coronary artery, Malignant neoplasm of rectum, Personal history of antineoplastic chemotherapy, Person boarding or alighting a pedal cycle injured in collision with railway train or railway vehicle Procedures: Percutaneous transluminal coronary angioplasty [PTCA], Insertion of non-drug-eluting coronary artery stent(s), Left heart cardiac catheterization, Coronary arteriography using two catheters, Insertion of two vascular stents, Procedure on two vessels Using Drugs: other analgesics and antipyretics, lipid modifying agents, plain, antithrombotic agents, i.v. solution additives, antidepressants, potassium, viral vaccines, vasodilators used in cardiac diseases, belladonna and derivatives, plain Answer:
Output: 1

Table 10: Example of Fine-Tuning Data Construction for Mortality Prediction

Instruction: Given the patient information, predict the mortality of the patient. Answer 1 if the patient will die, answer 0 otherwise. Answer with only the number
Input: Patient information: Age: 64 Gender: female Conditions: Osteoarthrosis, localized, not specified whether primary or secondary, lower leg, Other and unspecified angina pectoris, Unspecified essential hypertension, Mitral valve disorders, Psoriatic arthropathy Procedures: Total knee replacement Using Drugs: selective calcium channel blockers with mainly vascular effects, beta blocking agents, lipid modifying agents, plain, vitamin a and d, incl. combinations of the two, antihistamines for systemic use, other nutrients, macrolides, lincosamides and streptogramins, anticholinergic agents, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation, opioids, antidepressants, other analgesics and antipyretics, i.v. solutions, antivaricose therapy, antiemetics and antinauseants, antiinflammatory and antirheumatic products, non-steroids, antacids, calcium, angiotensin ii receptor blockers (arbs), plain, antithrombotic agents Will the patient die because of the above situation? Answer:
Output: 0

Table 11: Example of Fine-Tuning Data Construction for Readmission Prediction

Instruction:
Given the patient information, predict the readmission of the patient.
Answer 1 if the patient will be readmitted to the hospital within two weeks, answer 0 otherwise.
Answer with only the number

Input:
Patient information:
Age: 91
Gender: female
Conditions: Open wound of forehead, without mention of complication, Hyposmolality and/or hyponatremia, Open wound of knee, leg [except thigh], and ankle, without mention of complication, Contusion of thigh, Contusion of scapular region, Open wound of scalp, without mention of complication, Unspecified essential hypertension, Insomnia, unspecified, Unspecified deficiency anemia, Acquired absence of both cervix and uterus, Acquired absence of intestine (large) (small), Personal history of malignant melanoma of skin, Fall from other slipping, tripping, or stumbling
Procedures: Closure of skin and subcutaneous tissue of other sites
Using Drugs: selective calcium channel blockers with mainly vascular effects, vitamin a and d, incl. combinations of the two, antidepressants, other analgesics and antipyretics, beta blocking agents, drugs for constipation, antithrombotic agents, antacids, calcium, i.v. solution additives
Will the patient be readmitted to the hospital within two weeks?
Answer:

Output:
0

D Examples of LLM-Based Clinical Prediction

D.1 Length-of-Stay Prediction

D.1.1 Directly Prompting

Table 12: Example of LLMs with Directly Prompting for Length-of-Stay Prediction.

Input:
Patient information:
Age: 62
Gender: male
Conditions: Streptococcal septicemia, Pressure ulcer, lower back, Acute osteomyelitis, pelvic region and thigh, Unknown Diagnosis, Cocaine abuse, unspecified, Acute kidney failure, unspecified, Hemorrhage complicating a procedure, Hyperpotassemia, Chronic osteomyelitis, pelvic region and thigh, Sepsis, Paraplegia, Unspecified essential hypertension, Colostomy status, Other artificial opening of urinary tract status, Anemia of other chronic disease, Other iatrogenic hypotension, Candidiasis of other urogenital sites, Late effects of injury purposely inflicted by other person
Procedures: Local excision of lesion or tissue of bone, other bones, Local excision of lesion or tissue of bone, femur, Nonexcisional debridement of wound, infection or burn, Venous catheterization, not elsewhere classified, Transfusion of packed cells
Using Drugs: anesthetics, general, i.v. solution additives, other antibacterials, anterior pituitary lobe hormones and analogues, antithrombotic agents, other analgesics and antipyretics, drugs for constipation, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), antiseptics and disinfectants, antiinfectives and antiseptics, excl. combinations with corticosteroids, other nutrients, opioids, other beta-lactam antibacterials, other mineral supplements, other diagnostic agents, antipsychotics, i.v. solutions, antivaricose therapy, beta-lactam antibacterials, penicillins
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer:

Output for Llama3-8B:
3

Output for Gemma2-9B:
3

Output for Meditron-7B:
3

Output for Internist-7B:
3

D.1.2 Chain-of-Thought Prompting

Table 13: Example of LLMs with Chain of Thought for Length-of-Stay Prediction.

Input:
Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension
Procedures: Partial shoulder replacement, Transfusion of packed cells
Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Please provide your concise reasoning steps for the prediction(no more than 3 steps), and finally answer 1 if the patient will stay no more than one week, answer 2 if more than one week but not more than two weeks, answer 3 if more than two weeks.

Output for Llama3-8B:
Please note that the patient's condition is quite complex and the prediction may not be accurate. This is a hypothetical scenario and not a real patient.
Answer: 3
Reasoning steps:
1. The patient has a closed fracture of the surgical neck of humerus, which is a serious injury that requires surgical intervention. This suggests that the patient will require a significant amount of care and treatment in the hospital.
2. The patient also has a number of other conditions, including pneumonia, acute posthemorrhagic anemia, and iatrogenic pulmonary embolism and infarction, which are all serious and potentially life-threatening. This suggests that the patient will require ongoing medical care and monitoring in the hospital.
3. Given the complexity of the patient's condition and the need for ongoing medical care and monitoring, it is likely that the patient will require more than two weeks of hospitalization.
Note: This is a hypothetical scenario and not a real patient. The actual length of stay may vary depending on a variety of factors, including the patient's response to treatment, the severity of their condition, and the availability of hospital resources.

Output for Gemma2-9B:

Continued on next page

Table 13 – continued from previous page

Reasoning: 1. **Severity of Conditions:** The patient has multiple serious conditions including congestive heart failure, mitral valve disorders, atrial fibrillation, chronic kidney disease, and diabetes. These require significant medical management. 2. **Complex Procedures:** The patient underwent major surgeries like mitral valve replacement, coronary bypass, and transvenous lead insertion. Recovery from these procedures takes considerable time. 3. **Post-Operative Care:** The patient received multiple transfusions, medications for various conditions, and likely required intensive monitoring post-surgery. This indicates a prolonged hospital stay for recovery and management. **Answer:** 3
Output for Meditron-7B: "," ## Prediction ## Prediction ## Prediction ## Prediction ## Prediction ## Prediction ## Prediction ## Prediction
Output for Internist-7B: Reasoning: 1. Analyze the patient's conditions and procedures: - The patient has multiple fractures and injuries, indicating a high likelihood of a prolonged hospital stay. - The patient has a history of diabetes and hyperlipidemia, which may require close monitoring and management. - The patient has undergone a partial shoulder replacement, which is a surgical procedure that typically requires a longer recovery period. 2. Consider the patient's medications and drug history: - The patient is taking multiple medications, including opioids, beta-blockers, and antithrombotic agents. - These medications may require close monitoring and adjustment during the hospital stay. 3. Evaluate the patient's overall health and risk factors: - The patient has a history of diabetes and hyperlipidemia, which may increase the risk of complications and prolong the hospital stay. - The patient has a history of other diseases of the circulatory system, which may also contribute to a longer hospital stay. Based on the above analysis, it is likely that the patient will require a longer hospital stay. Therefore, the answer is 3, indicating a stay of more than two weeks. It is important to note that this prediction is based on the information provided and may vary depending on the specific circumstances and progress of the patient's condition. Regular monitoring and assessment by healthcare professionals will be necessary to determine the actual duration of the hospital stay. Please note that this is a prediction based on the given information and may not be accurate in all cases. It is important to consider individual patient factors and consult with healthcare professionals for a more accurate assessment.

D.1.3 Self-Reflection Prompting

Table 14: Example of LLMs with Self Reflection for Length-of-Stay Prediction.

Input:
Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension
Procedures: Partial shoulder replacement, Transfusion of packed cells
Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
First answer with a number. Then conduct a concise reflection. Finally output your answer again with a number.

Output for Llama3-8B:
Answer: 3
Reflection: The patient has a complex medical history with multiple conditions, including a closed fracture of the surgical neck of humerus, pneumonia, and iatrogenic pulmonary embolism. They have undergone a partial shoulder replacement and have received multiple medications. Given the severity of their conditions and the need for ongoing medical care, it is likely that they will require a prolonged hospital stay of more than two weeks.
Answer: 3
“‘python
......

Output for Gemma2-9B:
Answer: 3
Reflection:
The patient presents with multiple severe injuries, including a closed fracture of the surgical neck of the humerus, facial bones, and a hematoma complicating a procedure. These injuries necessitate surgical intervention (partial shoulder replacement) and extensive post-operative care. The patient also has a history of diabetes, emphysema, and hypertension, which can complicate recovery. The presence of pneumonia and acute posthemorrhagic anemia further suggests a prolonged hospital stay.
Answer: 3

Output for Meditron-7B:

Continued on next page

Table 14 – continued from previous page

Answer 1
Answer 2
Answer 3
Answer 1
Answer 2
Answer 3
Answer 1
Answer 2
Answer 3
Answer 1
Answer 2
Answer 3
Answer 1
Answer 2
Answer 3

Output for Internist-7B:

Reasoning:To predict the number of weeks of stay in hospital, we need to analyze the patient's conditions, procedures, and medications.

1. Conditions: The patient has multiple conditions, including a closed fracture of the surgical neck of the humerus, other emphysema, closed fractures of other facial bones, closed fractures of the malar and maxillary bones, hematoma complicating a procedure, iatrogenic pulmonary embolism and infarction, acute posthemorrhagic anemia, pneumonia, and fall from other slipping, tripping, or stumbling. These conditions indicate that the patient may require a longer stay in the hospital.
2. Procedures: The patient has undergone a partial shoulder replacement and a transfusion of packed cells. These procedures may require a longer recovery period and may contribute to a longer hospital stay.
3. Medications: The patient is using various medications, including i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, insulins and analogues, blood glucose lowering drugs, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation. These medications may also contribute to a longer hospital stay.

Based on the patient's conditions, procedures, and medications, it is likely that the patient will require a longer stay in the hospital. Therefore, the correct answer is 3, indicating a stay of more......

D.1.4 Role-Playing Prompting

Table 15: Example of LLMs with Role-Playing for Length-of-Stay Prediction.

Input:
Imagine that you are a doctor. Today, you're seeing a patient with the following profile:
Patient information:
Age: 62
Gender: male
Conditions: Streptococcal septicemia, Pressure ulcer, lower back, Acute osteomyelitis, pelvic region and thigh, Unknown Diagnosis, Cocaine abuse, unspecified, Acute kidney failure, unspecified, Hemorrhage complicating a procedure, Hyperpotassemia, Chronic osteomyelitis, pelvic region and thigh, Sepsis, Paraplegia, Unspecified essential hypertension, Colostomy status, Other artificial opening of urinary tract status, Anemia of other chronic disease, Other iatrogenic hypotension, Candidiasis of other urogenital sites, Late effects of injury purposely inflicted by other person
Procedures: Local excision of lesion or tissue of bone, other bones, Local excision of lesion or tissue of bone, femur, Nonexcisional debridement of wound, infection or burn, Venous catheterization, not elsewhere classified, Transfusion of packed cells
Using Drugs: anesthetics, general, i.v. solution additives, other antibacterials, anterior pituitary lobe hormones and analogues, antithrombotic agents, other analgesics and antipyretics, drugs for constipation, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), antiseptics and disinfectants, antiinfectives and antiseptics, excl. combinations with corticosteroids, other nutrients, opioids, other beta-lactam antibacterials, other mineral supplements, other diagnostic agents, antipsychotics, i.v. solutions, antivaricose therapy, beta-lactam antibacterials, penicillins
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer:

Output for Llama3-8B:
3

Output for Gemma2-9B:
3

Output for Meditron-7B:
3

Output for Internist-7B:
3

D.1.5 In-Context Learning

Table 16: Example of LLMs with In Context Learning for Length-of-Stay Prediction.

Input:

Patient information:
Age: 74
Gender: male
Conditions: Subendocardial infarction, initial episode of care, Intestinal infection due to Clostridium difficile, Congestive heart failure, unspecified, Coronary atherosclerosis of native coronary artery, Unspecified essential hypertension, Pure hypercholesterolemia, Abdominal aneurysm without mention of rupture, Percutaneous transluminal coronary angioplasty status, Personal history of malignant neoplasm of large intestine
Procedures: Unknown Procedure, Combined right and left heart cardiac catheterization, Coronary arteriography using two catheters
Using Drugs: i.v. solution additives, cardiac stimulants excl. cardiac glycosides, other nutrients, antithrombotic agents, potassium, expectorants, excl. combinations with cough suppressants, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), hypnotics and sedatives, other analgesics and antipyretics, opioids, lipid modifying agents, plain, belladonna and derivatives, plain, other mineral supplements, other diagnostic agents, agents against amoebiasis and other protozoal diseases, anxiolytics, quinolone antibacterials, ace inhibitors, plain, beta blocking agents, other antibacterials
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one ,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer: 1

Patient information:
Age: 36
Gender: male
Conditions: Thoracic aneurysm without mention of rupture, Aortic valve disorders, Congenital insufficiency of aortic valve, Cardiac complications, not elsewhere classified, Other specified cardiac dysrhythmias, Unspecified essential hypertension
Procedures: Open and other replacement of aortic valve, Resection of vessel with replacement, thoracic vessels, Other operations on vessels of heart, Extracorporeal circulation auxiliary to open heart surgery
Using Drugs: drugs for functional gastrointestinal disorders, antiarrhythmics, class i and iii, drugs for constipation, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), other analgesics and antipyretics, antiinflammatory and antirheumatic products, non-steroids, opioids, other nutrients, anesthetics, general, antiinfectives and antiseptics, excl. combinations with corticosteroids, calcium, arteriolar smooth muscle, agents acting on, cardiac stimulants excl. cardiac glycosides, i.v. solution additives, other beta-lactam antibacterials, insulins and analogues, propulsives, hypnotics and sedatives, other mineral supplements, other diagnostic agents, beta blocking agents, high-ceiling diuretics, potassium, antithrombotic agents, antacids, urologicals, throat preparations, ace inhibitors, plain, angiotensin ii receptor blockers (arbs), plain
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer: 2

Table 17: Example of LLMs with In Context Learning for Length-of-Stay Prediction (Cont.).

Input (Cont.):
Patient information:
Age: 73
Gender: male
Conditions: Closed fracture of shaft of femur, Acute posthemorrhagic anemia, Subendocardial infarction, initial episode of care, Other postoperative infection, Unspecified septicemia, Severe sepsis, Septic shock, Unknown Diagnosis, Cardiogenic shock, Cardiac complications, not elsewhere classified, Paroxysmal ventricular tachycardia, Ventricular fibrillation, Congestive heart failure, unspecified, Atrial fibrillation, Unknown Diagnosis, Pneumonia, organism unspecified, Unspecified fall, Coronary atherosclerosis of native coronary artery, Cardiac pacemaker in situ
Procedures: Open reduction of fracture with internal fixation, femur, Closed reduction of fracture without internal fixation, femur, Left heart cardiac catheterization, Coronary arteriography using two catheters, Continuous invasive mechanical ventilation for 96 consecutive hours or more, Venous catheterization, not elsewhere classified, Enteral infusion of concentrated nutritional substances
Using Drugs: i.v. solution additives, opioids, other analgesics and antipyretics, beta blocking agents, selective calcium channel blockers with direct cardiac effects, antithrombotic agents, vasodilators used in cardiac diseases, agents for treatment of hemorrhoids and anal fissures for topical use, angiotensin ii receptor blockers (arbs), plain, antidepressants, drugs for constipation, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), anti-dementia drugs, lipid modifying agents, plain, antipsychotics, antipruritics, incl. antihistamines, anesthetics, etc., anxiolytics, antiarrhythmics, class i and iii, quinolone antibacterials, antacids, urologicals, expectorants, excl. combinations with cough suppressants, antiinfectives and antiseptics, excl. combinations with corticosteroids, calcium, other beta-lactam antibacterials, i.v. solutions, antivaricose therapy, other mineral supplements, other diagnostic agents, high-ceiling diuretics, anesthetics, general, muscle relaxants, peripherally acting agents, other antibacterials, other nutrients, hypnotics and sedatives, beta-lactam antibacterials, penicillins, cardiac stimulants excl. cardiac glycosides, anterior pituitary lobe hormones and analogues, other drugs for obstructive airway diseases, inhalants, adrenergics, inhalants, potassium, antifungals for topical use, ace inhibitors, plain
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer: 3

Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension

Table 18: Example of LLMs with In Context Learning for Length-of-Stay Prediction (Cont.).

Input (Cont.):
Procedures: Partial shoulder replacement, Transfusion of packed cells
Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Predict the number of weeks of stay in hospital.
Answer 1 if no more than one week,
Answer 2 if more than one week but not more than two weeks,
Answer 3 if more than two weeks.
Answer with only the number. Answer:

Output for Llama3-8B:
2

Output for Gemma2-9B:
2

Output for Meditron-7B:
2

Output for Internist-7B:
1

D.2 Mortality Prediction

D.2.1 Directly Prompting

Table 19: Example of LLMs with Directly Prompting for Mortality Prediction.

Input:
Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension
Procedures: Partial shoulder replacement, Transfusion of packed cells
Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Will the patient die because of the above situation?
Answer 1 if yes, 0 if no. Answer with only the number.
Answer:

Output for Llama3-8B:
1

Output for Gemma2-9B:
0

Output for Meditron-7B:
1

Output for Internist-7B:
0

D.2.2 Chain-of-Thought Prompting

Table 20: Example of LLMs with Chain of Thought for Length-of-Stay Prediction.

Input:
Patient information:
Age: 61
Gender: male
Conditions: Unspecified septicemia, Acute pancreatitis, Alcohol abuse, unspecified, Acute kidney failure, unspecified, Acute and subacute necrosis of liver, Acidosis, Defibrination syndrome, Shock, unspecified, Unknown Diagnosis, Unspecified protein-calorie malnutrition, Acute respiratory failure, Severe sepsis, Anemia, unspecified, Acute alcoholic hepatitis, Personal history of noncompliance with medical treatment, presenting hazards to health, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Long-term (current) use of insulin, Abdominal pain, other specified site
Procedures: Continuous invasive mechanical ventilation for less than 96 consecutive hours, Insertion of endotracheal tube, Arterial catheterization, Venous catheterization, not elsewhere classified, Transfusion of packed cells
Using Drugs: insulins and analogues, other nutrients, i.v. solution additives, other beta-lactam antibacterials, anxiolytics, anterior pituitary lobe hormones and analogues, quinolone antibacterials, other antibacterials, vitamin b1, plain and in combination with vitamin b6 and b12, other diagnostic agents, vitamin b12 and folic acid, vitamin k and other hemostatics, other mineral supplements, beta-lactam antibacterials, penicillins, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), hypnotics and sedatives, anesthetics, general, posterior pituitary lobe hormones, calcium, cardiac stimulants excl. cardiac glycosides
Will the patient die because of the above situation?
Please provide your concise reasoning steps for the prediction(no more than 3 steps), and finally answer 1 if the patient will die and 0 otherwise."
Please provide your concise reasoning steps for the prediction(no more than 3 steps), and finally answer 1 if the patient will die and 0 otherwise

---

Output for Llama3-8B:
Here is the reasoning steps:
Step 1: The patient has severe sepsis, acute kidney failure, and liver failure, which are all life-threatening conditions.
Step 2: The patient also has a history of non-compliance with medical treatment, which may worsen the patient's condition.
Step 3: Given the severity of the patient's conditions and the potential for non-compliance with treatment, it is likely that the patient will die.
Answer: 1

Output for Gemma2-9B:
Reasoning:
1. **Severe Sepsis and Multi-Organ Failure:** The patient presents with severe sepsis, a life-threatening condition, and multiple organ failures (pancreas, liver, kidneys, respiratory system).
2. **Complicating Factors:** Alcohol abuse, diabetes, and malnutrition further weaken the patient's immune system and ability to fight infection.
3. **Aggressive Treatment:** The patient is receiving intensive care interventions (ventilation, transfusions, multiple medications), indicating a critical condition requiring urgent medical attention.
Answer: 1
Disclaimer: I am an AI chatbot and cannot provide medical advice. This prediction is based on the information provided and should not be considered a definitive prognosis. The patient's actual outcome depends on various factors and individual medical management.

Output for Meditron-7B:

Continued on next page

Table 20 – continued from previous page

1. The patient is 83 years old.
2. The patient has pneumonia.
3. The patient has organism unspecified.
4. The patient has postinflammatory pulmonary fibrosis.
5. The patient has atrial fibrillation.
6. The patient has congestive heart failure, unspecified.
7. The patient has acute respiratory failure.
8. The patient has coronary atherosclerosis of native coronary artery.
9. The patient has esophageal reflux.
10. The patient is using macrolides, lincosamides and streptogramins.
11. The patient is using other beta-lactam antibacterials.
12. The patient is using expectorants, excl. combinations with cough suppressants.

Output for Internist-7B:
Reasoning:
1. Assess the patient's overall condition: The patient has multiple comorbidities, including septicemia, acute pancreatitis, alcohol abuse, acute kidney failure, acute and subacute necrosis of liver, acidosis, defibrination syndrome, shock, unspecified, unknown diagnosis, unspecified protein-calorie malnutrition, acute respiratory failure, severe sepsis, anemia, unspecified, acute alcoholic hepatitis, personal history of noncompliance with medical treatment, presenting hazards to health, diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, long-term (current) use of insulin, and abdominal pain.
2. Consider the patient's medical history and current condition: The patient has a history of noncompliance with medical treatment, presenting hazards to health, and long-term use of insulin. The patient also has multiple comorbidities, including septicemia, acute pancreatitis, alcohol abuse, acute kidney failure, acute and subacute necrosis of liver, acidosis, defibrination syndrome, shock, unspecified, unknown diagnosis, unspecified protein-calorie malnutrition, acute respiratory failure, severe sepsis, anemia, unspecified, acute alcoholic hepatitis, diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, and abdominal pain.
3. Evaluate the patient's current treatment and procedures: The patient has undergone continuous invasive mechanical ventilation for less than 96 consecutive hours, insertion of an endotracheal tube, arterial catheterization, venous catheterization, not elsewhere classified, transfusion of packed cells, and has been using various medications, including insulins and analogues, other nutrients, i.v. solution additives, other beta-lactam antibacterials, anxiolytics, anterior pituitary lobe hormones and analogues, quinolone antibacterials, other antibacterials, vitamin b1, plain and in combination with vitamin b.....

D.2.3 Self-Reflection Prompting

Table 21: Example of LLMs with Self Reflection for Mortality Prediction.

Input:
Patient information:
Age: 61
Gender: male
Conditions: Unspecified septicemia, Acute pancreatitis, Alcohol abuse, unspecified, Acute kidney failure, unspecified, Acute and subacute necrosis of liver, Acidosis, Defibrination syndrome, Shock, unspecified, Unknown Diagnosis, Unspecified protein-calorie malnutrition, Acute respiratory failure, Severe sepsis, Anemia, unspecified, Acute alcoholic hepatitis, Personal history of noncompliance with medical treatment, presenting hazards to health, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Long-term (current) use of insulin, Abdominal pain, other specified site
Procedures: Continuous invasive mechanical ventilation for less than 96 consecutive hours, Insertion of endotracheal tube, Arterial catheterization, Venous catheterization, not elsewhere classified, Transfusion of packed cells
Using Drugs: insulins and analogues, other nutrients, i.v. solution additives, other beta-lactam antibacterials, anxiolytics, anterior pituitary lobe hormones and analogues, quinolone antibacterials, other antibacterials, vitamin b1, plain and in combination with vitamin b6 and b12, other diagnostic agents, vitamin b12 and folic acid, vitamin k and other hemostatics, other mineral supplements, beta-lactam antibacterials, penicillins, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), hypnotics and sedatives, anesthetics, general, posterior pituitary lobe hormones, calcium, cardiac stimulants excl. cardiac glycosides
Will the patient die because of the above situation?
Answer 1 if yes, 0 if no.
First answer with a number. Then conduct a concise reflection. Finally output your answer again with a number.

Output for Llama3-8B:
Answer: 1
Reflection: The patient has a multitude of severe conditions, including septicemia, acute pancreatitis, and acute kidney failure, which are all life-threatening. Additionally, the patient has a history of non-compliance with medical treatment, which may have contributed to the severity of their condition. The patient is also receiving multiple medications, including antibiotics, insulin, and other treatments, which may be helping to manage their condition. However, the patient's overall prognosis is poor, and it is likely that they will die due to the severity of their condition.
Answer: 1
"'python
# Define the patient information as a dictionary
patient_info = {
""age"": 61,
......

Output for Gemma2-9B:
0
It is impossible to determine with certainty whether a patient will die based solely on the provided information. While the patient presents with a multitude of severe and life-threatening conditions, including septicemia, pancreatitis, kidney failure, liver necrosis, and shock, the course of these conditions and the effectiveness of treatment are crucial factors. The patient's age, medical history, and response to interventions all play a role in the prognosis. A definitive answer requires a comprehensive medical evaluation and ongoing monitoring.
0

Output for Meditron-7B:

Continued on next page

1
2
3
4
5
6
7
8
9
10
11
12
13
14

Output for Internist-7B:

Reasoning:The patient in this scenario has multiple medical conditions and has undergone several procedures. The question asks if the patient will die because of these factors.

To determine the answer, we need to consider the patient's medical history, current conditions, and the procedures they have undergone.

The patient has a history of alcohol abuse, acute pancreatitis, acute kidney failure, acute and subacute necrosis of the liver, acidosis, defibrination syndrome, shock, unspecified protein-calorie malnutrition, acute respiratory failure, severe sepsis, anemia, acute alcoholic hepatitis, and a personal history of noncompliance with medical treatment.

The patient has also undergone several procedures, including continuous invasive mechanical ventilation for less than 96 consecutive hours, insertion of an endotracheal tube, arterial catheterization, venous catheterization, and transfusion of packed cells.

Given the patient's multiple medical conditions and the procedures they have undergone, it is likely that the patient's overall health is compromised. The combination of these factors increases the risk of complications and potentially death.

Therefore, the answer is 1, indicating that the patient may die due to the above situation.

Reflecting on the answer, it is important to consider the patient's medical history, current conditions, and the procedures they have undergone. The combination of these factors increases the risk of complications and potentially death. It is crucial to provide appropriate medical care and support to the patient to improve their chances of survival.

Outputting the answer again with a number, the answer is still 1, indicating that the patient may die due to the above situation.

In summary, based on the patient's medical history, current conditions, and the procedures they have undergone, there is a high risk of complications and potentially death. It is important to provide appropriate medical care and support to the patient to improve their chances of survival. The answer is 1.

Note: It is important to consider the patient's overall health and medical history when determining the likelihood of death. This answer is based on the information provided in the scenario and may vary in real-life situations. Always consult with a healthcare professional for accurate medical......

D.2.4 Role-Playing Prompting

Table 22: Example of LLMs with Role-Playing for Mortality Prediction.

Input:
Imagine that you are a doctor. Today, you're seeing a patient with the following profile:
Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension
Procedures: Partial shoulder replacement, Transfusion of packed cells
Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Will the patient die because of the above situation?
Answer 1 if yes, 0 if no. Answer with only the number.
Answer:

Output for Llama3-8B:
1

Output for Gemma2-9B:
1

Output for Meditron-7B:
1

Output for Internist-7B:
1

D.2.5 In-Context Learning

Table 23: Example of LLMs with In Context Learning for Mortality Prediction.

Input:
Patient information:
Age: 43
Gender: female
Conditions: Coronary atherosclerosis of native coronary artery, Intermediate coronary syndrome, Diabetes mellitus without mention of complication, type I [juvenile type], not stated as uncontrolled, Unspecified essential hypertension, Pure hypercholesterolemia, Tobacco use disorder
Procedures: (Aorto)coronary bypass of two coronary arteries, Left heart cardiac catheterization, Extracorporeal circulation auxiliary to open heart surgery, Coronary arteriography using two catheters, Angiocardiography of left heart structures
Using Drugs: other analgesics and antipyretics, antipsychotics, vasodilators used in cardiac diseases, antacids, urologicals, anxiolytics, antidepressants, potassium, iron preparations, lipid modifying agents, plain, beta blocking agents, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), dopaminergic agents, thyroid preparations, opioids, other nutrients, cardiac stimulants excl. cardiac glycosides, drugs for constipation, i.v. solution additives, calcium, propulsives, anti-inflammatory and antirheumatic products, non-steroids, other antibacterials, high-ceiling diuretics, antithrombotic agents, other beta-lactam antibacterials, other mineral supplements Will the patient die because of the above situation?
Answer 1 if yes, 0 if no. Answer with only the number.
Answer: 0

Patient information:
Age: 86
Gender: male
Conditions: Intracerebral hemorrhage, Pneumonitis due to inhalation of food or vomitus, Unspecified essential hypertension, Aortocoronary bypass status, Coronary atherosclerosis of unspecified type of vessel, native or graft
Procedures: Continuous invasive mechanical ventilation for less than 96 consecutive hours
Using Drugs: beta blocking agents, vitamin b1, plain and in combination with vitamin b6 and b12, i.v. solution additives, antiepileptics, quinolone antibacterials, other antibacterials, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), other mineral supplements, other diagnostic agents, anxiolytics, anesthetics, general, opioids, antiemetics and antinauseants, hypnotics and sedatives
Will the patient die because of the above situation?
Answer 1 if yes, 0 if no. Answer with only the number.
Answer: 1

Patient information:
Age: 73
Gender: female
Conditions: Closed fracture of surgical neck of humerus, Other emphysema, Closed fracture of other facial bones, Closed fracture of malar and maxillary bones, Hematoma complicating a procedure, Iatrogenic pulmonary embolism and infarction, Acute posthemorrhagic anemia, Pneumonia, organism unspecified, Fall from other slipping, tripping, or stumbling, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Other specified surgical operations and procedures causing abnormal patient reaction, or later complication, without mention of misadventure at time of operation, Personal history of other diseases of circulatory system, Other and unspecified hyperlipidemia, Unspecified essential hypertension
Procedures: Partial shoulder replacement, Transfusion of packed cells

Continued on next page

Table 23 – continued from previous page

Using Drugs: i.v. solution additives, opioids, selective calcium channel blockers with mainly vascular effects, beta blocking agents, high-ceiling diuretics, antithrombotic agents, lipid modifying agents, plain, insulins and analogues, blood glucose lowering drugs, excl. insulins, other analgesics and antipyretics, antiemetics and antinauseants, i.v. solutions, antivaricose therapy, other beta-lactam antibacterials, antipsychotics, anxiolytics, adrenergics, inhalants, other drugs for obstructive airway diseases, inhalants, other antibacterials, corticosteroids for systemic use, plain, other diagnostic agents, vitamin b12 and folic acid, vitamin b1, plain and in combination with vitamin b6 and b12, other mineral supplements, quinolone antibacterials, antidepressants, potassium, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), drugs for constipation
Will the patient die because of the above situation?
Answer 1 if yes, 0 if no. Answer with only the number.
Answer:

Output for Llama3-8B:
1

Output for Gemma2-9B:
1

Output for Meditron-7B:
1

Output for Internist-7B:
1

D.3 Readmission Prediction

D.3.1 Directly Prompting

Table 24: Example of LLMs with Directly Prompting for Readmission Prediction.

Input:
Patient information:
Age: 76
Gender: female
Conditions: Other specified intestinal obstruction, Malignant neoplasm of anus, unspecified site, Stenosis of rectum and anus, Unknown Diagnosis, Coronary atherosclerosis of native coronary artery, Esophageal reflux, Unspecified essential hypertension, Personal history of malignant neoplasm of rectum, rectosigmoid junction, and anus
Procedures: Temporary colostomy, Combined right and left heart cardiac catheterization, Coronary arteriography using two catheters, Venous catheterization, not elsewhere classified
Using Drugs: opioids, antithrombotic agents, antiinfectives and antiseptics, excl. combinations with corticosteroids, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), other antibacterials, other beta-lactam antibacterials, beta blocking agents, other mineral supplements, i.v. solution additives, other diagnostic agents, other nutrients, other analgesics and antipyretics, belladonna and derivatives, plain, i.v. solutions, antivaricose therapy, antiemetics and antinauseants, drugs for constipation
Will the patient be readmitted to the hospital within two weeks?
Answer 1 for yes, 0 for no. Answer with only the number.
Answer:

Output for Llama3-8B:
1

Output for Gemma2-9B:
1

Output for Meditron-7B:
1

Output for Internist-7B:
1

D.3.2 Chain-of-Thought Prompting

Table 25: Example of LLMs with Chain of Thought for Readmission Prediction.

Input:
Patient information:
Age: 68
Gender: male
Conditions: Malignant neoplasm of other specified sites of bladder, Hemorrhage complicating a procedure
Procedures: Other transurethral excision or destruction of lesion or tissue of bladder, Control of (postoperative) hemorrhage of bladder
Using Drugs: antiemetics and antinauseants, opioids, belladonna and derivatives, plain, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, other diagnostic agents
Will the patient be readmitted to the hospital within two weeks?
Please provide your concise reasoning steps for the prediction(no more than 3 steps), and finally answer 1 if the patient will die and 0 otherwise

---

Output for Llama3-8B:
Here is the solution:
Step 1: The patient has undergone a transurethral excision or destruction of lesion or tissue of bladder, which is a surgical procedure. This increases the likelihood of complications and readmission.

Step 2: The patient has a history of hemorrhage complicating a procedure, which is a significant complication that may require further treatment and hospitalization.
Step 3: The patient is taking a variety of medications, including opioids, antihistamines, and antipsychotics, which may increase the risk of adverse reactions and complications.
Answer: 1 (The patient will be readmitted to the hospital within two weeks)
Note: This prediction is based on the patient's medical history and procedures, and may not be accurate in all cases. A more detailed review of the patient's medical records and current condition would be necessary to make a more accurate prediction.

- - -

Output for Gemma2-9B:
Reasoning:
1. **Procedure:** The patient underwent a procedure for bladder cancer (""Other transurethral excision or destruction of lesion or tissue of bladder"") and experienced postoperative hemorrhage.

2. **Complications:** Postoperative hemorrhage is a potential complication of bladder cancer procedures and can require further intervention.
3. **Drug Usage:** The patient is on a variety of medications, including opioids and antiemetics, which are often used for pain management and nausea after surgery. This suggests ongoing recovery and potential need for further care.
Answer: 1

- - -

Output for Meditron-7B:

Continued on next page

Table 25 – continued from previous page

1. The patient is male. ## 2. The patient is 68 years old. ## 3. The patient has a malignant neoplasm of other specified sites of bladder. ## 4. The patient has a hemorrhage complicating a procedure. ## 5. The patient has other transurethral excision or destruction of lesion or tissue of bladder. ## 6. The patient has control of (postoperative) hemorrhage of bladder. ## 7. The patient has antiemetics and antinauseants. ## 8. The patient has opioids. ## 9. The patient has belladonna and derivatives. ## 10. The patient has plain. ......
Output for Internist-7B: Reasoning: 1. Assess the patient's condition: The patient is a 68-year-old male with a malignant neoplasm of other specified sites of the bladder and hemorrhage complicating a procedure. This indicates that the patient has undergone a procedure related to the bladder and is experiencing bleeding as a complication. 2. Consider the procedures and medications: The patient has undergone a transurethral excision or destruction of lesion or tissue of the bladder and control of postoperative hemorrhage of the bladder. The patient is also using various medications, including antiemetics and antinauseants, opioids, belladonna and derivatives, antipsychotics, antihistamines, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, and other diagnostic agents. 3. Evaluate the risk factors: The patient's age, gender, and medical history are important factors to consider. The patient is 68 years old, which may increase the risk of complications and readmission. The patient's gender and medical history are also relevant, as they may contribute to the overall risk assessment. 4. Analyze the likelihood of readmission: Based on the patient's condition, procedures, medications, and risk factors, it is important to assess the likelihood of readmission within two weeks. This may involve considering the severity of the patient's condition, the effectiveness of the procedures and medications, and the potential for complications or recurrence of the bleeding. 5. Make the prediction: Based on the analysis, if the likelihood of readmission within two weeks is high, the answer would be 1. If the likelihood is low, the answer would be 0. In this case, the patient's condition, procedures, medications, and risk factors suggest a high likelihood of readmission within two weeks. Therefore, the answer is 1. Note: It is important to consider the specific patient's condition, procedures, medications, and risk factors when making a prediction. This explanation is a general guide and may not apply to every individual case.

D.3.3 Self-Reflection Prompting

Table 26: Example of LLMs with Self Reflection for Readmission Prediction.

Input:
Patient information:
Age: 68
Gender: male
Conditions: Malignant neoplasm of other specified sites of bladder, Hemorrhage complicating a procedure
Procedures: Other transurethral excision or destruction of lesion or tissue of bladder, Control of (postoperative) hemorrhage of bladder
Using Drugs: antiemetics and antinauseants, opioids, belladonna and derivatives, plain, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, other diagnostic agents
Will the patient be readmitted to the hospital within two weeks?
Answer 1 for yes, 0 for no. Answer with only the number.
First answer with a number. Then conduct a concise reflection. Finally output your answer again with a number.

Output for Llama3-8B:
Please enter your answer.
“‘
1
“‘
Reflection: Based on the patient's conditions and procedures, it is likely that the patient will be readmitted to the hospital within two weeks. The patient has a malignant neoplasm of the bladder, which may require further treatment or monitoring. Additionally, the patient has undergone a transurethral excision or destruction of lesion or tissue of the bladder, which may lead to complications or side effects that require hospitalization. The patient is also taking a number of medications, which may increase the risk of readmission.
Answer:
“‘
1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1
“‘ “‘ 1"

Output for Gemma2-9B:
Reflection:
The patient underwent a procedure for a malignant bladder tumor and experienced post-operative hemorrhage. This suggests a potentially complex recovery and a higher risk of complications. The use of various medications, including opioids and antiemetics, points towards ongoing pain management and potential nausea/vomiting.
While it's impossible to say for certain without more information, the patient's condition and the nature of the procedure increase the likelihood of readmission within two weeks.
Answer: 1

Continued on next page

Table 26 – continued from previous page

Output for Meditron-7B:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Output for Internist-7B:

Reasoning:The patient in this scenario is a 68-year-old male with a malignant neoplasm of other specified sites of the bladder and hemorrhage complicating a procedure. The patient has undergone a transurethral excision or destruction of lesion or tissue of the bladder and control of postoperative hemorrhage of the bladder. The patient is also using various medications, including antiemetics and antinauseants, opioids, belladonna and derivatives, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, and other diagnostic agents.

Based on this information, we need to determine if the patient will be readmitted to the hospital within two weeks. To do this, we need to consider the patient's condition, the procedures performed, and the medications being used.

First, the patient has a malignant neoplasm of other specified sites of the bladder, which indicates a serious underlying condition. This suggests that the patient may require ongoing medical management and follow-up.

Second, the patient has experienced hemorrhage complicating a procedure. This indicates that the patient may have experienced bleeding during the transurethral excision or destruction of lesion or tissue of the bladder. This complication may require further intervention or monitoring.

Third, the patient is using various medications, including antiemetics and antinauseants, opioids, belladonna and derivatives, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, and other diagnostic agents. These medications may have side effects or interactions that could potentially lead to complications or the need for further medical attention.

Considering all of these factors, it is likely that the patient may require readmission to the hospital within two weeks......

D.3.4 Role-Playing Prompting

Table 27: Example of LLMs with Role-Playing for Readmission Prediction.

Input:
Imagine that you are a doctor. Today, you're seeing a patient with the following profile:
Patient information:
Age: 68
Gender: male
Conditions: Malignant neoplasm of other specified sites of bladder, Hemorrhage complicating a procedure
Procedures: Other transurethral excision or destruction of lesion or tissue of bladder, Control of (postoperative) hemorrhage of bladder
Using Drugs: antiemetics and antinauseants, opioids, belladonna and derivatives, plain, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, other diagnostic agents
Will the patient be readmitted to the hospital within two weeks? Answer 1 for yes, 0 for no. Answer with only the number.
Answer:

Output for Llama3-8B:
1

Output for Gemma2-9B:
1

Output for Meditron-7B:
1

Output for Internist-7B:
1

D.3.5 In-Context Learning

Table 28: Example of LLMs with In Context Learning for Readmission Prediction.

Input:

Patient information:
Age: 31
Gender: female
Conditions: Acute respiratory failure, Myasthenia gravis with (acute) exacerbation, Other specified cardiac dysrhythmias, Diarrhea, Unspecified essential hypertension, Iron deficiency anemia, unspecified
Procedures: Continuous invasive mechanical ventilation for 96 consecutive hours or more, Insertion of endotracheal tube, Injection or infusion of immunoglobulin, Non-invasive mechanical ventilation, Enteral infusion of concentrated nutritional substances, Transfusion of packed cells
Using Drugs: immunosuppressants, corticosteroids for systemic use, plain, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), other nutrients, i.v. solutions, antivaricose therapy, anticholinergic agents, beta blocking agents, anesthetics, general, parasympathomimetics, other antidiarrheals, calcium, other analgesics and antipyretics, antithrombotic agents, antacids, potassium, i.v. solution additives, sulfonamides and trimethoprim, ace inhibitors, plain, antipropulsives, antidepressants, belladonna and derivatives, plain, anxiolytics, hypnotics and sedatives, other cardiac preparations, antiseptics and disinfectants, antiepileptics
Will the patient be readmitted to the hospital within two weeks?
Answer 1 for yes, 0 for no. Answer with only the number.
Answer: 0

Patient information:
Age: 62
Gender: male
Conditions: Acute myocardial infarction of unspecified site, initial episode of care, Congestive heart failure, unspecified, Unknown Diagnosis, Acute kidney failure with lesion of tubular necrosis, Other and unspecified complications of medical care, not elsewhere classified, Pneumonia, organism unspecified, Late effects of cerebrovascular disease, hemiplegia affecting unspecified side, Unspecified pleural effusion, Unknown Diagnosis, Anticoagulants causing adverse effects in therapeutic use, Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled, Anemia, unspecified, Atrial fibrillation, Attention to tracheostomy, Unspecified essential hypertension, Pure hypercholesterolemia, Coronary atherosclerosis of unspecified type of vessel, native or graft, Aortocoronary bypass status
Procedures: Thoracentesis, Enteral infusion of concentrated nutritional substances, Infusion of drotrecogin alfa (activated), Continuous invasive mechanical ventilation for less than 96 consecutive hours
Using Drugs: antithrombotic agents, beta blocking agents, antiarrhythmics, class i and iii, other nutrients, lipid modifying agents, plain, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord), other analgesics and antipyretics, other beta-lactam antibacterials, other antibacterials, i.v. solution additives, hypnotics and sedatives, anesthetics, general, agents against amoebiasis and other protozoal diseases, quinolone antibacterials, ace inhibitors, plain, high-ceiling diuretics, other antidiarrheals, antacids, insulins and analogues, decongestants and other nasal preparations for topical use, intestinal antiinfectives, calcium, antifungals for topical use, anxiolytics, beta-lactam antibacterials, penicillins, adrenergics, inhalants, antidepressants, potassium
Will the patient be readmitted to the hospital within two weeks?
Answer 1 for yes, 0 for no. Answer with only the number.
Answer: 1

Patient information:

Continued on next page

Table 28 – continued from previous page

Age: 68
Gender: male
Conditions: Malignant neoplasm of other specified sites of bladder, Hemorrhage complicating a procedure
Procedures: Other transurethral excision or destruction of lesion or tissue of bladder, Control of (postoperative) hemorrhage of bladder
Using Drugs: antiemetics and antinauseants, opioids, belladonna and derivatives, plain, antipsychotics, antihistamines for systemic use, quinolone antibacterials, i.v. solutions, antivaricose therapy, hypnotics and sedatives, other analgesics and antipyretics, urologicals, other mineral supplements, i.v. solution additives, other diagnostic agents
Will the patient be readmitted to the hospital within two weeks? Answer 1 for yes, 0 for no. Answer with only the number.
Answer:

Output for Llama3-8B:
0

Output for Gemma2-9B:
0

Output for Meditron-7B:
0

Output for Internist-7B:
0