UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Deep Learning in Medical Applications

Permalink

https://escholarship.org/uc/item/5sk3p53c

Author

Zhang, Xinlu

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Deep Learning in Medical Applications

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Xinlu Zhang

Committee in charge:

Professor Linda Ruth Petzold, Chair
Professor William Yang Wang
Professor Lei Li

September 2024

The Dissertation of Xinlu Zhang is approved.

Professor William Yang Wang

Professor Lei Li

Professor Linda Ruth Petzold, Committee Chair

July 2024

Deep Learning in Medical Applications

Copyright © 2024

by

Xinlu Zhang

To my family, who have been my constant support through all of the challenges I have faced in life.

Acknowledgements

My deepest gratitude is reserved for my advisor, Prof. Linda Petzold, without whom I would not have been able to start and finish my PhD. I am grateful she introduced me to the fantastic field of deep learning. She is a supportive and kind lady who provided great support and freedom in my research journey. During these years, she has profoundly influenced my way of thinking not only about academia but also about life. To this day, I can still remember her advice, encouragement, and understanding throughout my entire PhD journey.

I would like to thank Prof. William Yang Wang and Prof. Lei Li for serving on my committee. They provided invaluable comments and feedback at every stage of my PhD career over the years.

I would also like to extend my thanks to all my collaborators. I have been truly fortunate to meet and work with many excellent individuals: Shiyang Li, Xianjun Yang, Zhiyu Chen, Chenxin Tian, Lichang Chen, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Zekun Li, Zhuowei Cheng, Yun Zhao, Rachael Callcut, Xifeng Yan, and Yao Qin. They have provided significant help in my work, and I learned a lot from them.

I would like to thank my colleagues and friends for all the helpful discussions, career advice, and emotional support. Here, I want to thank Feifei Zhao, Fangzheng Zhao, Qianyu Li, Ming Yin, Yingrui Yang, Hong Wang, Dengxian Yang, Yuqing Wang, and Zikun Yang, among many others whom I cannot list here, for the wonderful time we spent together.

Lastly, I want to thank my family members for their constant support and endless love, as always. To my husband, thank you for being by my side all the time with unwavering support. To my parents, thank you for your unconditional love.

Curriculum Vitæ

Xinlu Zhang

Education

Ph.D.in Computer Science	Sep. 2020 - July 2024
University of California, Santa Barbara	Advisor: Prof. Linda Petzold
M.S. in Statistics	Aug. 2018 - May 2020
Indiana University, Bloomington	GPA: 3.97/4.00
B.S. in Statistics and Mathematics	Aug. 2015 - July 2018
Indiana University, Bloomington, IN	GPA: 3.91/4.00

Preprints and Publications

1. **Unveiling the Impact of Coding Data Instruction Fine-Tuning on Large Language Models Reasoning**
   Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Lichang Chen, Xianjun Yang, William Yang Wang, and Linda Petzold
   Under Review

2. **A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law**
   Zhiyu Zoey Chen, Jing Ma, **Xinlu Zhang**, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang
   Under Review

3. **GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks**
   Xinlu Zhang*, Yujie Lu*, Weizhi Wang*, An Yan, Jun Yan, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Petzold (*Equal Contribution)
   Preprint

4. **AlpaCare: Instruction-tuned Large Language Models for Medical Applications**
   Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Petzold
   Preprint

5. **Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization**
   Xianjun Yang, Yan Li, **Xinlu Zhang**, Haifeng Chen, and Wei Wang
   Preprint

6. **Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting**
   Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Petzold
   ICLR 2024
7. **Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling**
   Xinlu Zhang*, Shiyang Li*, Zhiyu Chen, Xifeng Yan, and Linda Petzold (*Equal Contribution)
   ICML 2023
8. **Explanations from Large Language Models Make Small Reasoners Better**
   Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, **Xinlu Zhang**, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan
   SAI-AAAI Workshop 2024
9. **PcMSP: A Dataset for Scientific Action Graphs Extraction from Polycrystalline Materials Synthesis Procedure Text**
   Xianjun Yang, Ya Zhuo, Julia Zuo, **Xinlu Zhang**, Stephen Wilson, and Linda Petzold
   Findings of EMNLP 2022
10. **Domain Adaptation for Trauma Mortality Prediction in EHRs with Feature Disparity**
    Xinlu Zhang, Shiyang Li, Zhuowei Cheng, Rachael Callcut, and Linda Petzold
    Proceedings of BIBM 2021
11. **An Analysis of Relation Extraction from Wet Lab Protocols**
    Xianjun Yang, **Xinlu Zhang**, Julia Zuo, Stephen Wilson, and Linda Petzold
    Proceedings of BigData 2021
12. **Multiple Organ Failure Prediction with Classifier-Guided Generative Adversarial Imputation Networks**
    Xinlu Zhang*, Yun Zhao*, Rachael Callcut, and Linda Petzold (*Equal Contribution)
    KDD-BIOKDD Wrokshop 2021
13. **BERTSurv: BERT-Based Survival Models for Predicting Outcomes of Trauma Patients**
    Yun Zhao, Qinghang Hong, **Xinlu Zhang**, Yu Deng, Yuqing Wang, Paul K. Hansma, and Linda Petzold
    Proceedings of ICDM 2021

Work Experience

Research Scientist Intern	Sep. 2023 - Dec. 2023
Meta, Burlingame, CA	
Machine Learning Engineer Intern	June 2023 - Sep. 2023
LinkedIn, Sunnyvale, CA	
Machine Learning Engineer Intern	June 2022 - Sep. 2022
LinkedIn, Sunnyvale, CA	
Data Scientist Intern	May 2019 - Aug. 2019
The Bee Corp., Indianapolis, IN	

Abstract

Deep Learning in Medical Applications

by

Xinlu Zhang

The rapid advancement of deep learning has significantly impacted the medical domain, benefitting various applications including clinical decision-making, personalized treatment, and medical education. Deep learning applications in the medical domain can be categorized based on the data types used: 1) Numerical measurements modeling: building models on numerical clinical measurements, including static and time-series data; 2) Natural Language Processing (NLP): training models on medical textual data such as doctor-patient conversations and clinical notes; and 3) Multimodal learning: leveraging data from multiple modalities to enhance the model's medical capacity and performance. This thesis presents works in these three categories, aiming to advance AI systems that can assist clinicians in enhancing healthcare outcomes and efficiency.

In numerical measurements modeling, despite the effectiveness of deep learning models in decision support, many studies rely on extensive public datasets, overlooking the data scarcity in small hospital settings. We address this by utilizing domain adaptation techniques to improve modality prediction in ICU patients with limited data.

Concerning NLP, while Large Language Models (LLMs) like ChatGPT and GPT-4 have shown promising results, privacy concerns restrict their direct use in healthcare. We propose integrating medical knowledge from LLMs into local models for decision support to alleviate these privacy concerns. Furthermore, instruction tuning has become crucial in aligning LLMs with human intents and has shown potential in medical applications. However, existing medical LLMs ignore the diversity of tuning data, limiting their ability to follow medical instructions and generalize. This thesis presents a novel approach to generating a diverse, machine-generated medical instruction-following dataset and demonstrates that the model tuned on this dataset achieves superior performance in both medical and general domains.

For multimodal learning, although improvements have been seen in medical predictions using multimodal data, challenges in modeling irregularities within each modality and integrating irregular time information into the multimodal representation persist. We introduce strategies for addressing these challenges in multimodal electronic health records to enhance predictions for ICU patients.

Finally, we summarize the key findings and discuss future research directions to push the boundaries of deep learning in medical applications.

Contents

Curriculum Vitae **vi**

Abstract **ix**

1 Introduction **1**
- 1.1 Numerical Measurements Modeling 2
- 1.2 Natural Language Processing (NLP) 3
- 1.3 Multimodal Learning 5
- 1.4 Contributions 6

Numerical Measurements Modeling **6**

2 Domain Adaptation for Trauma Mortality Prediction in EHRs with Feature Disparity **7**
- 2.1 Introduction 7
- 2.2 Related Work 10
- 2.3 Background 11
- 2.4 Methodology 12
- 2.5 Experiments 20
- 2.6 Conclusion 26

Natural Language Processing **28**

3 Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting **29**
- 3.1 Introduction 29
- 3.2 Related Work 32
- 3.3 Methodology 34
- 3.4 Experiments 36
- 3.5 Analysis 42
- 3.6 Conclusion 48

4 **AlpaCare: Instruction Fine-tuned Large Language Models for Medical Applications** **49**
4.1 Introduction 49
4.2 Related Work 52
4.3 Method 54
4.4 Experimental Setup 56
4.5 Experiment Results 58
4.6 Human Study 64
4.7 Analysis & Case Study 65
4.8 Conclusion 69

Multimodal Learning **69**

5 **Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling** **70**
5.1 Introduction 70
5.2 Related Work 73
5.3 Method 75
5.4 Experiments 81
5.5 Conclusion 89

6 **Conclusion and Future Work** **91**
6.1 Numerical Measurement Modeling 91
6.2 NLP 92
6.3 Multimodal Learning 93

Appendix **94**

A **Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting** **95**
A.1 Data and code 95
A.2 SLM implementation and training details 95
A.3 Additional experimental results 96
A.4 Alternative measurement for privacy budget 96
A.5 General domain experimental setup 98
A.6 Prompt details 99

B **AlpaCare: Instruction Fine-tuned Large Language Models for Medical Applications** **116**
B.1 Medical Task Difficulty Level Scoring System 116
B.2 Prompt details for *MedInstruct-52k* generation 117
B.3 Training hyperparameter details 118
B.4 Additional experimental results 119

C Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling 123
C.1 Computation resource of UTDE 123
C.2 Data prepossessing 123
C.3 Baselines 124
C.4 Hyperparameters and training details 125

Bibliography 127

Chapter 1

Introduction

Recently, with the accelerated development of deep learning, a variety of deep learning techniques and frameworks have been applied to the medical domain with a range of applications, such as early-stage mortality prediction, medical exam question answering, and treatment recommendations, achieving state-of-the-art performance. To assist clinicians on various real-world medical applications, it is critical to learn medical information from diverse data sources, such as electronic health records (EHRs), pubmed articles, and transcripts of patient-doctor conversations, encompassing a variety of data modalities.

Due to the diverse biomedical data sources used to train deep neural networks for various medical applications, deep learning in biomedical applications can be categorized based on the different data modalities used for model training. In this thesis, I primarily focus on three areas: 1) Numerical Measurements Modeling, which involves training deep neural networks on biomedical numerical measurements, such as static and time-series physiological monitoring data found in Electronic Health Records (EHRs), to support doctors' decision-making; 2) Natural Language Processing (NLP), which focuses on building deep neural networks for medical textual data, including clinical notes, PubMed articles, and patient-doctor conversations, to assist in applications such as medical text generation, exam question-answering and patient diagnosis and treatment recommendation; and 3) Multimodal Learning, which utilizes data sources from more than one modality to obtain multimodal representations for downstream medical tasks, such as multimodal disease diagnosis and patient condition prediction.

In the following sections, I will dive into these three areas and summarize my works and contributions.

1.1 Numerical Measurements Modeling

In numerical measurements modeling, deep neural networks take both continuous and discrete numerical biomedical data as inputs and obtain outputs based on different applications. One of the major usages of biomedical numerical measurements modeling is medical predictions in intensive care unit (ICU) scenarios. The system aims to understand the internal relationships among various numerical features collected from ICU patients to support medical applications such as in-hospital mortality prediction, length-of-stay prediction and phenotype classification.

Training deep neural networks often requires a large labeled corpus to work well. However, limited data are available in a small, private medical system for model training, especially when further narrowed into a medical sub-domain, e.g. trauma, limiting the possible application of deep neural networks. Therefore, researchers often conduct experiments and draw conclusion on large, public EHR datasets, assuming that sufficient training data is available for different clinical tasks and the test set follows the same distribution of training data. However, direct inference models trained on large public datasets to small private datasets are challenging due to distribution shifts.

Domain adaptation (DA) is a subcategory of transfer learning that leverages knowledge from a different but related source domain as additional information to improve model performance for a target domain with limited training data. Although DA has shown progress in clinical predictions, previous studies considered overlapping data features in source and target domains and ignored important information provided by distinct features in each domain, which can hurt the prediction performance. In addition, how to align the target and source distributions when the target training set is extremely inadequate has not been well studied.

In this dissertation, I aim to fully utilize the feature information from both the source and target domains using DA techniques to enhance in-hospital mortality prediction performance in a small hospital setting using early-stage clinical numerical measurements.

1.2 Natural Language Processing (NLP)

NLP in medical domain employs deep neural networks by analyzing a wide array of biomedical textual data, including PubMed articles, doctor-patient dialogues, and clinical notes, to facilitate various medical applications including medical question-answering, free-form text generation and predictive modeling. Recently, advancements in large language models (LLMs) [1, 2, 3, 4, 5] have shown considerable promise in the medical domain, offering valuable insights and capabilities across various applications [6, 7, 8]. Despite the promise shown by LLMs in reshaping the medical domain, there emerge critical challenges, which have not been fully explored yet.

1.2.1 Privacy in medical LLM applications.

Restrictions on the use of medical data are strict due to privacy concerns, specifically prohibiting the sharing of private medical data with third parties [9], such as directly uploading it to ChatGPT [4]. While data usage policies aim to protect user data, their implementation varies across LLM APIs [4, 5, 10], leading to inconsistent levels of protection. On the other hand, small language models (SLMs) offer a way to maintain data privacy in the medical domain through local training, yet there is a significant performance gap between SLMs and LLMs.

In this dissertation, I focus on enhancing medical question-answering task performances, which require strong medical reasoning capacities, in a *privacy-restricted* scenario. I introduce a simple yet effective approach to leverage LLMs as a medical knowledge base. This involves querying LLMs for medical knowledge related to downstream tasks using medical keywords extracted from the original questions. The resulting contexts, enriched with strong medical information, are then fed into local SLMs to improve the models' medical decision-making capacity while addressing privacy concerns. Comprehensive experiments and analyses are conducted to demonstrate the effectiveness of the method.

1.2.2 Tuning data diversity in open-source medical LLMs.

Recent advancements in the training of LLMs have placed a significant emphasis on instruction-finetuning (IFT), a critical step in enabling pre-trained LLMs to effectively follow instructions [3, 11, 12]. To better align with human intent in the medical domain, [13, 14, 15] tune open-source LLMs [16] on different medical datasets to obtain medical LLMs. Although increasing in amounts, these datasets used often exhibit limited diversity, relying on benchmarks or narrow task scopes, restricting the effectiveness in medical instruction-following ability and generalizability.

In this dissertation, I propose to use stronger LLMs (e.g. GPT-4 and ChatGPT) as teachers to create a diverse, machine-generated medical IFT dataset for tuning LLMs, to better align with various user needs in medical applications. Comprehensive experiments are conducted to show that even using a smaller dataset, the model tuned with this diverse IFT dataset obtains better medical instruction-following ability and maintains strong generalizability. To foster further research, public access to the IFT dataset, *MedInstruct-52k*, a clinician-curated instruction test set, *MedInstruct-test* and our model *AlpaCare* are offered.

1.3 Multimodal Learning

Multimodal learning involves integrating information from multiple data sources across different modalities, such as text and numerical structured data, to enhance the understanding and performance of deep neural networks for various downstream tasks, including mortality prediction and diagnosis recommendations.

Health conditions among patients in ICUs are monitored via multimodal EHRs, composed of numerical time series and lengthy clinical note sequences, both taken at *irregular* time intervals. Although previous studies [17, 18, 19] show promising results on medical prediction by using multimodal EHRs comparing to only one modality, dealing with such irregularity in every modality and integrating irregularity into multimodal representations to improve medical predictions are still challenging problems.

In this dissertation, I propose to thoroughly model irregularity in multimodalities for improving medical predictions by first addressing irregularity in each single modality respectively and integrating irregularity in multimodal fusion with an interleaved attention mechanism across temporal steps. I conduct comprehensive experiments to showcase the effectiveness of the methods in tackling irregularity in each single modality and multimodal fusion scenarios, demonstrating the importance of considering irregularity in multimodal EHRs.

1.4 Contributions

Here, I briefly summarize my contributions during my Ph.D. studies. I have been working on building deep neural networks for different medical applications in numerical measurements modeling, NLP, and multimodal learning.

For numerical measurements modeling, in [20], I propose methods to improve mortality prediction in private, small hospital settings, addressing the challenge of data scarcity. For NLP, in [21], I study how to leverage medical knowledge in strong LLMs to support medical decision-making under privacy-restricted scenarios. Furthermore, in [22], I investigate the importance of data diversity in medical LLM tuning and propose a semi-automated process to generate a medical IFT dataset for better aligning medical LLM with user domain-specific intents. For multimodal learning, in [23], I explore the irregularity modeling in multimodal EHRs to improve medical predictions for ICU patients.

Chapter 2

Domain Adaptation for Trauma Mortality Prediction in EHRs with Feature Disparity

2.1 Introduction

Trauma is the leading cause of death from age 15 to 49 worldwide, resulting in the death of more than 5 million people each year [24]. After admission to intensive care units (ICUs), most of these deaths occur in the first several hours or days. Treatment decisions and actions in the first several minutes or hours after injury are critical [25], and medical decision errors are more likely to be made during this period than at later times [26]. Thus, tools that can provide efficient and real-time prediction are critical for clinicians to maximize the impact of treatment and improve survival rates.

Machine learning approaches are increasingly being used to detect adverse events in clinical settings. Different from risk scores, e.g. Modified Early Warning Score (MEWS) [27], which are developed on general populations with specific clinical measurements, machine learning techniques can be customized to different patient subpopulations or professional care facilities by training models on different datasets [28]. Recently, with the rapid development of deep learning (DL), a variety of DL techniques and frameworks have been applied to clinical predictions, such as in-hospital mortality prediction, length-of-stay prediction and phenotype classification [29] [30], achieving state-of-the-art performance. DL models often require a large corpus of labeled training data to work well [31]. Therefore, researchers often use DL approaches and draw conclusions on the basis of large public electronic health record (EHR) datasets, such as Medical Information Mart for Intensive Care (MIMIC III) [9], assuming that sufficient training data is available for different clinical tasks and the test set follows the same distribution of training data. Nevertheless, models trained on large public datasets often achieve suboptimal performance when directly deployed to smaller private EHRs due to distribution shift [32], e.g. differences in lab procedures and instrumentation, injury types and population groups based on location etc. On the other hand, only limited private EHRs are available to serve as the training set at a single medical system [28], especially when further narrowed into a medical sub-domain, e.g. trauma, limiting the possible application of DL methods.

Domain adaptation (DA) is a subcategory of transfer learning that leverages knowledge from a different but related source domain as additional information to improve model performance for a target domain with limited training data [33]. DA has made remarkable progress in computer vision [34] and natural language processing [35]. Some prior works [32][36][37] have also successfully applied DA to clinical predictions across multiple hospital systems, assuming that the distribution shift between the source and target domain of EHRs is caused by heterogeneous patient populations (*covariate shift*) and variations in data collection procedures (*systematic bias*). However, these methods are used only on datasets with overlapping features in both domains, and ignore information provided by distinct features in the target and source (*feature disparity*), which describe the characteristics of different datasets and can be essential for prediction performance. For example, clinicians tend to order particular blood tests for trauma patients, compared to other ICU patients, to identify disease states associated with coagulopathy [38], a known contributor to trauma mortality [39]. Two challenges remain to directly apply DA to multi-hospital system EHRs with feature disparity in a target data scarcity setting. First, not all clinical features included in the source and target datasets are precisely the same, and even some overlapping features collected from different hospital systems are represented in different ways. Second, it is impracticable to align the target and source distributions when the target training set is extremely inadequate.

In this chapter we aim to overcome the aforementioned challenges to fully utilize DA techniques to improve DL performance on a small target dataset, by leveraging domain-invariant knowledge from another different yet related large source dataset with feature disparity. Specifically, we introduce the private encoding technique to map target and source datasets from different feature spaces into the same hidden representation space, and utilize pairing sampling techniques [40],[41] to pair each target data point to abundant source data points, to effectively align the target and source distributions for applying different DA approaches. We demonstrate the effectiveness of DA in mortality prediction for trauma patients by taking feature disparity into account on two real-world datasets. To summarize, our contributions include:

- We extend DA techniques with a proposed private encoding to enable early-stage mortality prediction for trauma patients in multi-hospital system EHRs with feature disparity. To the best of our knowledge, this is the first work to consider feature disparity between the source and target domain in leveraging DA methods.
- Experimental results on two small target datasets show that DA techniques with our proposed methodology improve mortality prediction performance consistently

and significantly in various training scenarios, with F1-scores up to 100%.

- We provide an ablation study and a 2d t-SNE [42] analysis on two datasets, which further underscores the importance of modeling feature disparity and validates the effectiveness of DA techniques, respectively.

2.2 Related Work

Machine learning models have been trained for many clinical tasks, with several public datasets available, e.g. MIMIC III and eICU Collaborative Research Database (eICU-CRD)[43]. [44] presents benchmarking results for clinical prediction tasks such as mortality prediction, length of stay prediction, and ICD-9 code group prediction using the MIMIC III dataset. [45] performs a length of stay prediction, utilizing a Bayesian neural network, and conducts experiments on the eICU-CRD datasets. Although DL approaches attain state-of-the-art performance in medical domains, most of the works are only applied to large EHRs but difficult to yield similar performance when retrospective data are limited, which is common in real-world hospital settings.

DA leverages knowledge from a source domain to improve a target domain performance with limited training data, and has been applied in computer vision [34] and natural language processing [35]. One of the simplest DA methods is fine-tuning (FT) [46], which first pretrains model parameters on a source dataset and then updates them with a target dataset. Although widely used [47], FT tends to be sub-optimal, especially when target training data is insufficient [40], due to catastrophic forgetting [48]. More advanced techniques have been proposed to deal with the challenge in few-shot scenarios. In particular, [41] proposes a Siamese architecture [49] to address visually supervised DA by learning an embedding subspace, in which mapped raw feature domains are semantically aligned but maximally separated with few labeled target data samples available.

Under the same few-shot learning scenario, [40] provides a framework to exploit adversarial learning to identify the embedding subspace for feature alignment. However, it is not immediately apparent how to apply these methods to multi-hospital system EHRs with feature disparity because these EHRs are within a different input space.

For DA in multi-hospital system EHRs, [32] shows the effectiveness of FT on mortality prediction by pretraining on MIMIC III or eICU-CRD and fine-tuning on the other one with overlapping features. [36] seeks domain-invariant representations between two systems by adversarial learning for clinical task predictions. [37] proposes two causes of discrepancies between multi-hospital system EHRs: 1) covariate shift, caused by different patient distributions in different hospitals, and 2) systematic bias, caused by different administrative policies and workflows of different medical systems. However, they utilize only target and source datasets with overlapping features, and ignore the feature disparity among EHRs. This can significantly degrade performance in the target domain since the information provided by overlapping features between source and target is limited, and distinctive properties provided by domain-specific features are not fully utilized.

2.3 Background

Mortality Prediction Let $\mathcal{X}^{|d|}$ be a $|d|$-dimensional space, where d is a selected feature set based on particular hospital systems and/or sub-medical domains, and $\mathbf{x}$ is a data vector which represents clinical measurements, taking values in $\mathcal{X}^{|d|}$ following distribution $p(\mathbf{x})$. Supposing that $y \in \{0, 1\}$ is the binary outcome indicator for each sample, where $y = 0$ and $y = 1$ denote discharge and death respectively, we can represent an EHR dataset as a collection of N i.i.d. samples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$.

In early-stage mortality prediction, given selected clinical measurements $\mathbf{x}$ in the first several hours after admission, we would like to predict whether a patient will die after certain periods:

$$y = \Phi(\mathbf{x}), \tag{2.1}$$

where Φ is a probabilistic model. In this capter, we focus on the training data scarcity scenario for the problem of interest.

DA in EHRs To learn a representative Φ in Eq. 2.1 in a data-scarce EHR dataset, given that another relevant and large EHR dataset is accessible, we cast the problem in terms of a DA problem. In the DA setting, we have a source dataset $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$, where $\mathbf{x}_i^s \in \mathcal{X}^{|d_s|}$, and a target dataset $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{N_t}\}$, where $\mathbf{x}_i^t \in \mathcal{X}^{|d_t|}$, with $N_s \gg N_t$. Our goal is to produce an accurate survival outcome on the target domain t with training data scarcity, by leveraging knowledge from the source domain s with a sufficient amount of available data information, but with feature disparity, i.e. $|d_s \cup d_t| < |d_s| + |d_t|$, $|d_s| > |d_s \cap d_t|$ and $|d_t| > |d_s \cap d_t|$.

2.4 Methodology

In this section we introduce the private encoding technique to deal with feature disparity between multi-hospital system EHRs, and explain how it is equipped with three DA approaches: FT, ALPCA [40] and CLPSA [41] to fully utilize source knowledge and improve prediction performance in the target domain during inference.

2.4.1 Private Encoding

There is no universal representation for raw features in EHRs. We have $\mathcal{X}^{|d_s|} \neq \mathcal{X}^{|d_t|}$, given that EHRs from different hospital systems or sub-domain populations meet discrepancies. To leverage DA, which requires target and source representations in the same space, we define a new space $\tilde{\mathcal{X}}$ as the hidden feature space. Instead of obtaining the space utilizing only overlapping features between domains, private encoders, K_s and K_t, map all selected features, $\mathbf{x}^s$ and $\mathbf{x}^t$ from $\mathcal{X}^{|d_s|}$ and $\mathcal{X}^{|d_t|}$, respectively, to $\tilde{\mathcal{X}}$ to obtain the hidden feature vectors $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$, by

$$\tilde{\mathbf{x}}^s = K_s(\mathbf{x}^s) \tag{2.2}$$

$$\tilde{\mathbf{x}}^t = K_t(\mathbf{x}^t), \tag{2.3}$$

following distributions $p(\tilde{\mathbf{x}}^s)$ and $p(\tilde{\mathbf{x}}^t)$. We assume that there is a covariate shift [50] between $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$, such that the distribution $p(\tilde{\mathbf{x}}^s) \neq p(\tilde{\mathbf{x}}^t)$ in $\tilde{\mathcal{X}}$. Encoding all selected features in both domains encourages $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$ to have more comprehensive representations for corresponding domains, and can further improve the mortality prediction performance.

2.4.2 DA techniques with private encoding

DA approaches attempt to align the distributions of the source and target domains to obtain more domain-invariant information from the source, improving model performance in the target domain when training data is scarce. To utilize DA techniques in EHRs with feature disparity, we first map both source and target domain data into the same hidden space by private encoders, and then utilize two more components for both domains: shared encoders $H_s, H_t : \tilde{\mathcal{X}} \to \mathcal{Z}$, which take outputs from the source and target private encoders, respectively, to obtain shared domain-invariant representations; and classifiers, $C_s, C_t : \mathcal{Z} \to \mathcal{Y}$, which take outputs from H_s and H_t to predict mortality for the source and target domains, respectively. To improve mortality prediction in the target domain, we optimize binary cross entropy losses in both domains,

$$\hat{y} = C(H(\tilde{\mathbf{x}})), \tag{2.4}$$

$$\mathcal{L}_C = -\big[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\big], \tag{2.5}$$

with different DA techniques, where $\hat{y}$, H and C are symbolic representations of the predicted mortality, shared encoders and classifiers, respectively, for both the source and target domains. To encourage representations of the source and target towards becoming domain-invariant in the embedding space $\mathcal{Z}$, we share parameters between H_s and H_t, i.e. $H_s = H_t = H$. To ensure that representations carry the domain-specific characteristics of source and target in $\mathcal{Y}$, we set $C_s \neq C_t$.

Fine-tuning (FT) FT is one of the most direct ways to apply DA. Specifically, FT is a method that adopts a model that has already been trained for a given task, and tunes or tweaks the model to perform on a second different but related task [51]. Here, we first pretrain networks with the source data for mortality prediction,

$$\hat{y}^s = C_s(H_s(\tilde{\mathbf{x}}^s)), \tag{2.6}$$

and then fine-tune with the available target data,

$$\hat{y}^t = C_t(H_t(\tilde{\mathbf{x}}^t)), \tag{2.7}$$

where H_t is initialized by H_s from Eq. 2.6.

Adversarial Learning with Pairing Classes Alignment (ALPCA) Traditionally, adversarial learning [52] introduces a min-max game training strategy to obtain domain-invariant knowledge by seeking a discriminator, D, that can identify samples from source and target distributions. After D is learned, H, in the role of a generator, is updated to render D unable to distinguish samples from the source and target domains. However,

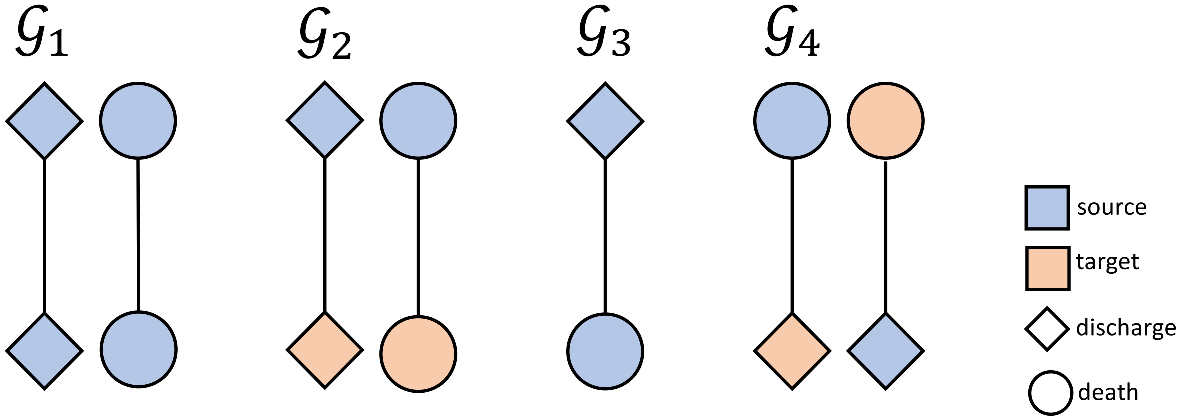


Figure 2.1: An example illustration of pairs in four groups for ALPCA.

due to data scarcity, H cannot estimate the target population accurately. Even with a perfect H, i.e. D cannot distinguish whether a sample is from the source or target domain, H still cannot guarantee that samples from different domains but with the same class label will map nearby in the embedding space, since no class information is provided to D in standard adversarial training [40].

We follow [40] to alleviate the target training sample shortage issue in adversarial domain adaptation and encourage networks to learn the properties of death and discharge patients in two domains. Specifically, a pairing strategy [40] is used to overcome the scarcity of training data by creating four groups based on domain and class information, and a multi-class discriminator D [40] is introduced by distinguishing four pairing groups to encourage H to generate domain-invariant representations with class information.

As shown in Figure 2.1 (encouraged by [40]), two groups ($\mathcal{G}_1$ and $\mathcal{G}_2$) consist of positive pairs and two ($\mathcal{G}_3$ and $\mathcal{G}_4$) consist of negative pairs. Each positive pair is composed of two samples with the same class, either `death-death` or `discharge-discharge`; while each negative pair is composed of two samples with different classes, i.e. `death-discharge`. Pairs in $\mathcal{G}_1$ and $\mathcal{G}_3$ consist of items both from the source domain, generated by randomly

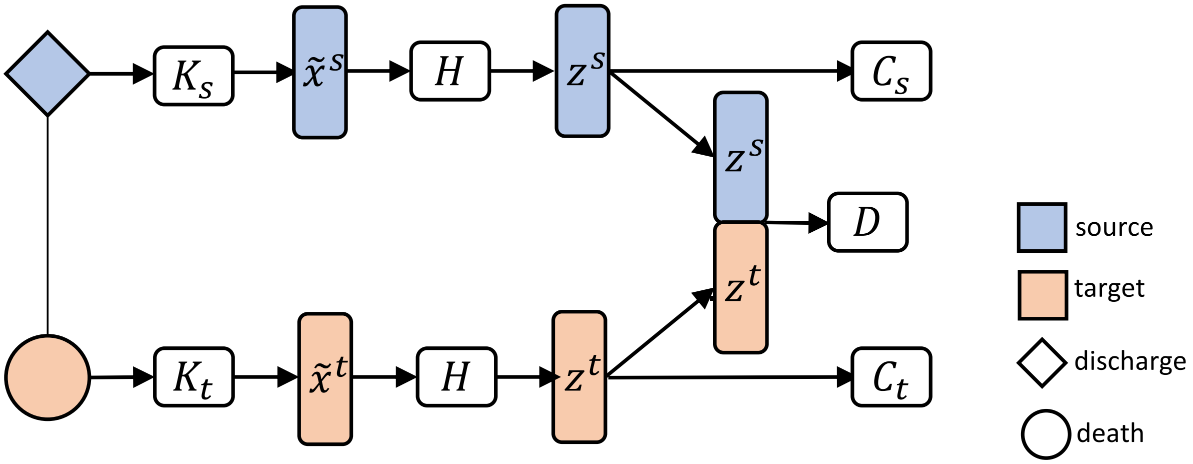


Figure 2.2: ALPCA with a `discharge-death` example from $\mathcal{G}_4$.

pairing samples drawn from the source distribution based on class information; while pairs in $\mathcal{G}_2$ and $\mathcal{G}_4$ consist of one item from the source and another from the target distribution, created by pairing each target sample with a number of randomly drawn source samples based on classes. In $\mathcal{G}_1$ and $\mathcal{G}_2$, we generate `death-death` and `discharge-discharge` with approximate ratio 1 : 1 to encourage the networks to learn more similarity information between death patients, which is hard to achieve with very imbalanced medical data. We construct each group of the same size for training D by matching the other three groups' size with the smallest one.

As demonstrated in Figure 2.2, H tries to fool D by taking hidden representations from corresponding private encoders (K_s and K_t) and outputting domain-invariant feature representations. D takes the concatenation of domain-invariant representations to distinguish which group a sample pair comes from, trained via a standard cross entropy loss,

$$\mathcal{L}_D = -E\big[\sum_{i=1}^{4} y_{\mathcal{G}_i} \log(D(H(K(\mathcal{G}_i))))\big], \tag{2.8}$$

where $y_{\mathcal{G}_i}$ denotes the label of group $i = 1, 2, 3, 4$, and K denotes a symbolic representation of private encoders based on different pair groups, i.e. K is two K_s for $\mathcal{G}_1$ and $\mathcal{G}_3$, while

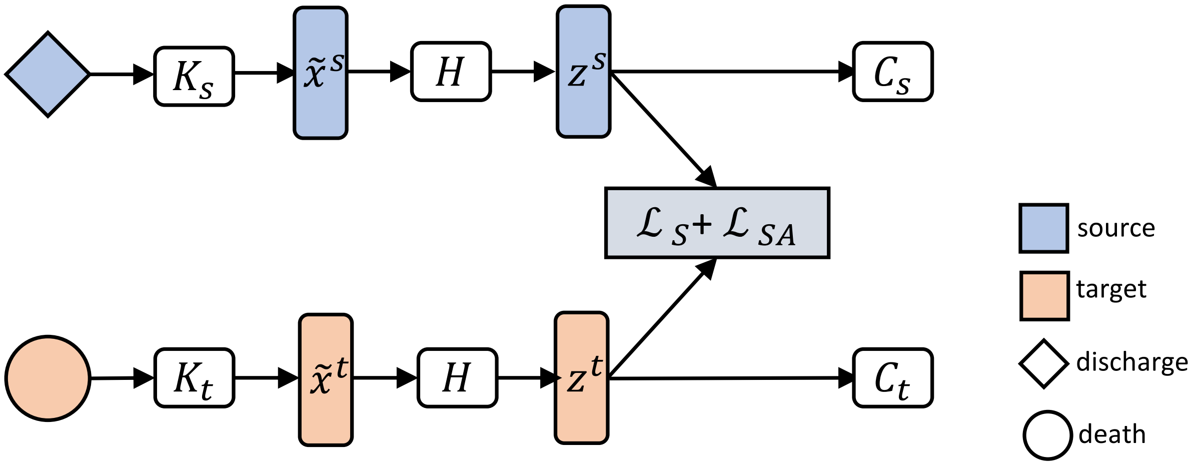


Figure 2.3: CLPSA with a negative example.

one K_t and one K_s for $\mathcal{G}_2$ and $\mathcal{G}_4$. To output domain-invariant representations carrying class information, H fools D to identify sample pairs from $\mathcal{G}_2$ as $\mathcal{G}_1$, and pairs from $\mathcal{G}_4$ as $\mathcal{G}_3$, so that target samples can have indistinguishable representations with more diverse source samples. Mathematically, H along with K_t and K_s are updated, with

$$\mathcal{L}_G = -E\big[y_{\mathcal{G}_1}\log(D(H(K(\mathcal{G}_2)))) + y_{\mathcal{G}_3}\log(D(H(K(\mathcal{G}_4))))\big]. \tag{2.9}$$

Minimizing (2.9) together with the source and target classification losses,

$$\mathcal{L}_{G\&C} = \alpha\mathcal{L}_G + \beta\mathcal{L}_{C_s} + \mathcal{L}_{C_t}, \tag{2.10}$$

where α and β are hyper-parameters, encourages networks to obtain domain-invariant representations with class information and achieve good performance on classification tasks by considering distribution differences in $\mathcal{Y}$, simultaneously.

Contrastive Learning with Pairing Semantic Alignment (CLPSA) Instead of training an additional D for aligning the feature distributions in $\mathcal{Z}$ to achieve the goal of

DA, CLPSA [41] learns $\mathcal{Z}$ by pulling samples from different domains but the same class as close together as possible yet pushing samples from different domains and classes as far away as possible. Following [41], as shown in Figure 2.3, H takes $\tilde{\mathbf{x}}^s$ and $\tilde{\mathbf{x}}^t$ to obtain domain-invariant representations by training with contrastive semantic alignment losses, consisting of a semantic alignment loss $\mathcal{L}_{SA}$ and a separation loss $\mathcal{L}_S$.

Specifically, to align the distributions of samples in the embedding space, a semantic alignment loss is introduced,

$$\mathcal{L}_{SA} = \sum_{a=0}^{1} d(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_a^t))), \tag{2.11}$$

where $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_a^t$ are vectors from the source and target domains in $\tilde{\mathcal{X}}$ belonging to the same class a. d is a metric to measure the distance between the distributions of $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_a^t$ in $\mathcal{Z}$. $\mathcal{L}_{SA}$ prompts samples with the same class from two different domains to map nearby in the embedding space.

Although pulling the same class samples from different domains close together in the embedding space encourages target groups to obtain information from similar points in the source, minimizing $\mathcal{L}_{SA}$ does not guarantee that points in different classes from two domains are separated enough, which would significantly degrade performance in the target domain. Therefore, we leverage a separation loss,

$$\mathcal{L}_S = \sum_{a=0,b=0|a\neq b}^{1} k(p(H(\tilde{\mathbf{x}}_a^s)), p(H(\tilde{\mathbf{x}}_b^t))) \tag{2.12}$$

where k is a metric to measure the similarity between the distributions $\tilde{\mathbf{x}}_a^s$ and $\tilde{\mathbf{x}}_b^t$ in $\mathcal{Z}$. $\mathcal{L}_S$ encourages class separation by pushing the representations of different classes in two domains farther away, i.e. adding a penalty if the distance between distributions $p(H(\tilde{\mathbf{x}}_a^s))$ and $p(H(\tilde{\mathbf{x}}_b^t))$ is small.

Finally, CLPSA is jointly trained with classification losses from both domains and contrastive semantic alignment losses,

$$\mathcal{L}_{CCSA} = \mathcal{L}_{C_t} + \gamma \mathcal{L}_{C_s} + \delta(\mathcal{L}_{SA} + \mathcal{L}_S), \tag{2.13}$$

where γ and δ are hyper-parameters.

Similar to ALPCA, it is difficult to minimize Eq. 2.13 when the target training data is scarce because $\mathcal{L}_{SA}$ and $\mathcal{L}_S$ depend on calculating distances and similarities between distributions, and those need to learn with sufficient samples.

Therefore we pair each target sample to a large number of randomly selected source samples and compute average pairwise distances between positive pairs

$$d(p(H(\tilde{\mathbf{x}}^s_a)), p(H(\tilde{\mathbf{x}}^t_a))) = \sum_{i,j} d(H(\tilde{x}^s_i), H(\tilde{x}^t_j)), \tag{2.14}$$

where $y^s_i = y^t_j = a$; or similarities between negative pairs

$$k(p(H(\tilde{\mathbf{x}}^s_a)), p(H(\tilde{\mathbf{x}}^t_b))) = \sum_{i,j} k(H(\tilde{x}^s_i), H(\tilde{x}^t_j)), \tag{2.15}$$

where $y^s_i = a \neq y^t_j = b$, between points in the embedding space to achieve semantic alignment. Here, the ratio between positive and negative sample pairs is 1:1, and the ratio between the `death-death` and `discharge-discharge` pairs is also 1:1 in the positive group, to encourage the network to learn more information from the records of patients that died, in an imbalanced dataset.

We implement $\mathcal{L}_{SA}$ and $\mathcal{L}_S$ with contrastive loss following [53]

$$d(H(\tilde{x}^s_i), H(\tilde{x}^t_j)) = \frac{1}{2}\|H(\tilde{x}^s_i) - H(\tilde{x}^t_j)\|^2, \tag{2.16}$$

$$k(H(\tilde{x}^s_i), H(\tilde{x}^t_j)) = \frac{1}{2} max(0, m - \|H(\tilde{x}^s_i) - H(\tilde{x}^t_j)\|)^2, \tag{2.17}$$

where $\|\cdot\|$ denotes the Frobenius norm and m is a margin to define the separability in the embedding space.

2.5 Experiments

2.5.1 Datasets

We conducted experiments on a source dataset extracted from MIMIC III and two target datasets: the UCSF dataset and the EICU dataset. For the missing data issue [54], each dataset was preprocessed in the same way. First, we excluded patients with more than 50% of features missing, and then we applied MICE [55] data imputation for the remaining missing values. The statistics of these processed datasets are summarized in Table 2.1, and details are described below.

Table 2.1: Dataset Statistics

	Source	**Target**	
	MIMIC III	UCSF	EICU
# patients	29914	2069	2565
# death	3063	342	204
# discharge	26851	1727	2361
# feature	21	31	35
#overlapping	21	18	12

MIMIC III is a public data source of de-identified EHRs, which contains 53,423 patients admitted to ICUs at a Boston-area hospital from 2001–2012 [9]. We extracted the data following [56] and selected the 17 most common clinical features (e.g. heart rate, blood pressure, temperature and respiratory rate, etc.),

as well as 4 demographic features (ethnicity, gender, age and weight) for the mortality prediction. As we focus on prediction with early-stage measurements, we took the first appearance of each clinical feature measurement in the first two hours after admission if it is available; otherwise, it was regarded as a missing value. After data preprocessing, the source dataset consisted of 29914 patients with 21 features in total. We randomly selected 80% of the data points as the training set, and the rest as the validation set.

UCSF Dataset, collected from the UCSF/San Francisco General Hospital and Trauma Center, contains 2,190 patients admitted to a Level I trauma center. We selected demographic information, injury score, physical vital signs and laboratory results[1] measured at the time of admission or during the first two hours after admission, as features for mortality prediction. After data preprocessing, we have 2,069 patients with 31 features, including 18 features overlapping with MIMIC III. We randomly selected 784 and 785 patients as the validation and testing sets, respectively, and further randomly drew different training sizes from the remaining patients to simulated data scarcity at various levels.

EICU Dataset is extracted from the eICU-CRD, a multi-center ICU database with high granularity data for over 200,000 admissions to ICUs monitored by eICU programs across the United States. We selected the first-time ICU admission of each adult patient (age $>$ 18) with a diagnosis related to trauma. Then we queried the minimum and maximum of clinical measurements (e.g. blood urea nitrogen, white blood cell count and hemoglobin, etc.) and demographics information taken in the first 24 hours of a patient's ICU stay, following the code shared by the eICU research community[2], to predict the mortality after the first admission day. After data prepossessing, we have 2565 trauma patients with a death and discharge ratio of approximately 1:12, and 35 features containing 12 overlaps with MIMIC III. 1032 and 1033 patients were randomly selected as validation and test sets, and the remaining were randomly drawn for various training size scenarios, which is the same as UCSF dataset.

---

[1] Lab tests focusing on trauma patients: Protein C, D-Dimer, ATIII, Factor II, V, VII, VIII, IX and X

[2] `https://github.com/mit-lcp/eicu-code`

2.5.2 Experiment settings

Evaluation Metric

We measured the performance of three DA methods and baselines by the F1-score, with 0.5 as the prediction threshold following the previous work [57].

Model configurations

- **MLP-target (baseline)**: This is a multi-layer perceptron (MLP) composed of K_t, H and C_t, trained only on the target dataset.
- **FT**: An MLP model composed of K_s, H and C_s was first pretrained on source data. The best performing model, evaluated by the source validation set, was saved. Then another MLP model composed of K_t, H and C_t were trained on target data, with K_t and C_t randomly initialized and H inherited from the pretraining step.
- **ALPCA**: As shown in Figure 2.2, all six networks (K_s, K_t, H, D, C_s and C_t) in the framework are MLPs with random initialization, and the discriminator is trained with Eq. 2.8. The other five networks are trained with Eq. 2.10 following standard adversarial training schema [52].
- **CLPSA**: As shown in Figure 2.3, five networks (K_s, K_t, H, C_s and C_t) in the framework are MLPs with random initialization. They are trained with Eq. 2.13.

For the neural networks above, we used batch normalization to normalize the input layer by re-centering and re-scaling. For fair comparison, we assigned the same structures for K_t, H and C_t in the MLP-target and FT, respectively. In ALPCA and CLPSA, K_s, K_t, H, C_s and C_t have the same structure as FT. D in ALPCA is a two-layer MLP.

The size of each hidden layer in all networks was selected by grid search among $\{8, 16, 32\}$. We implemented all models in PyTorch [58] and all of the neural networks

Table 2.2: F1 comparison (%) of MLP-target, FT, ALPCA and CLPSA with different training data sizes on the UCSF and EICU datasets. Mean values along with their standard deviations in the subscript were calculated with 5 data splits.

	Training size	50	100	200	300	400	500
Dataset	**Model**						
UCSF	MLP-target	$36.2_{4.7}$	$36.4_{0.4}$	$39.5_{6.1}$	$46.8_{4.8}$	$46.7_{2.0}$	$47.4_{0.3}$
	FT	$50.3_{7.1}$	$51.4_{7.7}$	$58.1_{4.6}$	$59.9_{4.4}$	$60.1_{2.8}$	$58.0_{2.6}$
	ALPCA	$48.6_{7.3}$	$\mathbf{57.5}_{2.1}$	$\mathbf{59.4}_{1.4}$	$60.4_{1.2}$	$59.9_{0.9}$	$60.7_{1.7}$
	CLPSA	$\mathbf{52.3}_{3.4}$	$54.2_{4.3}$	$59.3_{2.6}$	$\mathbf{62.3}_{2.3}$	$\mathbf{64.5}_{1.5}$	$\mathbf{63.7}_{2.2}$
EICU	MLP-target	$17.7_{7.7}$	$20.6_{2.7}$	$21.4_{1.3}$	$21.9_{1.9}$	$20.4_{1.7}$	$22.3_{1.4}$
	FT	$23.5_{7.0}$	$21.8_{5.1}$	$26.3_{3.7}$	$26.7_{5.6}$	$25.7_{4.5}$	$27.9_{1.7}$
	ALPCA	$26.1_{5.4}$	$25.3_{6.5}$	$30.7_{3.2}$	$29.8_{3.2}$	$35.6_{4.5}$	$36.3_{1.9}$
	CLPSA	$\mathbf{28.1}_{6.4}$	$\mathbf{30.4}_{5.1}$	$\mathbf{34.0}_{4.5}$	$\mathbf{37.5}_{2.9}$	$\mathbf{40.8}_{3.5}$	$\mathbf{41.0}_{2.1}$

were trained with Adam [59], whose learning rates were selected by grid search among $\{0.0001, 0.0002, 0.0005\}$[3].

2.5.3 Performance comparison

Our results are summarized in Table 2.2. All three DA methods yielded better performance than the MLP-target in both datasets across various training data scenarios, demonstrating the effectiveness of DA in small training data regimes. Both ALPCA and CLPSA achieved better performance than FT. CLPSA outperformed ALPCA or achieved comparable results in the UCSF dataset, and consistently performed better than ALPCA in the EICU dataset, across the entire range of training data sizes. The result

---

[3] All other hyper-parameters, e.g. α, β and m etc. were selected by grid search in the same ranges for fair comparison. We omitted these due to space limitations.

Table 2.3: F1 comparison (%) of ablation study on three reasons for discrepancies of DA methods on the UCSF and EICU datasets.

Model	MLP-target	FT			ALPCA			CLPSA		
		`C-only`	`S+C`	`F+S+C`	`C-only`	`S+C`	`F+S+C`	`C-only`	`S+C`	`F+S+C`
UCSF	$47.4_{0.3}$	$29.0_{2.2}$	$38.9_{3.5}$	$\mathbf{58.0}_{2.6}$	$54.6_{4.1}$	$55.6_{1.7}$	$\mathbf{60.7}_{1.7}$	$54.8_{2.1}$	$53.7_{2.6}$	$\mathbf{63.7}_{2.2}$
EICU	$22.3_{1.4}$	$18.3_{0.7}$	$18.3_{0.7}$	$\mathbf{27.9}_{1.7}$	$17.34_{0.8}$	$15.5_{3.11}$	$\mathbf{36.3}_{1.9}$	$16.5_{2.4}$	$15.8_{0.7}$	$\mathbf{41.0}_{2.1}$

that ALPCA underperforms CLPSA may be primarily due to introducing an additional network (discriminator), making the whole framework more challenging to optimize, with more parameters to update [60]. Surprisingly, the three DA methods with 50 training data points consistently outperformed the MLP-target with 500 training data points (10 times larger), which further strengthens the powerful capability of DA in small EHR data regimes.

2.5.4 Ablation study on modeling feature disparity.

We have demonstrated the effectiveness of DA methods in Table 2.2, where feature disparity (`F`) is modeled by private encoding strategy, as well as systematic bias (`S`) and covariate shift (`C`). We denote this setting as `F+S+C`. To verify the effectiveness of modeling feature disparity in DA, we considered two variants for both EHRs in the 500 patients training data scenario. First, we considered both systematic bias and covariate shift but not feature disparity, denoting as `S+C`. The networks were trained on both the source and target data, only including overlapping features and not sharing parameters on private encoders. Second, we trained the networks on two domains with overlapping features and shared parameters on private encoders, counting only the discrepancies caused by covariate shift, which is denoted as `C-only`.

Table 2.3 presents the results of our ablation study on three different causes of discrepancies, as well as the MLP-target. The empirical analysis shows that `S+C` is not always better than `C-only`, or vice versa and they may even underperform MLP-target, which utilizes all features in target domain including overlapping ones. However, our method, `F+S+C`, consistently yields significantly better performance than `S+C` and `C-only` , demonstrating the importance of modelling feature disparity in DA of multi-hospital systems EHRs, especially when the overlapping features between systems are limited.

2.5.5 Analysis

To understand why DA improves prediction performance on target datasets and why CLPSA yields the best performance compared to other DA strategies, we further analyzed 2-dimensional (2-d) embeddings of testing sets on the UCSF and EICU datasets, respectively. Specifically, we obtain 2-d embeddings by reducing high dimensional representations before feeding into the C_t using t-SNE [61] in Scikit-learn [62]. Then we calculate the average difference between inter-cluster and intra-cluster distance in the death group, which is more important than the discharge group in the medical domain, yielding

$$d_{\texttt{diff}} = d_{\texttt{inter}} - d_{\texttt{intra}} = \frac{\sum_i^n \|e_i - c_{\texttt{discharge}}\|^2 - \sum_i^n \|e_i - c_{\texttt{death}}\|^2}{n},$$

where n represents the number of patients who died, e_i is i-th patient's 2-d embedding, and $c_{\texttt{discharge}}$ and $c_{\texttt{death}}$ are centers of the discharge and death clusters in the 2-d embedding space, respectively. Inter-cluster distance is an average distance between members of a cluster and another cluster's center; meanwhile, intra-cluster distance is an average distance between members and their own center. We want the inter-cluster distance to be large to push the cluster far away from the other, but the intra-cluster distance to be small to pull members in a cluster as close as possible. Thus the $d_{\texttt{diff}}$ should be large to make the cluster easy to identify by C_t.

Average $d_{\texttt{diff}}$ results corresponding to various training data sizes in Table 2.2 are summarized in Table 2.4. Consistent with the F1 scores in Table 2.2, CLPSA achieved a greater or comparable $d_{\texttt{diff}}$, compared to other methods for both the UCSF and EICU datasets, which indicates that hidden representations generated by CLPSA are prone to be identified by C_t compared to other models.

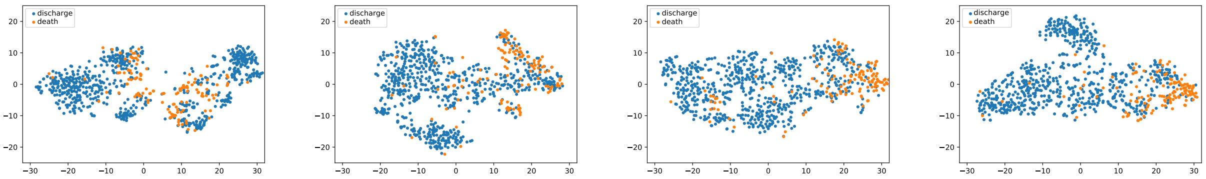
We further visualized the 2-d t-SNE of all models from the same data split on both datasets with 500 training points in Figure 2.4. The 2d t-SNEs of death and discharge patients on both the UCSF and the EICU dataset for MLP-target in Figure(a) and Figure(a′) are almost overlapping, indicating that it is difficult to find a straight line to distinguish the two groups, making C_t prediction of mortality challenge. For the other DA models, the cluster of the death group often aggregates at the right in each plot, making it more straightforward to separate with the discharge patients compared to the MLP-target, and illustrating the reason for improvement by utilizing DA. Comparing the 2-d t-SNEs of different DA methods, we find that the cluster of death patients with CLPSA in Figure(d) and Figure(d′) is more concentrated than that with FT in Figure(b) and Figure(b′) and with ALPCA in Figure(c) and Figure(c′), which explains the better performance of CLPSA compared to other DA strategies.

2.6 Conclusion

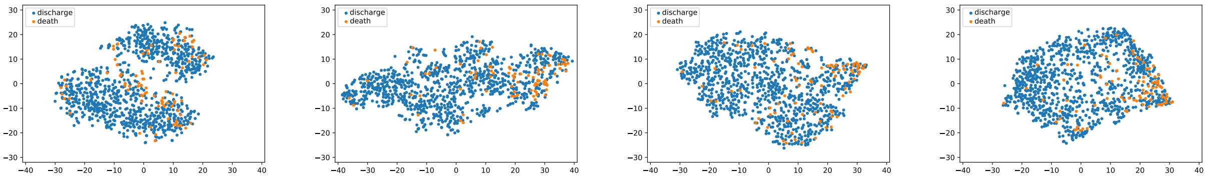
In this chapter we showed how DA methodologies, in particular FT, ALPCA and CLPSA, can be used to improve the performance of mortality prediction for trauma patients in regimes with limited training data. In contrast to existing DA methodologies in multi-hospital system EHR predictive tasks, which consider only the discrepancies caused by covariate shift and systematic bias, we bridge the gap of feature disparity by introducing a private encoding strategy that maps clinical measurements from different raw

Table 2.4: 2d t-SNE embedding distance evaluation of testing set in the UCSF and EICU datasets.

	Training Size	50	100	200	300	400	500
Dataset	**Model**						
UCSF	MLP-target	5.7	5.9	5.1	8.2	9.3	8.3
	FT	6.9	10.8	11.8	13.0	13.9	12.9
	ALPCA	**11.3**	11.7	**13.4**	11.1	13.8	11.9
	CLPSA	8.2	**12.1**	12.9	**14.5**	**15.8**	**14.0**
EICU	MLP -target	−0.2	0.8	2.3	1.7	2.7	2.9
	FT	3.7	3.6	4.7	3.1	3.6	6.1
	ALPCA	**4.2**	4.6	**4.9**	3.1	7.1	3.3
	CLPSA	3.6	**5.5**	4.1	**5.0**	**7.2**	**6.5**



(a) UCSF: MLP (b) UCSF: FT (c) UCSF: ALPCA (d) UCSF: CLPSA



(a′) EICU: MLP (b′) EICU: FT (c′) EICU: ALPCA (d′) EICU: CLPSA

Figure 2.4: Testing set 2d t-SNE embedding for the UCSF and EICU dataset with the 500 training data scenario. The blue and orange dots represent the discharged and dead patients, respectively.

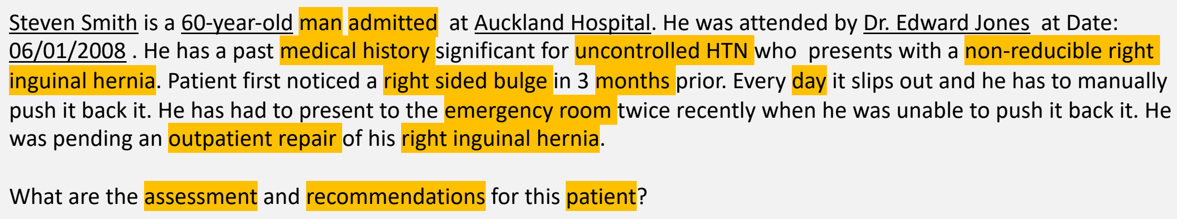
feature spaces to a hidden feature space and follows with various DA techniques. Extensive experimental results on two datasets demonstrate the usefulness of DA, and ablation studies and 2-d t-SNE analysis further explain the effectiveness of private encoding and DA methods, respectively.

Chapter 3

Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting

3.1 Introduction

Figure 3.1: **Synthetic medical data for illustration.** Though rich in domain-specific knowledge, medical data contains sensitive private information. We extract keywords to mitigate privacy concerns.



Steven Smith is a 60-year-old man admitted at Auckland Hospital. He was attended by Dr. Edward Jones at Date: 06/01/2008 . He has a past medical history significant for uncontrolled HTN who presents with a non-reducible right inguinal hernia. Patient first noticed a right sided bulge in 3 months prior. Every day it slips out and he has to manually push it back it. He has had to present to the emergency room twice recently when he was unable to push it back it. He was pending an outpatient repair of his right inguinal hernia.

What are the assessment and recommendations for this patient?

Large language models (LLMs) [1, 2, 3, 4, 5] have shown promise in the medical field [6, 7, 8]. However, concerns about medical data privacy prevent the direct use of LLMs' medical capabilities in healthcare domain, as illustrated in Figure 3.1. Despite data usage policies[1] to safeguard user data, the implementation of these policies varies among LLMs, creating inconsistent protection levels. Moreover, medical data usage agreements[2] are stringent, explicitly forbidding data sharing with third parties, like direct uploads to ChatGPT [4]. Therefore, developing methods to harness LLMs' medical knowledge while balancing data privacy in *privacy-restricted* scenarios is an urgent and underexplored research area.

Small language models (SLMs) [3] that are specific to the medical domain [64, 65, 66, 67, 68, 69] have shown superior in-domain task performance compared to general-domain SLMs [70, 71, 72], addressing the vital need for data privacy in the medical field through local training. However, a notable performance gap between SLMs and LLMs in medical tasks remains [6, 8]. A critical question arises: How can we bridge the performance gap between SLMs and LLMs for medical tasks in privacy-restricted scenarios?

One common strategy to narrow the performance gap between LLMs and SLMs involves leveraging generated rationales [73, 74] from LLMs to boost the performance of SLMs [63, 75, 76, 77]. However, previous research has often required feeding complete data information into LLMs, ignoring privacy concerns. Thus, it is essential to explore alternative methods that can effectively utilize LLM-generated context, which is rich in medical knowledge, while balancing privacy concerns for LLMs and performance enhancement for SLMs in the medical domain.

In this chapter, we present a simple yet effective pipeline that boosts SLM performance by incorporating medical contexts from LLMs in privacy-restricted scenarios. To the best of our knowledge, this is the first work that has utilized LLMs to improve SLM performance in such settings. While our primary focus is on multiple-choice medical

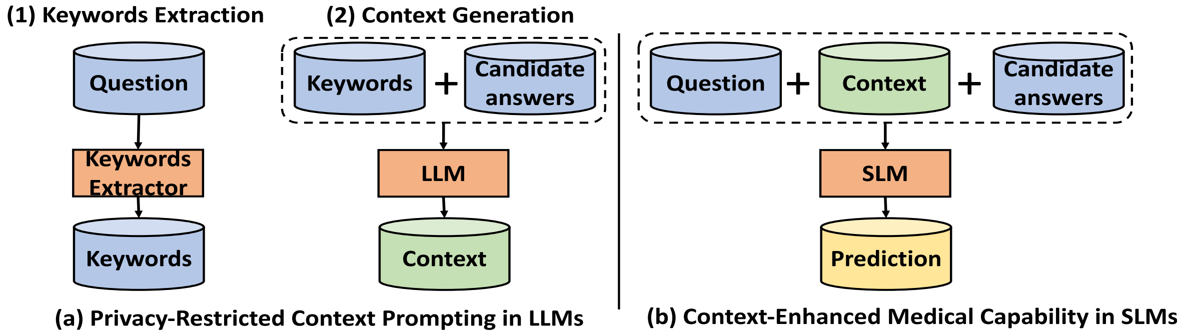
---

[1] https://openai.com/policies/api-data-usage-policies

[2] https://physionet.org/content/mimiciii/view-dua/1.4/

[3] Following [63], we argue that the definition of small and large models is context-dependent.

Figure 3.2: **Framework overview.** (a) To mitigate privacy leakage, we use a keyword extractor to obtain medical keywords. Clinicians then create several contexts based on these keywords and candidate answers, which the LLM uses to produce privacy-restricted contexts. (b) The generated contexts are used as additional input to enhance SLM medical decision-making capacity.



(a) Privacy-Restricted Context Prompting in LLMs (b) Context-Enhanced Medical Capability in SLMs

QA, our framework can be adapted to other tasks or domains. Figure 4.2 illustrates our framework. Specifically, we use existing named-entity recognition (NER) models [78] to extract keywords[4], thereby mitigating privacy risks. Based on these keywords and candidate answers, clinicians generate several medical contexts that mirror their thought processes. These clinician-written contexts serve as demonetisation to generate contexts for the remaining data by the LLM by in-context learning [1]. Finally, we integrate these contexts into the SLMs to improve their performance on medical tasks.

Overall, our main contributions can be summarized as follows:

1. We propose a simple yet effective pipeline that uses keywords and candidate answers to elicit medical knowledge from LLMs, which is then fed into SLMs to enhance medical capabilities.
2. We introduce a privacy-conscious prompting strategy that mimics clinicians' thinking

---

[4] Keywords can be extracted by other methods, e.g., a manually created dictionary based on domain expertise.

to generate medical knowledge-intensive contexts from LLMs in privacy-restricted scenarios.

3. Our method notably surpasses standard SLM fine-tuning without context in both full training and few-shot settings, achieving up to a 22.57% increase in accuracy, and obtains SOTA results in two medical tasks in privacy-restricted scenarios.

3.2 Related Work

LLMs in Biomedicine. LLMs excel in NLP tasks, including medical fields [1, 2, 3, 4, 5, 8, 79, 6]. [79] used ChatGPT for the US Medical Exam. [8] leverage GPT-3.5 models for medical reasoning tasks. MedPaLM, tuned from FlanPaLM [80], answered consumer medical questions comparably to clinicians [6]. However, privacy concerns limit LLMs in real-world medical uses, highlighting the need to utilize LLM medical knowledge in privacy-restricted setting.

Biomedical SLMs. Domain-specific SLMs, either extended from general-domain pretraining [67, 66, 81] or built from scratch on biomedical corpora [64, 65, 68], surpass general models [70, 71, 72] in biomedical tasks [64, 65, 66, 67, 68, 69]. They also enhance privacy through local training. However, a performance gap remains between domain-specific SLMs and LLMs in medical tasks [8, 6], emphasizing the need for strategies to reduce this gap.

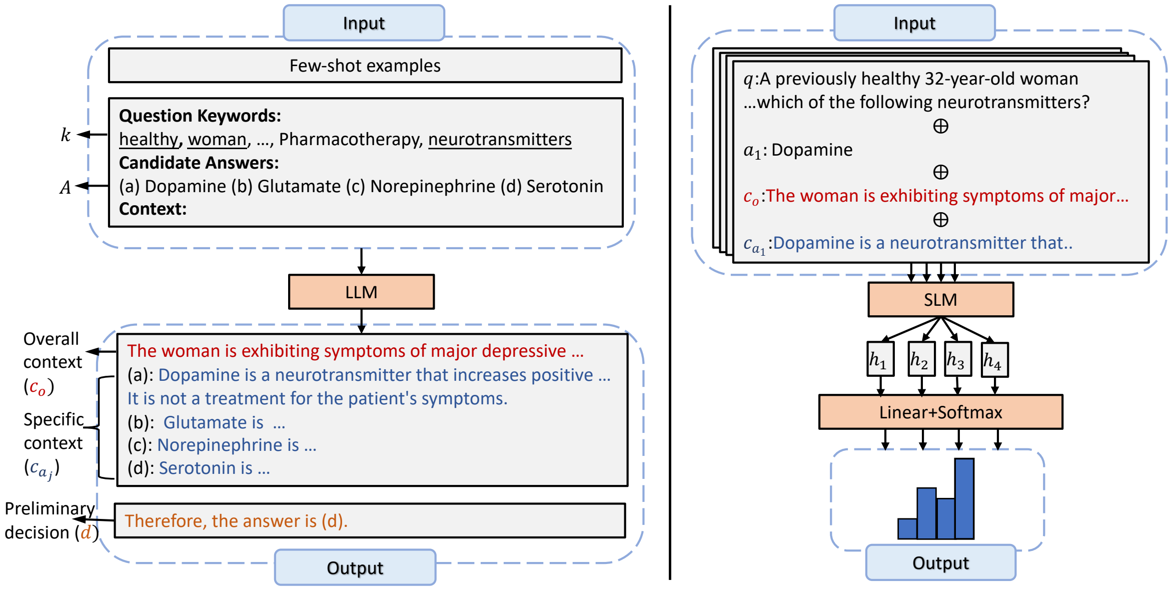
Knowledge Distillation from LLMs. Previous studies have investigated distilling knowledge [82] from LLMs to enhance smaller models' performance [76, 77, 83, 63]. [76] fine-tuned smaller models using InstructGPT-generated reasoning samples. [83] input LLM generated rationales into SLM for question-answering. [63] applied LLM-derived explanations for multi-task learning. However, these approaches involve prompting LLMs without considering privacy preservation, which is vital in real-world medical scenarios

[84, 85, 86]. We introduce a keyword-based prompting strategy to generate medical knowledge from LLMs, while balancing LLM privacy preservation and SLM performance enhancement.

Augmenting NLP Tasks with External Knowledge. Improving knowledge-intensive NLP tasks can be achieved by retrieving information from large evidence corpora like Wikipedia [87, 88, 89, 90, 91, 92, 93]. The retrieve-then-read model [94] utilizes retrievers, such as BM25 [95] and DPR [96], to identify relevant documents within a corpus. Then, a reader, like FiD [88], analyzes these documents to enhance NLP tasks [93, 97, 98, 99]. Some studies retrieve subgraphs from knowledge graphs to boost question-answering tasks [89, 100]. Recent research shows that pre-trained language models can ”retrieve” information via direct text generation [101, 102]. [103, 87] utilized LLMs to generate relevant contexts or background documents for question-answering tasks. Our work leverages LLMs as a knowledge base for medical knowledge retrieval.

Data Privacy in BioNLP. Biomedical data inherently contains sensitive information [104]. De-identification methods eliminate private details and replace them with synthetic data [105, 106, 107]. For example, [108, 109] treat de-identification as a NER problem, modeling by neural networks. Despite being de-identified, sharing restrictions still apply [9, 110]. Differential privacy (DP) provides theoretical bounds on individual data privacy while allowing aggregate statistical disclosure for the entire database [111, 112, 113]. [114] train a model using a DP-based approach to generate synthetic clinical notes for secure sharing. In this work, we migrate privacy concerns during LLM inference by prompting the LLM with medical keywords extracted from raw data.

Figure 3.3: **LLM generates privacy-restricted medical contexts to enhance SLM decision-making.** (a) The LLM generates medical knowledge-intensive context for each instance using clinicians' few-shot demonstrations, extracted keywords from raw data (k), and candidate answers (A). The generation output comprises: overall context (c_o), specific context of each candidate answer (c_{a_j}); and preliminary decision of LLM (d). (b) The overall and specific contexts are then concatenated ($\oplus$) with the question as additional input to fine-tune an SLM, enhancing its medical decision-making.


Input
Few-shot examples
Question Keywords:
healthy, woman, ..., Pharmacotherapy, neurotransmitters
Candidate Answers:
(a) Dopamine (b) Glutamate (c) Norepinephrine (d) Serotonin
Context:
k
A
LLM
Overall context (c_o)
Specific context (c_{a_j})
Preliminary decision (d)
The woman is exhibiting symptoms of major depressive ...
(a): Dopamine is a neurotransmitter that increases positive ... It is not a treatment for the patient's symptoms.
(b): Glutamate is ...
(c): Norepinephrine is ...
(d): Serotonin is ...
Therefore, the answer is (d).
Output
Input
q:A previously healthy 32-year-old woman ...which of the following neurotransmitters?
⊕
a_1: Dopamine
⊕
c_o:The woman is exhibiting symptoms of major...
⊕
c_{a_1}:Dopamine is a neurotransmitter that..
SLM
h_1
h_2
h_3
h_4
Linear+Softmax
Output

(a) (b)

3.3 Methodology

3.3.1 Privacy-Restricted Context Prompting in LLMs

We consider a dataset $D = \{(q_i, A_i, y_i)\}^N$, where N denotes the total number of instances, q_i a problem, A_i a set of candidate answers, and y_i the correct answer. A set of keywords k_i is extracted from each problem q_i using a medical NER model [78]. To mitigate privacy leakage, we feed k_i and A_i into the LLM to generate medical context

[76, 63].

We have a small set of clinician-written instances $E = \{(k_i^p, A_i^p, C_i^p, d_i^p)\}^M$, where both C_i^p, denoting medical contexts, and d_i^p, representing preliminary decisions, are generated based on partial data information k_i^p and A_i^p, with $M \ll N$[5]. The medical context C_i^p consists of an overall context and specific contexts for each candidate's answer. The overall context encapsulates high-level medical knowledge based on keywords and candidate answers, while specific contexts provide detailed information for each candidate answer and its relationship to the overall context. The preliminary decision represents clinicians' predictions informed by these contexts. This prompt strategy simulates clinicians' reasoning steps by producing high-level guidance based on partial data, examining individual candidate answers in-depth, and ultimately making an initial determination given prior analysis, as shown in Figure 3.3 (a).[6].

We utilize the LLM with E as demonstration for in-context learning to generate medical contexts and preliminary decisions for all instances in D. For $1 \leq i \leq N$, we concatenate all instances in E, k_i, and A_i and then feed this concatenated string into the LLM for decoding. After this, we parse the decoded sentence into two parts: the context part C_i, which includes overall and specific contexts, and the preliminary decision part d_i[7], which serves as a performance metric for the LLM. All instances of C_i are used for SLM training, regardless of the LLM's preliminary decision accuracy. We find that these contexts contain valuable medical knowledge even with incorrect decisions. We defer more details of our experiments into Section 3.5.1.

---

[5] We set $M = 5$ for our experiments.

[6] See Appendix A.6 for prompt details.

[7] We only use d_i to evaluate LLM performance and do not add this into SLM training.

3.3.2 Context-Enhanced Medical Capability in SLMs

Given an augmented dataset $D' = \{(q_i, A_i, C_i, y_i)\}^N$, our goal is to leverage the medical context C_i generated by the LLM to enhance the SLM's medical proficiency and predict y_i for each instance. We omit i for simplicity. Inspired by previous works [83], we treat C as additional input for the SLM to aid the decision-making. For each candidate answer $a_j \in A$, $0 \leq j < |A|$, we provide both the overall context, c_o, and the specific context of the answer, c_{a_j}, concatenating these contexts with the question q and answer a_j. The SLM generates a contextual representation vector $\mathbf{h}_j$ for each choice, which is then fed into a linear layer to produce s_j, a prediction score for the correctness of the answer choice:

$$\mathbf{h}_j = \mathrm{SLM}([q \oplus a_j \oplus c_o \oplus c_{a_j}]),\ s_j = \mathrm{Linear}(\mathbf{h}_j).$$

For each a_j, the score s_j is computed and normalized using the softmax function across all candidate answers, as shown in Figure 3.3 (b). During training, models are optimized to maximize correct answer scores employing standard cross-entropy loss between predictions and ground truths. In the inference phase, s_j is calculated for each a_j, and the answer with the highest score is the predicted answer.

3.4 Experiments

3.4.1 Experimental Setup

Datasets. We evaluate our methods on the first three datasets for in-domain performance and on all four datasets for out-of-domain performance: **1. MedQA** [115] contains 4-way multiple-choice questions from the US Medical Licensing Exam. It has 10,178/1,272/1,273 instances in the training/development/test sets. Results on the development and test sets are reported. **2. HEADQA** [116] features multiple-choice questions from specialized Spanish healthcare exams conducted between 2013 and 2017. The dataset has 2,657/1,366/2,742 instances in the training/development/test sets. We report results on the development and test sets. **3. MedMCQA** [117] is a 4-way multiple-choice dataset from Indian medical school entrance exams. It has 182.8k/4.2k/6.1k instances in the training/development/test sets. We use a randomly selected subset of 10,000 training instances and report results on the development set, following previous work [7]. **4. MMLU-professional medicine** [118] is a 4-way biomedical multiple-choice dataset, with 5/31/272 instances in the training/validation/test sets. We evaluate the Out-of-Domain (OOD) performance of our method on the test set without adaptation.

Context Generation with LLM. We use the *gpt-3.5-turbo* via OpenAI API[8] and employ greedy decoding for in-context learning. Each dataset has five-shot medical examples, shown in Figure 3.3.

Training SLMs with Contextual Information. After acquiring data from LLMs, we utilize BioLinkBert-Base [65], BioLinkBert-Large [65], and BioMedLM [68] as SLM backbones for Fine-Tuning with Context (FTC). We compare FTC with privacy-restricted baselines that leverage additional knowledge to aid medical decision making: QA-GNN [89], GreaseLM [100], DRAGON [90], MurKe [86], MOEBQA [119], HDRN [120] and VOD [121]. Also, we perform SLM standard fine-tuning (SFT) without any external knowledge and LLM prompting with keywords and candidate answers (LLM) to validate our method's efficacy. To ensure a fair comparison, we keep the backbones and hyperparameters consistent for both FTC and SFT approaches. For BioLinkBERT-Base, We conduct three separate runs for each setting and report the average results along with the standard deviation. We report only a single run for BioLinkBERT-Large and BioMedLM due to high computational cost.[9]. The performance are measured by accuracy (%).

---

[8] `https://platform.openai.com/docs/models/gpt-3-5`

[9] We provide implementation and training details in Appendix A.2.

Table 3.1: Performance comparison (%) on MedQA, HEADQA and MedMCQA. Best results are bold and second best are underlined. †: results from their original papers. ¶: results from [90]. ‡ results from [68]. §: VOD uses 180k training instances for MedMCQA, in contrast to our approach which utilizes only 10k instances.

	MedQA		**HEADQA**		**MedMCQA**
	dev	test	dev	test	
QA-GNN [89]	-	45.0¶	-	-	-
GREASELM [100]	-	45.1¶	-	-	-
DRAGON [90]	-	47.5¶	-	-	-
HDRN[120]	-	47.6†	-	-	-
MurKe [122]	-	-	-	46.7†	-
MOEBQA [119]	39.9†	41.6†	44.3†	46.7†	-
VOD [121]					
+ BioLinkBERT&BM25	41.0†	40.4†	-	-	51.6$^{\dagger\ \S}$
+ BioLinkBERT& BioLinkBERT	53.6†	55.0†	-	-	**58.3**$^{\dagger\ \S}$
LLM	38.30	41.70	47.60	47.50	35.20
SFT (w/o additional knowledge)					
+ BioLinkBERT-Base	41.22$_{0.48}$	42.21$_{0.91}$	39.14$_{1.88}$	41.00$_{0.34}$	32.15$_{2.23}$
+ BioLinkBERT-Large	-	45.1‡	39.53	41.61	35.86
+ BioMedLM	-	50.3‡	48.68	50.33	43.63
FTC (Ours)					
+ BioLinkBERT-Base	50.73$_{0.35}$	50.17$_{0.42}$	61.35$_{0.16}$	60.21$_{0.47}$	49.20$_{0.45}$
+ BioLinkBERT-Large	51.02	53.10	62.30	62.18	50.38
+ BioMedLM	**53.85**	**55.90**	**63.10**	**63.17**	52.09

3.4.2 Superior Medical Decision Performance of FTC

Table 3.1 compare results between Fine-Tuning with Context (FTC) and baselines. FTC significantly outperforms standard fine-tuning (SFT), with improvements of up to 7.96%, 21.23%, and 17.05% in absolute accuracy on the test sets of MedQA and HEADQA, and the development set of MedMCQA, respectively. Furthermore, FTC exceeds LLM by 14.20%, 15.75%, and 16.90% on these datasets in privacy-restricted scenarios. These results indicate that SLMs effectively leverage the medical knowledge provided by the LLM to aid decision-making, highlighting the benefits of incorporating context from the LLM in SLMs training process.

FTC with BioMedLM achieves SOTA performance on both MedQA and HEADQA datasets. Particularly, for HEADQA, FTC with BioMedLM backbone outperforms the best baseline MOEBQA [119] by 18.80% and 16.55% in absolute accuracy on the development and test sets, respectively. This demonstrates the considerable impact of the FTC on enhancing the performance of SLMs in medical tasks. Compared to the complex VOD [121] baseline, a retriever-and-reader framework with multiple training strategies, our approach is more straightforward. We simply fine-tune SLMs and include only one context per candidate answer generated by the LLM. Despite this, our method achieves superior performance in MedQA and secures second place in MedMCQA, utilizing less than 6% of the training data required by the VOD [10].

3.4.3 Few-Shot Learning Enhancement with FTC

Real medical environments often face scarce training data [20, 123]. In this section, we explore if additional contexts boost SLM's medical proficiency in few-shot setting. We experiment on BioLinkBERT-Base with training sample sizes of $\{100, 200, 500\}$ for all datasets. For every size, we randomly generate three data splits from the entire training set, performing a single run for each split. Results are shown in Table 3.2.

FTC consistently surpasses SFT by a considerable margin, achieving absolute accuracy enhancements of up to 14.12%, 22.57%, and 11.81% for the test sets of MedQA

---

[10] In theory, FTC could be integrated with VOD [121]. We intend to do this once their code is ready.

Table 3.2: Results (%) of LLM, SFT and FTC under different training sizes. [11]

	MedQA				**HEADQA**				**MedMCQA**			
	100	200	500	full	100	200	500	full	100	200	500	full
LLM	41.70				47.50				35.20			
SFT	$31.40_{0.79}$	$33.79_{1.34}$	$35.27_{0.63}$	$42.21_{0.91}$	$36.54_{0.30}$	$39.00_{0.61}$	$34.88_{2.00}$	$41.00_{0.34}$	$29.07_{0.16}$	$28.31_{0.86}$	$31.43_{1.32}$	$32.15_{2.23}$
FTC	$\mathbf{45.52}_{1.03}$	$\mathbf{46.29}_{0.32}$	$\mathbf{47.87}_{1.41}$	$\mathbf{50.17}_{0.42}$	$\mathbf{55.03}_{1.10}$	$\mathbf{56.18}_{0.76}$	$\mathbf{57.45}_{0.53}$	$\mathbf{60.21}_{1.47}$	$\mathbf{38.82}_{1.03}$	$\mathbf{40.12}_{0.58}$	$\mathbf{43.06}_{1.92}$	$\mathbf{49.20}_{0.45}$

and HEADQA, and the development set of MedMCQA, respectively. These consistent gains demonstrate that our method not only enhances performance in full-training but also proves highly beneficial when training data is limited. Interestingly, SFT with full training data does not surpass LLM in HEADQA and MedMCQA, and achieves only a comparable performance to LLM in MedQA.

In contrast, our FTC method consistently surpasses LLM and SFT with full training data in all tasks, even with as few as 100 training data points. This underscores the efficacy of prompting medical knowledge from the LLM to boost the SLM's medical capacities.

3.4.4 Out-of-Domain (OOD) Generalizability Boost with FTC

To evaluate the generalizability of our approach, we investigate the OOD performance of FTC using BioLinkBERT-Base as the backbone, without additional training. The best model from the source domain in Section 3.4.2 is directly applied to the target domain. Table 3.3 presents the OOD performance.

The OOD performance of SFT is inferior compared to LLM prompting in the target domain, indicating its limited generalization capabilities. Conversely, FTC consistently outperforms both SFT in OOD settings and LLM prompting baselines. This underscores

---

[11] We present results for the test sets of all datasets, excluding MedMCQA. For those datasets with available development set results, the results are provided in the Appendix A.3.

Table 3.3: Accuracy comparison (%) between LLM on the target domain (upper), and FTC and SFT trained on a source domain (lower) and applied directly to the target domain.

	MedQA		**HEADQA**		**MedMCQA**		**MMLU**		
	HEADQA	MedMCQA	MedQA	MedMCQA	MedQA	HEADQA	MedQA	HEADQA	MedMCQA
LLM	41.70		47.50		35.20		52.94		
SFT	35.57$_{0.24}$	31.32$_{2.38}$	35.62$_{3.19}$	34.34$_{4.07}$	30.14$_{1.07}$	31.77$_{0.37}$	41.30$_{6.89}$	38.84$_{2.13}$	32.97$_{6.07}$
FTC	**47.26**$_{0.96}$	**49.25**$_{0.17}$	**55.27**$_{1.28}$	**61.90**$_{0.38}$	**41.14**$_{0.26}$	**45.98**$_{0.78}$	**58.95**$_{1.73}$	**54.53**$_{2.11}$	**54.66**$_{0.17}$

Table 3.4: Accuracy comparison (%) in general domain.

	CSQA	**OBQA**
LLM	41.25	51.60
SFT	62.63$_{0.17}$	56.93$_{0.25}$
FTC	65.87$_{0.23}$	68.60$_{1.43}$

the enhanced generalizability achieved by incorporating the medical context generated by the LLM into the SLM.

3.4.5 Strong General Applicability of FTC

Privacy concerns not only appear in the medical domain. In this section, we investigate whether our method is generally applicable beyond the medical domain. We use two general domain datasets, CommonsenseQA (CSQA) [124] and OpenbookQA (OBQA) [125], under privacy-restricted settings in the full training scenarios. We adopt T5-base [71] as the SLM backbone following previous works [63, 83] and utilize Fusion-in-Decoder

[88] to incorporate contexts. We conduct three separate runs for each setting[12]. Table 3.4 shows the results.

FTC consistently outperforms LLM and SFT baselines on two datasets. Specifically, FTC performs 3.24% and 11.67% better than its standard finetuning counterpart, SFT, in CSQA and OBQA, respectively. This demonstrates the effectiveness of our method, utilizing the LLM as a strong knowledge base and prompting knowledge within the LLM in privacy-restricted scenarios, which in turn enhances the SLM's knowledge capacities and improves decision-making.

3.5 Analysis

3.5.1 Context Analysis

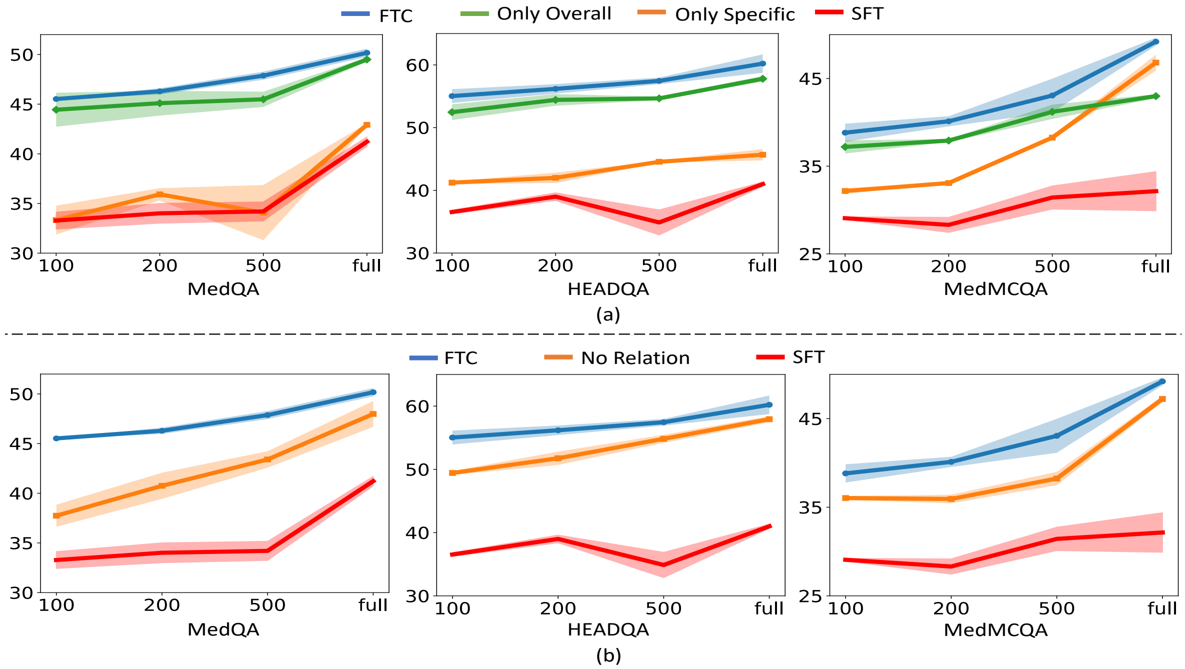
To further understand the effectiveness of context generated by the LLM on SLM performance, we perform ablation studies with BioLinkBert-Base as the SLM backbone.

Role of Context Parts. We investigate the effectiveness of each part of the context by separately providing (1) overall context (*Only Overall*) and (2) specific context for each candidate answer (*Only Specific*) as additional information to the SLM in both few-shot and full-training settings. Results are shown in the upper part (a) of Figure 3.4. Despite decreased performance when providing only overall or specific context compared to FTC, SLMs with added medical context still outperform SFT baselines across various training settings, demonstrating the importance of both context aspects in informed decision-making. The more pronounced performance drop for overall context in most settings could be attributed to it offering general medical knowledge, while specific context provides tailored knowledge for each candidate answer.

---

[12] We defer detailed experiment settings to the Appendix A.5.

Figure 3.4: Results of ablation studies [13]. The upper part examines the effect of context components on SLM training, while the lower part investigates the impact of relationships within the context.

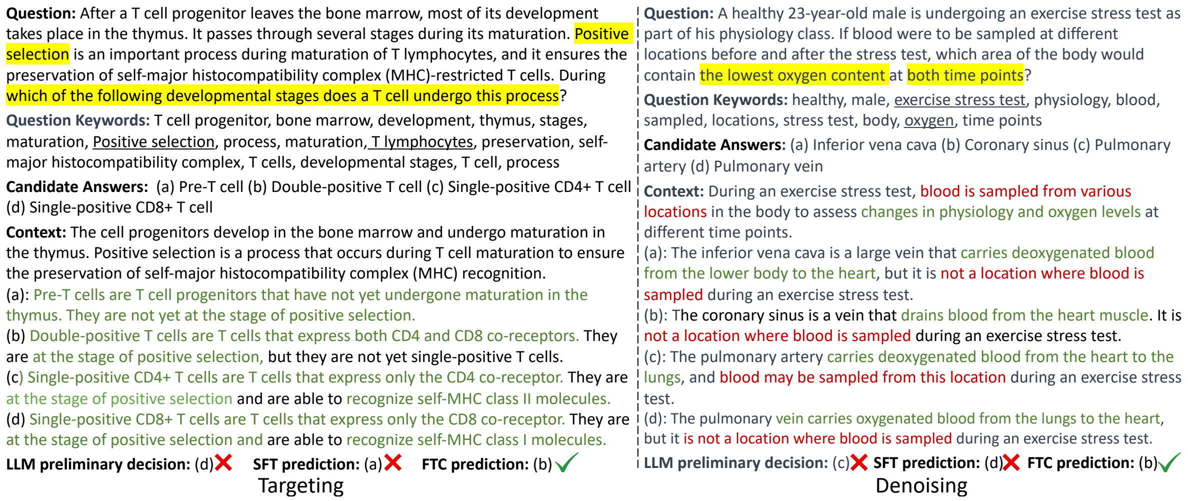


(a)

(b)

Content Learned by SLM. Specific context for each candidate answer includes its relationship to overall context, as shown in Figure 3.3. For example, the relation of answer (a) is ”It is not a treatment for the patient’s symptoms.” We explore whether the SLM learns from medical knowledge or merely cherry-picks relationships by removing all relationship information in specific contexts (*No Relation*), retaining only the knowledge content. Then, we train the SLM using these modified contexts in both few-shot and full-training scenarios. Results are shown in the lower part (b) of Figure 3.4. When relationships are removed, performance declines compared to the FTC. However, even without any relationship information, the SLM with medical contexts consistently and significantly outperforms the SFT. This suggests that, although relationships assist in decision-making, the SLM prioritizes medical knowledge from context over simply repli-

Figure 3.5: Case Study: MedQA test set contexts and predictions. yellow highlights important local information; underlined indicates LLM-selected keywords for context generation; green and red signify correct and incorrect contexts that could aid or confuse the SLM, respectively. FTC succeeded 113 instances where SFT and LLM failed: 45 *Targeting* (left) and 68 *Denoising* (right).



Question: After a T cell progenitor leaves the bone marrow, most of its development takes place in the thymus. It passes through several stages during its maturation. Positive selection is an important process during maturation of T lymphocytes, and it ensures the preservation of self-major histocompatibility complex (MHC)-restricted T cells. During which of the following developmental stages does a T cell undergo this process?

Question Keywords: T cell progenitor, bone marrow, development, thymus, stages, maturation, Positive selection, process, maturation, T lymphocytes, preservation, self-major histocompatibility complex, T cells, developmental stages, T cell, process

Candidate Answers: (a) Pre-T cell (b) Double-positive T cell (c) Single-positive CD4+ T cell (d) Single-positive CD8+ T cell

Context: The cell progenitors develop in the bone marrow and undergo maturation in the thymus. Positive selection is a process that occurs during T cell maturation to ensure the preservation of self-major histocompatibility complex (MHC) recognition.
(a): Pre-T cells are T cell progenitors that have not yet undergone maturation in the thymus. They are not yet at the stage of positive selection.
(b) Double-positive T cells are T cells that express both CD4 and CD8 co-receptors. They are at the stage of positive selection, but they are not yet single-positive T cells.
(c) Single-positive CD4+ T cells are T cells that express only the CD4 co-receptor. They are at the stage of positive selection and are able to recognize self-MHC class II molecules.
(d) Single-positive CD8+ T cells are T cells that express only the CD8 co-receptor. They are at the stage of positive selection and are able to recognize self-MHC class I molecules.

LLM preliminary decision: (d)✗ **SFT prediction:** (a)✗ **FTC prediction:** (b)✓

Targeting

Question: A healthy 23-year-old male is undergoing an exercise stress test as part of his physiology class. If blood were to be sampled at different locations before and after the stress test, which area of the body would contain the lowest oxygen content at both time points?

Question Keywords: healthy, male, exercise stress test, physiology, blood, sampled, locations, stress test, body, oxygen, time points

Candidate Answers: (a) Inferior vena cava (b) Coronary sinus (c) Pulmonary artery (d) Pulmonary vein

Context: During an exercise stress test, blood is sampled from various locations in the body to assess changes in physiology and oxygen levels at different time points.
(a): The inferior vena cava is a large vein that carries deoxygenated blood from the lower body to the heart, but it is not a location where blood is sampled during an exercise stress test.
(b): The coronary sinus is a vein that drains blood from the heart muscle. It is not a location where blood is sampled during an exercise stress test.
(c): The pulmonary artery carries deoxygenated blood from the heart to the lungs, and blood may be sampled from this location during an exercise stress test.
(d): The pulmonary vein carries oxygenated blood from the lungs to the heart, but it is not a location where blood is sampled during an exercise stress test.

LLM preliminary decision: (c)✗ **SFT prediction:** (d)✗ **FTC prediction:** (b)✓

Denoising

cating relationships directly.

Context Case Study. To examine what medical knowledge is generated by the LLM using keywords and candidate answers, and how FTC reasons with these generated contexts, we analyzed instances from the MedQA test set where the FTC correctly predicted answers while both the SFT and LLM failed. Clinicians identified two distinct categories among the cases: (1) *Targeting*, where the LLM successfully refines the target scope and generates high-quality medical contexts, albeit arriving at an incorrect answer; the SLM integrates these contexts with the raw question and correctly predicts the answer. (2) *Denoising*, where the LLM fails to figure out the correct relationship of the correct answer and generates noisy medical knowledge; the SLM effectively obtains useful information, combines it with localized data, and ultimately makes the correct prediction. Figure 3.5 provides examples of each case. This case study demonstrates that LLMs

Table 3.5: Performance comparison of FTC and FTCR in the full-training setting [14].

	MedQA	**HEADQA**	**MedMCQA**
FTCR	$48.12_{0.66}$	$57.79_{0.65}$	$46.55_{0.28}$
FTC	$\mathbf{50.17}_{0.42}$	$\mathbf{60.21}_{1.47}$	$\mathbf{49.20}_{0.45}$

can generate valuable medical information even when making incorrect decisions based on partial data, and that the FTC can extract useful medical knowledge from noisy contexts, thereby enhancing SLM medical reasoning capabilities.

Context Quality. To further quantitatively demonstrate that LLM-generated context retains medical knowledge even with incorrect preliminary decisions. We compared FTC and ***F****ine-****T****uning with* ***C****ontext and* ***R****ejection* (FTCR) in full training settings. FTCR uses context for SLM training only if the LLM's preliminary decision is correct; otherwise, no additional context is provided. The results are in Table 3.5. FTC consistently outperforms FTCR across three medical tasks, implying valuable medical knowledge remains in contexts even with LLM's incorrect decisions. SLM can harness these insights from imperfect contexts to enhance medical capabilities.

3.5.2 Privacy Analysis

We conduct a privacy analysis on MedQA using BioLinkBERT-Base and introduce the *privacy budget*, a metric estimating information usage, presented in Table 3.6. The privacy budget is calculated as the ratio of the number of words provided to the LLM to the total words in the original question. Lower privacy budgets signify better privacy preservation [15].

---

[14] We present results for the test sets of all datasets, excluding MedMCQA. For those datasets with available development set results, the results are provided in the Appendix A.3.

[15] We defer BPC evaluation result on privacy to appendix A.4

Table 3.6: Privacy budget statistics in MedQA. Avg. K and Avg. Q are the average word count for keywords and raw questions, respectively, across the dataset. Budget is privacy budget.

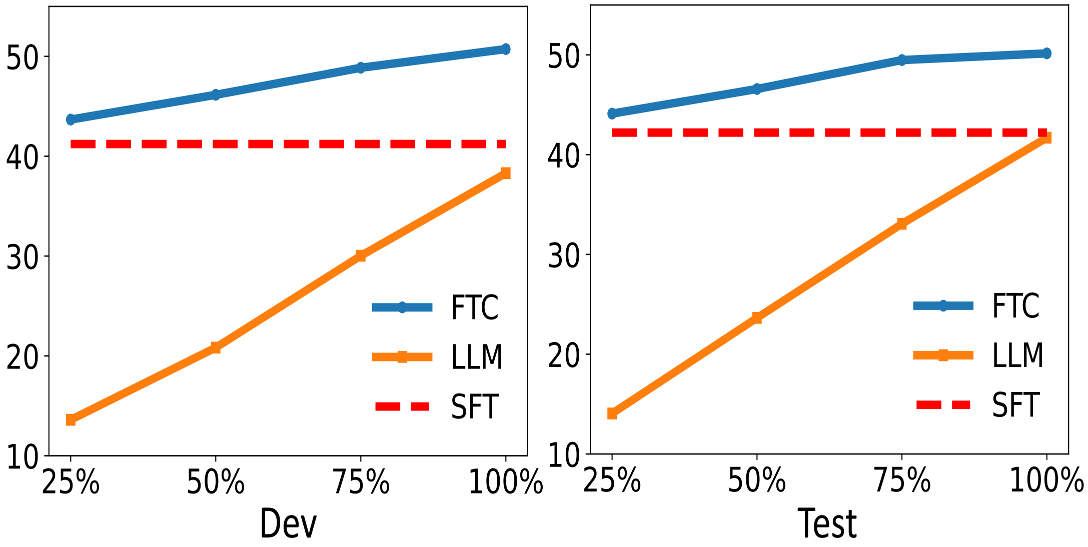
	Avg. K	Avg. Q	Budget
Train + Dev	49.1	116.2	42.3%
Test	50.7	119.6	42.4%

Table 3.7: Results of different information representation methods, maintaining the same privacy budget on MedQA.

	Dev	Test
LLM prompting		
+ Random Span	27.59	28.52
+ Random Words	30.03	30.48
+ Keywords	38.30	41.70
SLM fine-tuning		
SFT	$35.67_{0.47}$	$33.99_{0.87}$
FTC		
+ Random Span	$42.95_{0.33}$	$44.10_{0.80}$
+ Random Words	$44.18_{1.11}$	$45.06_{1.38}$
+ Keywords	$46.42_{0.28}$	$47.91_{0.51}$

Why Keywords? We evaluate the effectiveness of using keywords (*Keywords*) to represent raw data when querying LLMs and compare it to two other methods of raw data representation: (1) random consecutive word spans (*Random Span*), and (2) random word bags from the original data (*Random Words*). Given the same privacy budget, we

Figure 3.6: LLM and FTC results under different keyword usage ratios on MedQA. Standard divisions of FTC and SFT are omitted for simplicity.



randomly select a shared set of 1000 training instances for each method, use these to query the LLM for medical contexts, and then use these contexts as additional input for SLM training. Table 3.7 displays the accuracy of LLM prompting and SLM fine-tuning. All FTC methods consistently outperform the SFT baselines, demonstrating their effectiveness. Keywords perform better than Random Span and Random Words, providing a more efficient representation of medical knowledge within the same privacy budget. LLM prompting performance parallels SLM training, underscoring the importance of raw data representation in context generation and effective SLM training.

Privacy Budget-Model Performance Trade-off. We analyze the trade-off between privacy budget and model performance by generating context from the LLM using randomly selected {25%, 50%, 75%, 100%} of keywords and training the SLM with full training data and corresponding context. Figure 3.6 displays the accuracy for LLM prompting and SLM fine-tuning at various privacy budgets. As privacy budget increases, performance improves. Impressively, FTC outperforms SFT using context from just 25%

of keywords, resulting in LLM prompting performance below 15%—significantly lower than the 25% random guess rate. This suggests that LLM-generated context maintains essential medical knowledge despite limited raw data information, and the SLM effectively learns from it.

3.6 Conclusion

We introduce a simple yet effective pipeline that enhances the SLM performance in medical tasks by using medical keywords to prompt LLMs within privacy-restricted scenarios. Our experimental results across three medical tasks in various training settings underscore the effectiveness of our proposed approach. Through a comprehensive analysis, we gain a deeper understanding of our method capabilities and the impact of LLMs on SLM performance in privacy-restricted scenarios.

Chapter 4

AlpaCare: Instruction Fine-tuned Large Language Models for Medical Applications

4.1 Introduction

Recent advancements in the training of large language models (LLMs) have placed a significant emphasis on instruction-finetuning (IFT), a critical step in enabling pre-trained LLMs to effectively follow instructions [3, 11, 12]. However, relying solely on NLP benchmarks to create instructional datasets can lead to 'game-the-metric' issues, often failing to meet actual user needs [3]. To better align with human intent, [126] introduces the concept of fine-tuning LLMs using diverse machine-generated instruction-response pairs. Subsequent works further highlight the importance of diversity in IFT datasets [127, 128, 129, 130, 131]. However, how to improve dataset diversity in the medical domain for aligning with various user inquiries is still underexplored.

LLMs have demonstrated significant potential in the medical domain across various

```
topic: Pharmacology
view: Pharmacy Student
type: Classifications
difficulty: 1

instruction: Classify these drugs as either
antibiotics, antivirals or antifungals.
input: Penicillin, Lamivudine, Fluconazole
```

Figure 4.1: **Selected example from the clinician-crafted seed set.** We focus on 4 perspectives: *topic*, *viewpoint*, *task type*, and *difficulty level*, to improve the seed set diversity. The set is further used to query GPT-4 to generate medical tasks.

applications [7, 132, 6, 133, 8, 21]. To alleviate privacy concerns and manage costs, several medical open-source LLMs [13, 14, 15, 134] have been developed by tuning LLaMA [16, 135] on medical datasets. Even substantial volumes, these datasets are limited in task scopes and instructions, primarily focusing on medical benchmarks or specific topics, due to the high cost of collecting real-world instruction datasets [126], particularly when extending further into the medical domain[136, 137]. This lack of diversity can negatively impact the models' ability to follow instructions in various medical applications and their effectiveness in the general domain. Therefore, there is an urgent need for a method to generate diverse medical IFT datasets that align with various domain-specific user inquiries while balancing cost.

To bridge this gap, inspired by [126], we propose a semi-automated process that uses GPT-4 [5] and ChatGPT [4] to create a diverse medical IFT dataset for tuning a medical LLM, which can better align with various domain-specific user intents. Initially, to guide the overall task generations with meaningful medical instructions and considering different user needs, we create a high-quality seed set of 167 clinician-curated tasks spanning various medical topics, points of view, task types, and difficulty levels, as shown in Figure 4.1. To automatically generate a broader array of tasks for training, we prompt

GPT-4 to create instructions for new medical tasks by leveraging the existing clinician-curated tasks as demonstrations. After generating tasks and conducting deduplications, we employ ChatGPT to provide responses to the valid tasks. Consequently, we compile a 52k medical self-instruct dataset, *MedInstruct-52k*, which supervises the tuning on the LLaMA series models [16, 135], resulting in *AlpaCare*. Due to the limited number of test sets available for evaluating medical LLMs in terms of instruction-following ability and medical capacity, we introduce a new clinician-crafted free-form instruction evaluation test set, *MedInstruct-test*, covering medical tasks across different difficulty levels.

Our comprehensive experiments within medical and general domains reveal that *AlpaCare*, solely tuned on the 52k diverse medical IFT dataset, exhibits enhanced performance on medical applications and strong generalizability. It achieves up to a 38.1% absolute gain over the best baselines in medical free-form instruction evaluations and a 6.7% absolute gain averaged over multiple general domain benchmarks. Moreover, our human study on free-form instruction evaluations shows that *AlpaCare* consistently produces better responses compared to existing medical LLMs by a large margin in terms of both correctness (+12%) and helpfulness (+49%).

This chapter makes the following contributions:

- To address the challenge of generating cost-effective, high-quality IFT data for LLM alignment in various medical applications, we propose a pipeline to create a diverse medical machine-generated IFT dataset for tuning LLMs.

- We conduct extensive experiments in medical and general domains, demonstrating that tuning LLMs with a diverse medical IFT dataset can boost their capacity in medical applications and generalization simultaneously.

- We release, *MedInstruct-52K*, a diverse medical IFT dataset comprising 52K instruction-response pairs and, *MedInstruct-test*, a test set of 216 clinician-crafted novel medical tasks, to facilitate the building and evaluation of medical LLMs.

4.2 Related Work

IFT. Closed-form IFT [12, 11] creates IFT datasets from existing NLP benchmarks to improve model generalization on new tasks. For instance, FLAN [12], T0 [138], and Flan-T5 [80] construct their IFT datasets using existing NLP tasks with carefully designed instructions and demonstrate that fine-tuning with diverse task instructions enhances performance on unseen tasks. However, these closed-form instructions are often simpler compared to cases in real-world scenarios, making their models fail to align with various real-world user intentions. Alternatively, [3] collects a diverse IFT dataset with real-world instructions and responses, which is rich in both instruction forms and task types, and trains GPT-3 [139] to obtain InstructGPT [3] on this dataset, showing promising results in aligning with diverse actual user needs. Due to the closed-source propriety of strong LLMs (e.g. ChatGPT and GPT-4), various open-source instruction fine-tuned models [127, 128, 129, 140] have been proposed to fine-tune open-source LLMs using datasets obtained from these strong teacher models to enhance their instruction-following abilities. Alpaca [127] creates a 52k diverse machine-generated IFT dataset by distilling knowledge from the "teacher" Text-Davinci-003 [82, 63]. [140] utilizes the same instructions with Alpaca but adopts GPT-4 as the "teacher" LLM to generate higher-quality and more diverse responses to improve the model's alignment on 3H (Helpfulness, Honesty, and Harmlessness) [141]. Vicuna [129] is trained on the ShareGPT data [142], which contains actual ChatGPT users' diverse instructions, obtaining strong response quality and instruction-following ability. However, creating diverse IFT datasets for aligning models with various domain-specific user intentions in the medical domain remains underexplored.

LLMs in Biomedicine. Closed-source LLMs have demonstrated significant proficiency in the medical domain [4, 5, 133, 6]. ChatGPT demonstrates promise in the US Medical Exam [79] and serves as a knowledge base for medical decision-making [21]. The MedPaLM [6, 133] have shown performance in answering medical questions on par with that of medical professionals. GPT-4 [5] obtains strong medical capacities without specialized training strategies in the medical domain or engineering for solving clinical tasks [7, 132]. Due to privacy concerns and high costs, several open-source medical LLMs [134, 13, 14, 15] have been built by tuning open-source base model, such as LLaMA [16, 135], on medical corpus. However, due to the high cost of collecting diverse real-world user instructions [126], their datasets are limited in diversity and primarily focus on medical benchmarks or narrow task scopes, such as doctor-patient conversations. ChatDoctor [14] is fine-tuned using 100k online doctor-patient dialogues, while Baize-Healthcare [134] employs about 100k Quora and MedQuAD dialogues. MedAlpaca [13] utilizes a 230k dataset of question-answer pairs and dialogues. PMC-LLAMA [15] continually trains LLaMA with millions of medical textbooks and papers, and then tunes it with a 202M-token dataset formed by benchmarks and dialogues in the IFT stage. However, these datasets are limited in diversity, mainly focusing on benchmarks or within certain topics, hampering models' medical instruction-following ability and generalizability. To address this, we propose building a cost-effective diverse medical machine-generated IFT dataset by using GPT-4 and ChatGPT for model tuning to better align the model with various medical user intents. Follow-up works by others [143, 144] consistently show the benefits of tuning medical LLMs with diverse machine-generated datasets.

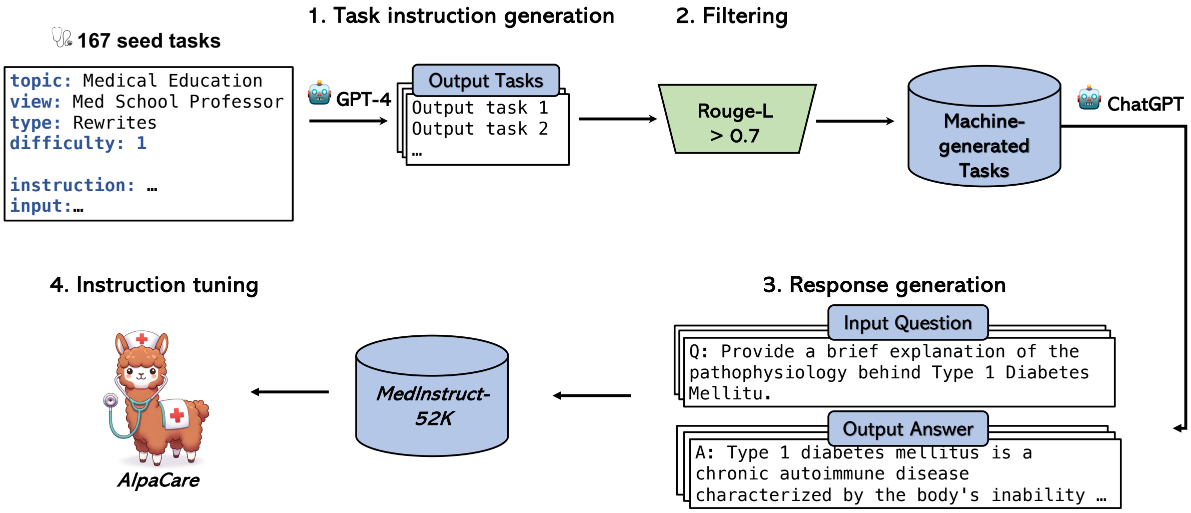


Figure 4.2: **The pipeline of *AlpaCare*.** The process starts with a small set of clinician-curated seed tasks. **1. Task instruction generation:** GPT-4 iteratively generates a series of new task instructions using 3 tasks from the seed set. **2. Filtering:** Ensures textual diversity by removing similar instructions via Rouge-L. **3. Response generation:** ChatGPT creates responses for each task, forming *MedInstruct-52K*. **4. Instruction tuning:** The dataset is used to fine-tune LLaMA models, developing *AlpaCare*.

4.3 Method

Collecting a large-scale medical IFT dataset is challenging because it necessitates 1) a deep understanding of the specific domain knowledge and 2) creativity in designing novel and diverse tasks by considering different real-world medical needs. To mitigate human effort while maintaining high quality, we propose a pipeline by instructing GPT-4 and ChatGPT to create a machine-generated dataset containing diverse domain-specific tasks. The process starts with utilizing a small set of high-quality clinician-curated seed tasks with 167 instances to prompt GPT-4 in generating medical tasks. Similar instructions are removed from the generated medical tasks, preserving 52k instances which are subsequently inputted into ChatGPT for response generation. The instruction-response pairs dataset, *MedInstruct-52k*, is used to tune the LLaMA, resulting in *AlpaCare* with superior medical instruction-following ability and generalizability. The pipeline is shown in Figure 4.2.

4.3.1 Clinician-curated seed dataset

A diverse and high-quality seed task is essential for prompting LLMs in automatic task generation [126]. We focus on 4 key areas, taking into account various user intents in medical applications, to improve the diversity of seed instructions: *topic*, *view*, *type*, and *difficulty level*. Specifically, the *topic* covers various submedical domains, such as radiology, genetics, and psychophysiology. The *view* is derived from diverse medical personnel, including researchers, medical students, and patients, who have different inquiries, to ensure a comprehensive range of viewpoints based on various levels of domain knowledge. For the *type*, we include various task formats, such as summarization, rewriting, single-hop, and multi-hop reasoning, to align with different application needs. Lastly, each task is categorized by its medical *difficulty level*, ranging from 1 to 5 (low to high), to ensure that the seed tasks can prompt new tasks on a wide range of expertise levels. We defer the explanation of the difficulty score to Appendix for further clarification. A clinician crafts each task considering these 4 dimensions, and each task contains instruction and may have a corresponding input, which could be a detailed medical example to further elucidate the instruction and enhance task diversity. Examples are shown in Figure 4.1.

4.3.2 Medical IFT dataset generation and LLM tuning

We utilize GPT-4 for in-context learning by randomly selecting 3 tasks from the seed set and generating 12 tasks for each run. To ensure generated task diversity, we instruct GPT-4 to consider the 4 aspects outlined in 4.3.1. Detailed prompt instructions are provided in Table B.2 in the Appendix. To further amplify textual diversity, instructions with a Rouge-L similarity above 0.7 to any other generated task are discarded [126]. Due to the lengthy propriety of medical text, we separately generate responses for each task using ChatGPT (GPT-3.5-turbo), which has demonstrated efficacy in the medical domain [21]. Finally, we result 52k machine-generated medical instruction-response pairs, *MedInstruct-52k*. To verify *medInstcut-52k*'s data quality, we randomly select 50 instances for a clinician to evaluate, resulting in 49 out of 50 responses being graded as correct, which demonstrates the dataset's high quality.

In the IFT stage, we adopt the same training prompt and hyper-parameter setup as [127] to fine-tune LLaMA models on *MedInstruct-52k*, Specifically, we employ instructions and inputs (when available) as inputs to tune the model to generate corresponding response outputs through a standard supervised fine-tuning with cross-entropy loss. We defer hyper-parameter setup into Appendix B.3.

4.4 Experimental Setup

4.4.1 Free-form Instruction Evaluation

Table 4.1: ***MedInstruct-test* statistics.** The distribution of task counts across various difficulty levels in *MedInstruct-test* is approximately equal to comprehensively evaluate medical proficiency.

Difficulty Level	1	2	3	4	5
Count	44	46	41	41	44

Datasets. To evaluate the effectiveness of LLMs in a user-oriented manner, we conduct free-form instruction evaluations on two medical datasets. (1) iCliniq[1], a dataset comprising transcripts of real patient-doctor conversations collected from an online website [14]. In this task, the model processes patient inquiries as input and then simulates a doctor to provide corresponding answers. (2) *MedInstruct-test*, a dataset created by our clinicians, includes 216 medical instructions. These instructions mimic inquiries posed by different medical personnel, varying in difficulty on a scale from 1 to 5, with 1 being the simplest and 5 being the most challenging. We present the statistics of this dataset in Table 4.1, and the description of the difficulty levels is provided in Appendix 4.3.1.

Evaluation Metric. We conduct auto-evaluation by employing GPT-3.5-turbo to serve as a judge [145]. The judge pairwise compares responses from a model with reference responses produced by another LLM API for each instruction in the test sets. To conduct a holistic evaluation, we employ reference outputs generated by 4 different APIs: Text-davinci-003, GPT-3.5-turbo, GPT-4 and Claude-2, respectively. To ensure unbiased evaluation and avoid positional bias [146], we evaluate each output comparison twice, alternating the order of the model output and the reference output. We follow [147] to score our models and baselines by calculating the win rate. To ensure fair comparisons, we set the maximum token length to 1024 and utilize greedy decoding for the generation of all model outputs and reference responses.

4.4.2 Benchmark Evaluation

Datasets. We further evaluate *AlpaCare* on 4 medical multiple-choice benchmarks, namely MedQA [136], HeadQA [148], PubmedQA [137], and MedMCQA [117], as well as a summarization dataset, i.e., MeQSum [149][2], to assess the model's medical capacity.

Evaluation Metric. Following [150], we conduct the multiple-choice benchmark evalu-

---

[1] We randomly selected 1,000 instances for evaluation from the 10,000 instances proposed by [14].

[2] We randomly selected 200 out of 1000 instances in MeQSum.

ation and report the accuracy. For the summarization task, we utilize greedy decoding with a maximum token length of 1024 to generate outputs and report the ROUGE-L score.

4.4.3 Baselines

We evaluate the performance of *AlpaCare* by comparing it with both general and medical LLMs based on the LLaMA models. We consider a range of models including: (1) Alpaca, tuning on 52k general domain machine-generated samples with responses from Text-davinci-003; (2) ChatDoctor, fine-tuning with 100k real patient-doctor dialogues; (3) MedAlpaca, utilizing approximately 230k medical instances such as Q&A pairs and doctor-patient conversations; (4) PMC-LLaMA (PMC), a two-step tuning model that was first trained on 4.8 million biomedical papers and 30k medical textbooks, then instruction-tuned on a corpus of 202 million tokens; and (5) Baize-Healthcare (Baize-H), training with around 100k multi-turn medical dialogues.

4.5 Experiment Results

4.5.1 Main Results

Free-form Instruction Evaluation Performance. The evaluation results for 4 reference models on both datasets are summarized in Table 4.2. *AlpaCare* outperforms its general domain counterpart, Alpaca, demonstrating that domain-specific training bolsters medical capabilities. Despite tuning with only 52k medical instruction-response pairs, *AlpaCare* consistently and significantly surpasses other medical models, which are trained on considerably larger datasets, across various reference LLMs. Specifically, for average scores across reference models, *AlpaCare* demonstrates a relative gain of 130%

Table 4.2: **Comparative analysis of free-form instruction evaluation.** Performance comparison of *AlpaCare* and instruction-tuned baselines. GPT-3.5-turbo acts as a judge for pairwise auto-evaluation. Each instruction-tuned model is compared with 4 distinct reference models: Text-davinci-003, GPT-3.5-turbo, GPT-4, and Claude-2. 'AVG' denotes the average performance score across all referenced models in each test set.

	iCliniq					**MedInstruct**				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	38.8	30.4	12.8	15.6	24.4	25.0	20.6	21.5	15.6	22.5
ChatDoctor	25.4	16.7	6.5	9.3	14.5	35.6	18.3	20.4	13.4	18.2
Medalpaca	35.6	24.3	10.1	13.2	20.8	45.1	33.5	34.0	29.2	28.1
PMC	8.3	7.2	6.5	0.2	5.5	5.1	4.5	4.6	0.2	4.6
Baize-H	41.8	36.3	19.2	20.6	29.5	35.1	22.2	22.2	15.6	26.6
AlpaCare	**66.6**	**50.6**	**47.4**	**49.7**	**53.6**	**67.6**	**49.8**	**48.1**	**48.4**	**53.5**

on iCliniq and 90% on MedInstruct, respectively, compared to the best baselines. These results highlight the advantages of improving medical proficiency by training with a diverse, domain-specific IFT dataset. Surprisingly, medical LLMs don't always outperform general ones in medical tasks, and some even fail to generate useful responses, possibly due to their limited training scope restricting conversational skills.

Benchmark Evaluation Performance. Table 4.3 presents an extensive evaluation of *AlpaCare* on 5 medical benchmarks. *AlpaCare* obtain the best performance on average, highlighting its robust capability in the medical domain. Benchmarks evaluate a model's intrinsic knowledge[150], which is mainly gained in LLM pretraining instead of instruction fine-tuning [151]. *AlpaCare*'s strong medical capability, combined with its superior ability to follow medical instructions, enables it to meet a wide range of medical application needs effectively.

Table 4.3: **Results on medical benchmarks.** 'AVG' represents the mean performance score across tasks.

	MEDQA	HeadQA	PubmedQA	MEDMCQA	MeQSum	AVG
Alpaca	35.7	29.1	**75.4**	29.2	24.4	38.8
ChatDoctor	34.3	30.0	73.6	**33.5**	27.1	39.7
Medalpaca	**38.4**	30.3	72.8	31.3	11.0	36.8
PMC	34.2	28.1	68.2	26.1	9.2	33.2
Baize-H	34.5	29.3	73.8	32.5	8.1	35.6
AlpaCare	35.5	**30.4**	74.8	**33.5**	**29.0**	**40.6**

4.5.2 Generalizability Evaluation

Training models with specific data may lead to catastrophic forgetting, limiting their generalizability kirkpatrick2017overcoming. Our approach, instruction tuning a model with a diverse, domain-specific dataset, aims to improve its generalizability simultaneously. We test this using *AlpaCare* in AlpaFarm [152], MMLU [118], BBH [153] and TruthfulQA [154]. We compare *AlpaCare* with 4 reference LLMs in AlpaFarm and report the average score, and follow [155] to holistically evaluate models' general domain knowledge on MMLU (5-shot) and BBH (3-shot), receptively; and evaluate the model truthfulness on TruthfulQA (0-shot) with eval-harness. The results are shown in Table 4.4. The detailed score for each reference model on AlpaFarm and more general domain experiment are deferred to Table B.5 and Table B.6 in the Appendix.

Medical LLMs typically lag behind or achieve results comparable to the general domain LLM, Alpaca, in terms of generalizability. However, AlpaCare significantly outperforms both medical and general domain baselines in multiple general tasks on average. Specifically, *AlpaCare* shows a significant relative improvement of 57.8% on AlpacaFarm compared to the best baseline, demonstrating strong general instruction-following ability.

Table 4.4: **Performance on general domain tasks.** AlpacaFarm is a free-form instruction evaluation, MMLU and BBH are knowledge benchmarks and TruthfulQA is a truthfulness task. 'AVG' denotes the average score across all tasks.

	AlpacaFarm	MMLU	BBH	TruthfulQA	AVG
Alpaca	22.7	40.8	32.4	25.6	30.4
ChatDoctor	21.2	34.3	31.9	**27.8**	28.8
Medalpaca	25.8	41.7	30.6	24.6	30.7
PMC	8.3	23.6	30.8	23.8	21.6
Baize-H	18.3	36.5	30.1	23.5	27.1
AlpaCare	**40.7**	**45.6**	**34.0**	27.5	**37.0**

Moreover, *AlpaCare* scores higher in general knowledge tasks and maintains comparable truthfulness scores compared to other baselines, indicating strong generalization abilities due to high data diversity.

4.5.3 Ablation Study

To further understand the effectiveness of *AlpaCare*, we conduct systematic ablation studies on two medical free-form instruction evaluations and report the average score of each task across 4 reference models, receptively. The results of each reference model are detailed in the Appendix B.4.3.

***AlpaCare* consistently delivers superior performance in 13B model comparisons.** To explore the impact of scaling up the LLM backbone, we fine-tune *AlpaCare*-13B on LLaMA-13B and compare its performance against other 13B IFT baselines.The results are shown in Table 4.5.

Table 4.5: **Result comparison on 13B instruction-tuned models.** Comparing the average performance of *AlpaCare*-13B and its 13B baselines with GPT-3.5-turbo as the judge across 4 distinct reference models.

	Alpaca	Medalpaca	PMC	AlpaCare
iCliniq	31.3	3.9	25.4	**54.4**
MedInstruct	26.9	0.1	34.7	**54.5**

AlpaCare-13B consistently outperforms other 13B IFT models in both tasks. This reaffirms the conclusion drawn from the 7B model comparison: tuning models with a diverse medical instruction-following dataset can better align the model with user needs across different medical applications.

***AlpaCare* achieves superior performance across various backbones.** To explore the effect of different LLM backbones, we tune Alpaca-LLaMA2/3 and *AlpaCare*-LLaMA2/3 by training LLaMA2-7B touvron2023llama-2 and LLaMA3-8B llama3modelcard on Alpaca data and *MedInstruct-52k*, respectively. Table 4.6 compares the performance of Alpaca and *AlpaCare* based on different LLM backbone families.

Consistent with the results of using LLaMA-1 as the backbone, *AlpaCare*-LLaMA2/3 consistently and significantly outperforms Alpaca-LLaMA2/3 in both datasets. This further underscores the backbone agnostic property of our method and emphasises tuning with a diverse medical IFT dataset can bolsters models' medical capabilities.

***AlpaCare* shows robust performance across different judges.** Recent studies have highlighted potential biases in the LLM evaluator [146]. ChatGPT may give higher preference on outputs from ChatGPT and GPT-4, which both trained by OpenAI. To robustly evaluate our method, we introduce an alternative judge, Claude-2[10] from An-

Table 4.6: **Results on different LLM backbones.** Comparing the performance of Alpaca and *AlpaCare* using different LLM backbones.

		iCliniq	MedInstruct
LLaMA	Alpaca	24.4	23.2
	AlpaCare	**53.6**	**53.5**
LLaMA-2	Alpaca	30.3	26.8
	AlpaCare	**53.7**	**54.2**
LLaMA-3	Alpaca	26.8	20.7
	AlpaCare	**56.9**	**56.6**

thropic, to mitigate the potential biases of relying on a single family of judges. The results are shown in Table 4.7.

Table 4.7: **Results evaluated by the different judge.** Free-form instruction evaluation with Claude-2 as the judge.

	iCliniq	MedInstruct
Alpaca	26.7	23.5
ChatDoctor	17.4	21.7
Medalpaca	26.7	23.1
PMC	1.3	1.8
Baize-H	25.5	19.8
AlpaCare	**38.8**	**31.5**

Upon evaluation by Claude-2, it is observed that *AlpaCare* consistently outperforms its IFT baselines with a large margin. This aligns with findings from assessments using GPT-3.5-turbo as the judge. Such consistency underscores the superior medical proficiency of our approach.

4.6 Human Study

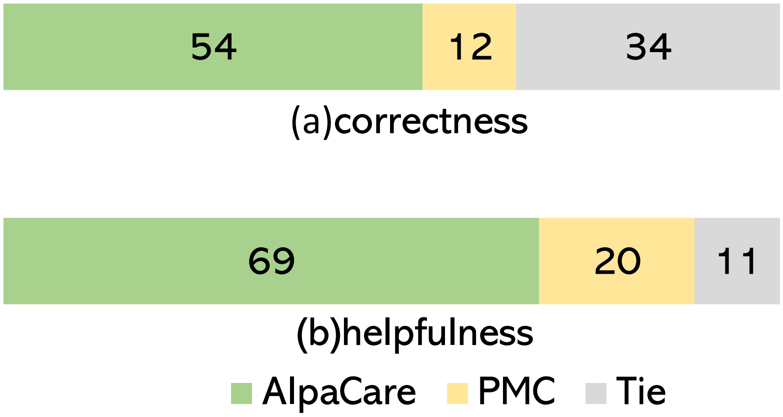


Figure 4.3: **Human study results.** Head-to-head clinician preference comparison between *AlpaCare*-13B and PMC-13B on 100 medical free-form instruction instances on (a) correctness and (b) helpfulness.

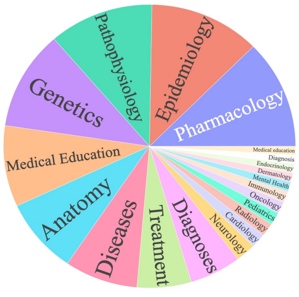
We further conduct human studies to label question-and-answer pairs in medical free-form instruction evaluation. Three annotators with MD degrees in progress are involved in the study to perform pairwise comparisons for each question and answer pair. Specifically, we randomly select 50 prompts from each test set, totaling 100 prompts. These prompts, along with the responses generated by both *AlpaCare*-13B and PMC-13B, the best baseline in the 13B models, are presented to the annotators for evaluation. The evaluation is based on two criteria: correctness and helpfulness. Correctness evaluates whether the response provides accurate medical knowledge to address the question posed, while helpfulness measures the model's ability to assist users concisely and efficiently, considering the user intent. In practical terms, an answer can be correct but not necessarily helpful if it is too verbose and lacks guidance. To determine the final result for each criterion of each evaluation instance, we employ a majority vote method. If at least two of the annotators share the same opinion, their preference is considered the final answer; otherwise, we consider the outputs of the two models to be tied. The results are shown in Figure 4.3.

Consistent with previous results, the *AlpaCare*-13B obtains better performance than PMC-13B in the human evaluation. The output generated by *AlpaCare*-13B is more accurate than that from PMC-13B, with 54% answers preferred by expert annotators. This result shows a superior medical capacity of *AlpaCare*. Beyond correctness, *AlpaCare* generates answers that are significantly more helpful compared to PMC-13B. 69% answers generated by *AlpaCare* are preferred by domain experts, highly exceeding the 11% preference rate for PMC-13B outputs, demonstrating better practical usability. The higher relative gain in helpfulness of *AlpaCare* compared to correctness is expected because the goal of IFT is to obtain the instruction-following ability of LLMs to align the model with various user needs, rather than gaining new knowledge [151].

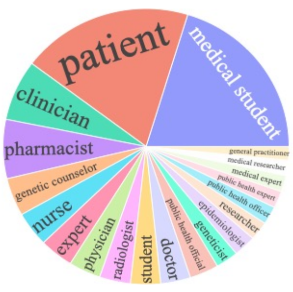
4.7 Analysis & Case Study

4.7.1 Instruction-following Dataset Diversity Analysis

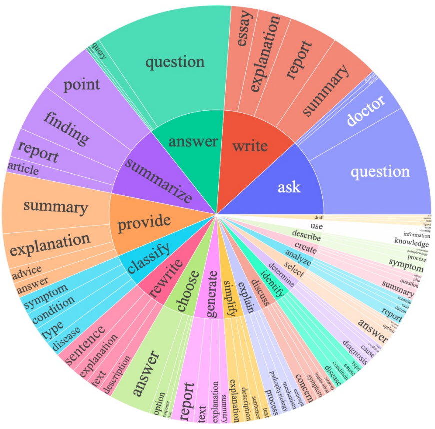
Training a model with diverse instructions enhances its ability to follow instructions [126, 130]. However, current medical IFT models often have training data lacking in instructional diversity, typically using repetitive instructions across different instances [14, 13, 15]. To examine the diversity in our dataset, we plot the distributions of 4

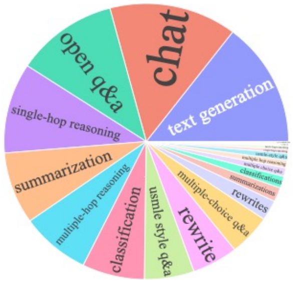


(a) Topic

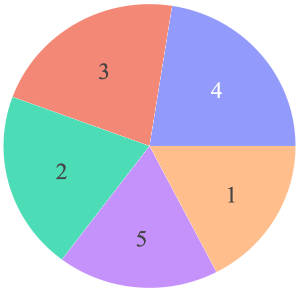


(b) View





(c) Type



(d) Difficulty

(e) Language

Figure 4.4: **Analysis of diversity in the *MedInstruct-52k*.** In panels (a-c), the top 20 entries for topic, view, and type are displayed, respectively. Panel (d) shows the distribution of instruction medical difficulty levels. Panel (e) analyzes linguistic diversity to depict the top 20 root verbs in the inner circle and their 4 primary direct noun objects in the outer circle in the generated instructions.

key areas for instruction generation from *MedInstruct-52k*, shown in Figure 4.4 (a)- (d). Specifically, we present the top 20 topics, views, types, and difficulty levels from 1 to 5, offering insight into training data distribution. We further analyze instruction linguistic diversity by showing the root verbs and their corresponding direct-object nouns from each instruction. The top 20 root verbs and their 4 most common direct-object nouns are displayed in Figure 4.4 (e), representing 22% of the total dataset. Our findings show quite diverse medical intents and textual formats in our *MedInstruct-52k*.

To quantitatively showcase our dataset's diversity in comparison to the tuning data of other medical IFT models, we calculate the linguistic entropy in the instructions of instruction-following datasets used for medical IFT models. Higher entropy values signify greater diversity. Specifically, we analyze the top 20 root verbs and their 4 primary direct noun objects for each dataset and calculate verb-noun pair entropy, as shown in Table 4.8. *AlpaCare*'s dataset, *MedInstruct-52K*, exhibits the highest entropy, underscoring its superior diversity, which enhances the model's instruction-following capabilities.

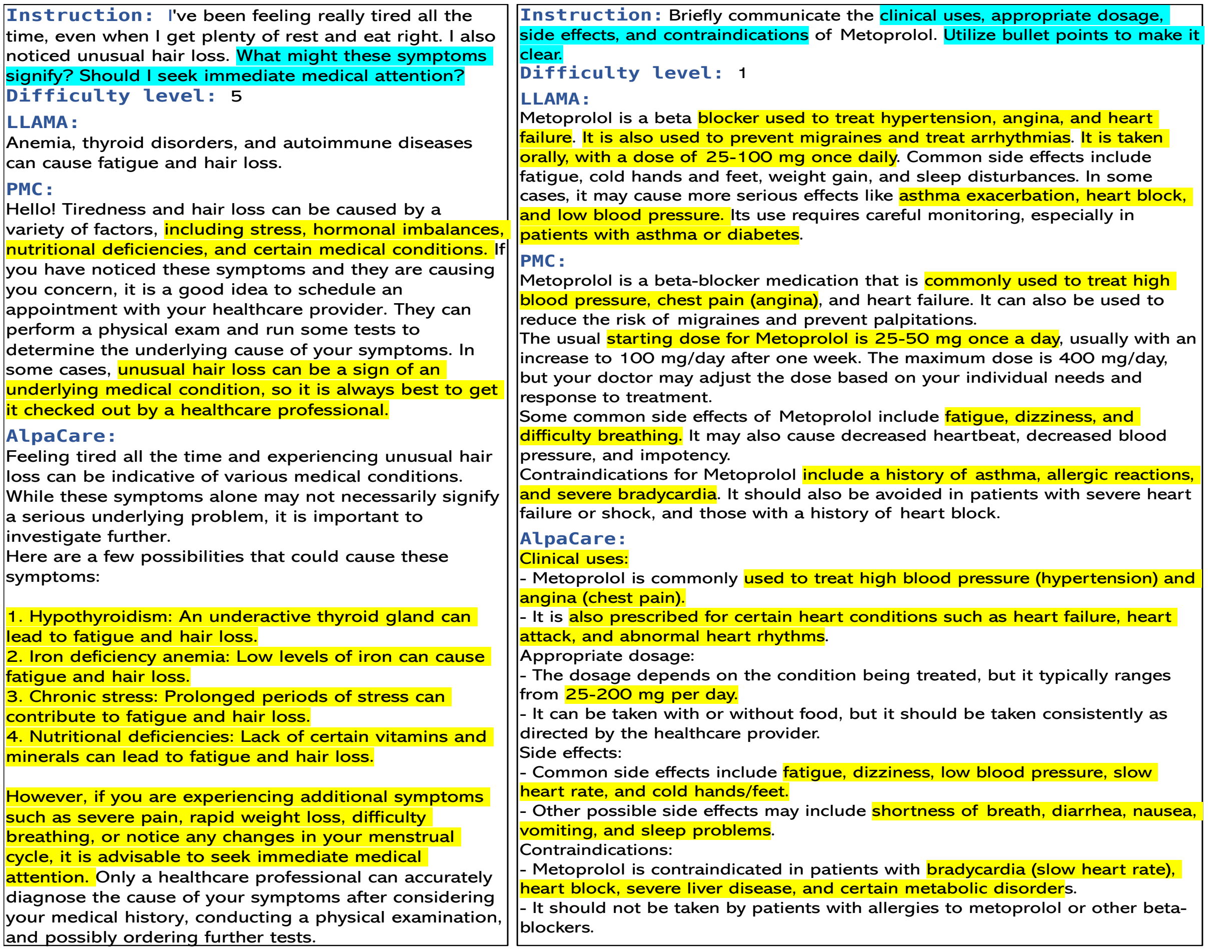
Table 4.8: **Quantitative comparison of linguistic diversity in medical instructional datasets.** Comparing linguistic entropy of each instructional dataset for medical IFT models. The higher value represents better diversity.

	ChatDoctor	Medalpaca	PMC	Baize-H	AlpaCare
Entropy	0	0	2.85	3.45	**5.57**

4.7.2 Generation Case Study

We randomly selected one win case from *MedInstruct-test* for correctness and another for helpfulness, as described in Section 4.6. Figure 4.5 displays the instructions and outputs of the base model, LLAMA-13B, and 13B medical models, *AlpaCare* and PMC.

Figure 4.5(a) showcases a case on correctness with a high medical difficulty level. LLAMA-13B struggles to provide correct answers to the questions in the instruction. In contrast, Medical LLMs demonstrate a marked improvement in the response accuracy. PMC offers a general overview, citing common causes like stress, hormonal imbalances, and nutritional deficiencies. However, it falls short by using ambiguous terms like "certain medical conditions" and "an underlying medical condition," which lack specificity and fail to deliver actionable medical insights. Conversely, *AlpaCare* provides a detailed analysis, pointing specific conditions such as hypothyroidism and iron deficiency anemia that could explain the patient's symptoms and offering a nuanced understanding that significantly



Instruction: I've been feeling really tired all the time, even when I get plenty of rest and eat right. I also noticed unusual hair loss. What might these symptoms signify? Should I seek immediate medical attention?
Difficulty level: 5

LLAMA:
Anemia, thyroid disorders, and autoimmune diseases can cause fatigue and hair loss.

PMC:
Hello! Tiredness and hair loss can be caused by a variety of factors, including stress, hormonal imbalances, nutritional deficiencies, and certain medical conditions. If you have noticed these symptoms and they are causing you concern, it is a good idea to schedule an appointment with your healthcare provider. They can perform a physical exam and run some tests to determine the underlying cause of your symptoms. In some cases, unusual hair loss can be a sign of an underlying medical condition, so it is always best to get it checked out by a healthcare professional.

AlpaCare:
Feeling tired all the time and experiencing unusual hair loss can be indicative of various medical conditions. While these symptoms alone may not necessarily signify a serious underlying problem, it is important to investigate further.
Here are a few possibilities that could cause these symptoms:

1. Hypothyroidism: An underactive thyroid gland can lead to fatigue and hair loss.
2. Iron deficiency anemia: Low levels of iron can cause fatigue and hair loss.
3. Chronic stress: Prolonged periods of stress can contribute to fatigue and hair loss.
4. Nutritional deficiencies: Lack of certain vitamins and minerals can lead to fatigue and hair loss.

However, if you are experiencing additional symptoms such as severe pain, rapid weight loss, difficulty breathing, or notice any changes in your menstrual cycle, it is advisable to seek immediate medical attention. Only a healthcare professional can accurately diagnose the cause of your symptoms after considering your medical history, conducting a physical examination, and possibly ordering further tests.

(a)

Instruction: Briefly communicate the clinical uses, appropriate dosage, side effects, and contraindications of Metoprolol. Utilize bullet points to make it clear.
Difficulty level: 1

LLAMA:
Metoprolol is a beta blocker used to treat hypertension, angina, and heart failure. It is also used to prevent migraines and treat arrhythmias. It is taken orally, with a dose of 25-100 mg once daily. Common side effects include fatigue, cold hands and feet, weight gain, and sleep disturbances. In some cases, it may cause more serious effects like asthma exacerbation, heart block, and low blood pressure. Its use requires careful monitoring, especially in patients with asthma or diabetes.

PMC:
Metoprolol is a beta-blocker medication that is commonly used to treat high blood pressure, chest pain (angina), and heart failure. It can also be used to reduce the risk of migraines and prevent palpitations.
The usual starting dose for Metoprolol is 25-50 mg once a day, usually with an increase to 100 mg/day after one week. The maximum dose is 400 mg/day, but your doctor may adjust the dose based on your individual needs and response to treatment.
Some common side effects of Metoprolol include fatigue, dizziness, and difficulty breathing. It may also cause decreased heartbeat, decreased blood pressure, and impotency.
Contraindications for Metoprolol include a history of asthma, allergic reactions, and severe bradycardia. It should also be avoided in patients with severe heart failure or shock, and those with a history of heart block.

AlpaCare:
Clinical uses:
- Metoprolol is commonly used to treat high blood pressure (hypertension) and angina (chest pain).
- It is also prescribed for certain heart conditions such as heart failure, heart attack, and abnormal heart rhythms.

Appropriate dosage:
- The dosage depends on the condition being treated, but it typically ranges from 25-200 mg per day.
- It can be taken with or without food, but it should be taken consistently as directed by the healthcare provider.

Side effects:
- Common side effects include fatigue, dizziness, low blood pressure, slow heart rate, and cold hands/feet.
- Other possible side effects may include shortness of breath, diarrhea, nausea, vomiting, and sleep problems.

Contraindications:
- Metoprolol is contraindicated in patients with bradycardia (slow heart rate), heart block, severe liver disease, and certain metabolic disorders.
- It should not be taken by patients with allergies to metoprolol or other beta-blockers.

(b)

Figure 4.5: **Case Study** of 13B models of *AlpaCare* and PMC focusing on (a) correctness and (b) helpfulness. Key points in the instructions are highlighted in blue, and primary responses are emphasized in yellow.

enhances the precision and utility of the guidance provided. Additionally, it emphasizes the importance of medical attention for severe symptoms, making its recommendations more actionable.

In the case of helpfulness illustrated in Figure 4.5(b), both the base model and medical models generate accurate medical information on Metoprolol from different perspectives. However, LLAMA and PMC do not adhere to the instruction to use a bullet-point format, which would enhance clarity. Conversely, *AlpaCare* clearly follows the instruction with well-structured formatting, demonstrating that fine-tuning with a diverse instruction-following dataset improves the ability to follow instructions and, in turn, increases helpfulness.

4.8 Conclusion

In conclusion, our study highlights the critical role of data diversity in enhancing medical IFT models. We produce a diversity set of medical instructions using GPT-4, leading to the creation of 52k medical machine-generated instruction-response pairs, termed *MedInstruct-52k*. This dataset is used to fine-tune LLaMA-series models, specifically *AlpaCare*, which exhibits strong medical capacity and robust generalization ability compared to other medical IFT models. Our approach demonstrates versatility across various metrics, including different LLM judges (such as ChatGPT and Claude-2), a range of base model families (like LLaMA-1,2,3), and various model sizes (7B and 13B). This underlines the significant benefits of incorporating diverse data in medical AI model development.

Chapter 5

Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling

5.1 Introduction

ICUs admit patients with life-threatening conditions, e.g. trauma [156], sepsis [157], and organ failure [158]. Care in the first few hours after admission is critical to patient outcomes. This period is also more prone to medical decision errors than later times [26]. Automated tools with effective and real-time predictions can be much beneficial in assisting clinicians in providing appropriate treatments. Recently, the health conditions of patients in ICUs have been recorded in EHRs [159], bringing the possibility of applying deep neural networks to healthcare [30, 29], e.g. mortality prediction [160] and phenotype classification [161]. EHRs contain multivariate irregularly sampled time series (MISTS) and irregular clinical note sequences, as shown in Figure 5.1. The multimodal structure and complex irregular temporal nature of the data present challenges for prediction. This

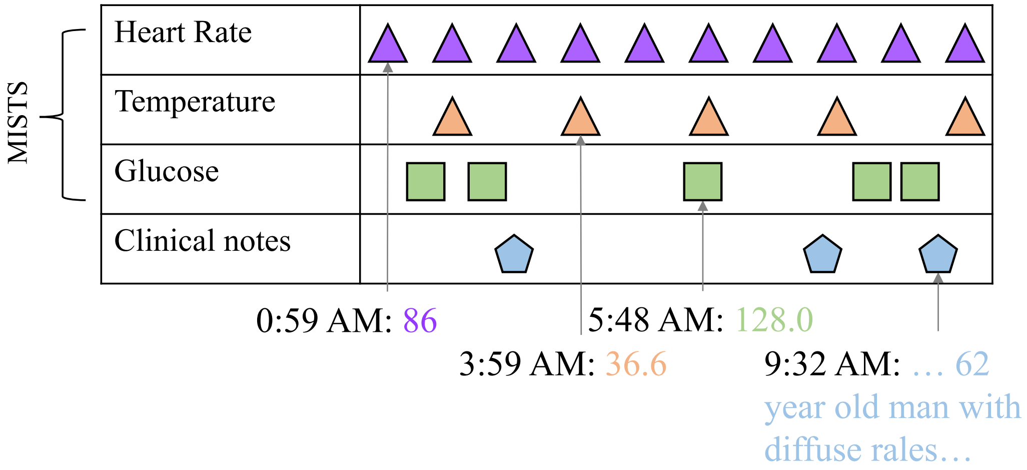


Figure 5.1: An example of a patient's ICU stay includes MISTS with three features and a series of clinical notes. For MISTS, heart rate and temperature are monitored regularly with different frequencies, and glucose is a laboratory test ordered at irregular time intervals based on doctors' decisions. Clinical notes are free text, collected with much sparser irregular time points than clinical measurements.

leads us to formulate two research objectives:

1. *Tackling irregularity in both time series and clinical notes*
2. *Integrating irregularity into multimodal representation learning*

To the best of our knowledge, none of the existing works has fully considered irregularity in multimodal representation learning.

We observed three major drawbacks for irregular multimodal EHRs modeling in existing works. 1) *MISTS models perform diversely.* While the numerous MISTS models have been proposed to tackle irregularity [162, 163, 164, 165, 166, 167], none of the approaches consistently outperforms the others. Even among ***T****emporal* ***d****iscretization-based* ***e****mbedding* (TDE) methods, including hand-crafted imputation [162] and learned interpolation [163, 164], which transform MISTS into regular time representations to interface with deep neural networks for regular time series, there is no clear superior approach. 2) *Irregularity in clinical notes is not well tackled.* Most existing works [17, 18] directly concatenate all clinical notes of each patient but ignore the note-taking time information. Although [168] proposes an LSTM variant to model time decay among clinical notes, this approach utilizes only a few trainable parameters, which could be less powerful. 3) *Exiting works ignore irregularity in multimodal fusion.* [169, 19] have demonstrated the effectiveness of combining time series and clinical notes for medical prediction tasks, however, these works are deployed only on multimodal data without considering irregularity. Their fusion strategies may not be able to fully integrate irregular time information into multimodal representations, which can be essential for prediction performance in real-world scenarios.

Our Contributions. To tackle the aforementioned issues, we separately model irregularity in MISTS and irregular clinical notes, and further integrate multimodalities across temporal steps, so as to provide powerful medical predictions based on the complicated irregular time pattern and multimodal structure of EHRs. Specifically, we first show that different TDE methods of tackling MISTS are complementary for medical predictions, by introducing a gating mechanism that incorporates different TDE embeddings specific to each patient. Secondly, we cast note representations and note-taking time as MISTS, and leverage a time attention mechanism [164] to model the irregularity in each dimension of note representations. Finally, we incorporate irregularity into multimodal representations by adopting a fusion method that interleaves self-attentions and cross-attentions [170] to integrate multimodal knowledge across temporal steps. To the best of our knowledge, this is the first work for a unified system that fully considers irregularity to improve medical predictions, not only in every single modality but also in multimodal fusion scenarios. Our approach demonstrates superior performance compared to baselines in both single modality and multimodal fusion scenarios, with notable relative improvements of 6.5%, 3.6%, and 4.3% in terms of F1 for MISTS, clinical notes, and multimodal fusion, respectively. Our comprehensive ablation study demonstrates that tackling irregularity in every single modality benefits not only their own modality but also multimodal fusion. We also show that modeling long sequential clinical notes further improves medical prediction performance.

5.2 Related Work

Multivariate irregularly sampled time series (MISTS). MISTS refer to observations of each variable that are acquired at irregular time intervals and can have misaligned observation times across different variables [171]. GRU-D [172] captures temporal dependencies by decaying the hidden states in gated recurrent units. SeFT [166] represents the MISTS to a set of observations based on differentiable set function learning. ODE-RNN [167] uses latent neural ordinary differential equations [173] to specify hidden state dynamics and update RNN hidden states with a new observation. RAINDROP [165] models MISTS as separate sensor graphs and leverages graph neural networks to learn the dependencies among variables. These approaches model irregular temporal dependencies in MISTS from different perspectives through specialized design. TDE methods are a subset of methods for handling MISTS, converting them to fixed-dimensional feature spaces, and feeding regular time representations into deep neural models for regular time series. Imputation methods [162, 161, 174] are straightforward TDE methods to discretize MISTS into regular time series with manual missing values imputation, but these ignore the irregularity in the raw data. To fill this gap, [163] presents interpolation-prediction networks (IP-Nets) to interpolate MISTS at a set of regular reference points via a kernel function with learned parameters. [164] further presents a time attention mechanism with time embeddings to learn interpolation representations. However, learned interpolation strategies do not always outperform simple imputation methods. This may be due to complicated data sampling patterns [166]. Inspired by Mixture-of-Experts (MoE) [175, 176], which maintains a set of experts (neural networks) and seeks a combination of the experts specific to each input via a gating mechanism, we leverage different TDE methods as submodules and integrate hand-crafted imputation embeddings into learned interpolation embeddings to improve medical predictions.

Irregular clinical notes modeling. [17, 18] concatenate each patient's clinical notes, divide them into blocks, and then obtain text representations by feeding a series of note blocks into BERT [177] variants [178, 64], ignoring the irregularity in clinical notes. [168] further proposes a time-awarded LSTM with trainable decay function to model irregular time information among clinical notes. However, this approach can be less powerful due to limited parameters. To fully model irregularity, we cast clinical note representations with irregular note-taking time as MISTS, such that each dimension of a series of clinical note representations is an irregular time series, and perform a time attention mechanism [164] to further model the irregularity.

Multimodal fusion. Combining both time series and clinical notes outperforms the results obtained when only one of them is used [179]. [180, 169, 19] directly concatenate representations from different modalities for downstream predictions. [181] utilizes an attention gate to fuse multimodal information. [182] selects multimodal fusion strategies from addition, concatenation and multiplication by a neural architecture search method. However, these fusion methods are only performed on EHRs without considering irregularity, failing to fully incorporate time information into multimodal representations, which is critical in real-world scenarios. To fill this gap, we first tackle irregularity in time series and clinical notes, respectively, and further leverage fusion module, which interleaves self-attentions and cross-attentions [170] to obtain multimodal interaction integrated with irregularity across temporal steps.

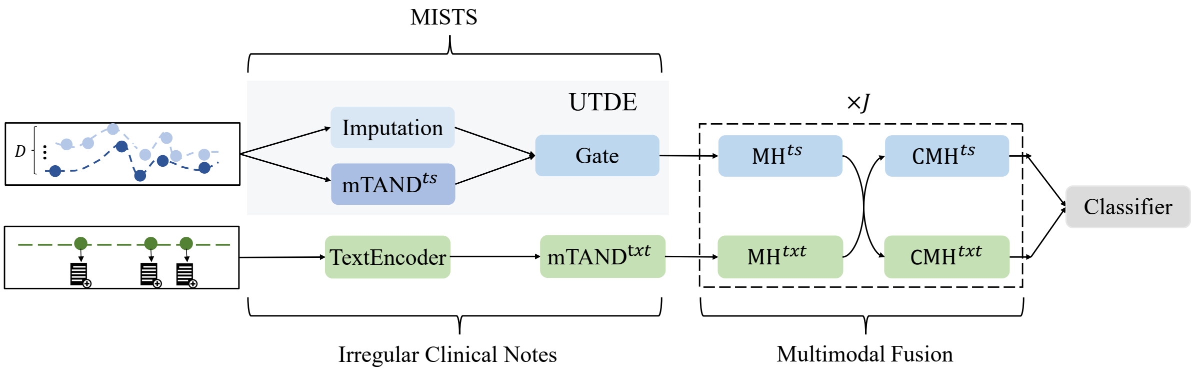


Figure 5.2: The model architecture, which encodes MISTS and clinical notes separately, and then performs a multimodal fusion. UTDE is a gating mechanism to obtain MISTS representations by dynamically fusing embeddings of imputation and a time attention module, mTAND^{ts}. Irregular clinical notes are encoded by a pretrained language model, TextEncoder, whose outputs are fed into mTAND^{txt} to obtain text interpolation representations. The multimodal fusion strategy contains J identical layers. Each layer interleaves self-attentions (MH) and cross-attentions (CMH) to integrate representations from multimodalities and incorporate irregularity into multimodal representations. A classifier with fully connected layers is used to predict patient outcomes.

5.3 Method

Our method models irregularity in three portions: MISTS, clinical notes, and multimodal fusion, as shown in Figure 5.2. In this section, we will illustrate each part thoroughly.

5.3.1 Problem setup

Denote $\mathcal{D} = \{(\mathbf{x}_i^{ts}, \mathbf{t}_i^{ts}), (\mathbf{x}_i^{txt}, \mathbf{t}_i^{txt}), \mathbf{y}_i\}_{i=1}^{N}$ to be an EHR dataset with N patients, where $(\mathbf{x}_i^{ts}, \mathbf{t}_i^{ts})$ is d_m-dimensional MISTS, $\mathbf{x}_i^{ts}$ being observations and $\mathbf{t}_i^{ts}$ being corresponding time points, $(\mathbf{x}_i^{txt}, \mathbf{t}_i^{txt})$ is a series of clinical notes with note-taking time and $\mathbf{y}_i$ is the target outcome, e.g. discharge or death for modality prediction. In the following part, we drop the patient index i for simplicity. Each dimension of the MISTS, $(\mathbf{x}_j^{ts}, \mathbf{t}_j^{ts})$, where $j = 1, \cdots, d_m$, has l_j^{ts} observations, and each patient's $(\mathbf{x}^{txt}, \mathbf{t}^{txt})$ includes l^{txt} clinical notes. In early-stage medical predictions, given $(\mathbf{x}^{ts}, \mathbf{t}^{ts})$ and $(\mathbf{x}^{txt}, \mathbf{t}^{txt})$ before a certain time point (e.g. 48-hour) after admission, α, we seek to predict $\mathbf{y}$ for every patient.

5.3.2 MISTS

TDE methods

We will describe two TDE methods to facilitate the introduction of our proposed MISTS embedding approach. An illustration is shown in Figure 5.3 for better understanding.

Imputation. We first discretize $\mathbf{x}^{ts}$ based on $\mathbf{t}^{ts}$, to hourly time intervals with a sequence of regular time points, $\boldsymbol{\alpha} = [0, 1, \cdots, \alpha - 1]$. Then, for each feature, we use the last observation, if multiple observations are in the same interval, and regard intervals without any observations as missingness. We impute missing values with the most recent observation if it exists, and to the global mean of all patients otherwise. For example, with $\boldsymbol{\alpha} = [0, 1, 2, 3]$ being the first 4-hour prediction, a feature with observations $[10, 8, 12]$ collected at $[1.2, 1.5, 3.7]$ hours after admission is discretized to $[\text{miss}_1, 8, \text{miss}_2, 12]$, where miss_1 and miss_2 will be imputed by global mean and the previous observed value, respectively. The regular time series is fed into a 1D causal convolutional layer with stride 1 to obtain imputation embeddings with hidden dimension d_h, $\mathbf{e}^{ts_{imp}} \in \mathbb{R}^{\alpha \times d_h}$.

Discretized multi-time attention (mTAND). We leverage a discretized multi-time attention (mTAND) module [164] to re-represent MISTS into $\boldsymbol{\alpha}$. To incorporate irregular time knowledge of MISTS, a time representation, Time2Vec [183], is learned to transform each value in a list of continuous time points, $\boldsymbol{\tau}$, with arbitrary length, $l_{\boldsymbol{\tau}}$, to a vector of size d_v and obtain a series of time embeddings $\theta(\boldsymbol{\tau}) \in R^{l_{\tau} \times d_v}$,

$$\theta(\boldsymbol{\tau})[i] = \begin{cases} \omega_i \boldsymbol{\tau} + \phi_i & \text{if } i = 1 \\ sin(\omega_i \boldsymbol{\tau} + \phi_i), & \text{if } 1 < i \leq d_v, \end{cases}$$

where $\theta(\boldsymbol{\tau})[i]$ is the i-th dimension of Time2Vec, and $\{\omega_i, \phi_i\}_{i=1}^{d_v}$ are learnable parameters. The sine function captures periodic patterns while the linear term captures non-periodic behaviors, conditional on the progression of time [183].

The mTAND module leverages V different Time2Vec, $\{\theta_v(\cdot)\}_{v=1}^{V}$, to produce interpolation embeddings at $\boldsymbol{\alpha}$, based on a time attention mechanism. Specifically, similar to the multi-head attention [170], $\{\theta_v(\cdot)\}_{v=1}^{V}$ are performed on $\boldsymbol{\alpha}$ and all dimensions of MISTS to embed all time points to V different d_v-dimensional hidden spaces simultaneously, capturing various characteristics of different time points with regard to the overall time information in different time subspaces. For each $\theta_v(\cdot)$, a time attention mechanism is performed for each dimension of the MISTS simultaneously, which takes $\boldsymbol{\alpha}$ as queries, $\mathbf{t}_j^{ts}$ as keys and $\mathbf{x}_j^{ts}$ as values, and acquires $\hat{\mathbf{x}}_j^{ts} \in \mathbb{R}^{\alpha}$, a series of interpolations of corresponding univariate time series at $\boldsymbol{\alpha}$. Therefore, an interpolation matrix $\mathbf{o}_v^{ts} \in \mathbb{R}^{\alpha \times d_m}$ is obtained by

$$\begin{aligned} \mathbf{o}_v^{ts} &= [\hat{\mathbf{x}}_1^{ts}, \hat{\mathbf{x}}_2^{ts}, \cdots, \hat{\mathbf{x}}_{d_m}^{ts}] \\ \hat{\mathbf{x}}_j^{ts} &= \text{Attn}(\theta_v(\boldsymbol{\alpha})\mathbf{w}_v^q, \theta_v(\mathbf{t}_j^{ts})\mathbf{w}_v^k, \mathbf{x}_j^{ts}) \end{aligned}$$

where $j = 1, \cdots, d_m$, and $\mathbf{w}_v^q$ and $\mathbf{w}_v^k$ are learned parameters. Afterwards, $\mathbf{o}_1^{ts}, \mathbf{o}_2^{ts}, \cdots, \mathbf{o}_V^{ts}$ are further concatenated and linearly projected to obtain mTAND embeddings, $\mathbf{e}^{ts_{attn}} \in \mathbb{R}^{\alpha \times d_h}$.

Unifying TDE methods

The imputation approach ignores the irregularity of the time series, while mTAND could result in worse performance, probably due to different time series sampling strate-

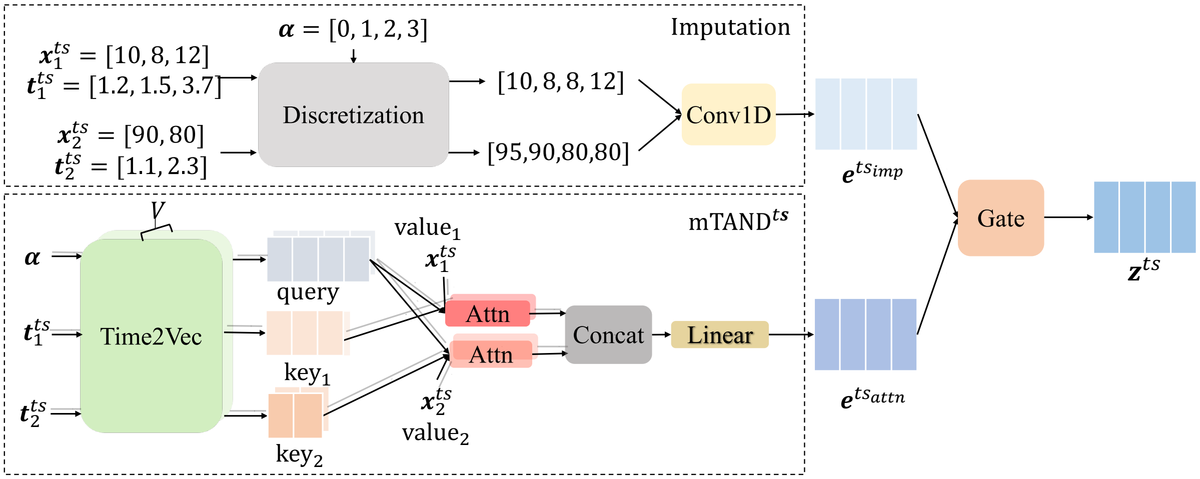


Figure 5.3: Architecture of UTDE module with two input features. UTDE incorporates two TDE methods: Imputation and mTAND^{ts}, as submodules, and learns to integrate different embeddings that are best suited to patients for a given task, via a gating mechanism.

gies [166]. We propose a Unified TDE module, UTDE, via a gate mechanism to take advantage of both, for tackling complex time patterns in EHRs. The architecture of UTDE is illustrated in Figure 5.3. UTDE incorporates Imputation and mTAND as submodules, and learns to dynamically integrate $\mathbf{e}^{ts_{imp}}$ into $\mathbf{e}^{ts_{attn}}$ to obtain compounding embeddings $\mathbf{z}^{ts} \in \mathbb{R}^{\alpha \times d_h}$. Formally,

$$\mathbf{z}^{ts} = \mathbf{g} \odot \mathbf{e}^{ts_{imp}} + (1 - \mathbf{g}) \odot \mathbf{e}^{ts_{attn}}$$

$$\mathbf{g} = f(\mathbf{e}^{ts_{imp}} \oplus \mathbf{e}^{ts_{attn}}),$$

where $f(\cdot)$ is a gating function implemented by MLP for simplicity, $\oplus$ is the concatenation operator and $\odot$ is point-wise multiplication. Specifically, we perform UTDE in 3 levels in which $\mathbf{g}$ has different dimensions : 1) patient level with $\mathbf{g} \in \mathbb{R}$, 2) temporal level with $\mathbf{g} \in \mathbb{R}^{\alpha}$, and 3) hidden space level with $\mathbf{g} \in \mathbb{R}^{\alpha \times d_h}$. The $\mathbf{g}$ on the hidden space level can be more powerful than temporal and patient levels, while it introduces more parameters to update, making the whole module more challenging to optimize. In the experiment section, we use validation sets to decide the level on which to operate.[1] In principle, UTDE can be applied to any two TDE methods. Here, we utilize Imputation and mTAND as submodules based on empirically results.

5.3.3 Irregular clinical notes

To extract relevant knowledge from the clinical notes, we first encode the notes by a in-domain pretrained language model, TextEncoder. Then we extract the representation of the [CLS] token for each encoded clinical note, to obtain a series of note representations, $\mathbf{e}^{txt} \in \mathbb{R}^{l^{txt} \times d_t}$,where d_t is the hidden dimension of the encoded text. Formally,

$$\mathbf{e}^{txt} = \text{TextEncoder}(\mathbf{x}^{txt}).$$

To tackle irregularity, we sort $\mathbf{e}^{txt}$ by $\mathbf{t}^{txt}$ and cast $(\mathbf{e}^{txt}, \mathbf{t}^{txt})$ as MISTS, such that each hidden dimension of $\mathbf{e}^{txt}$ is a time series sequence and every time series sequence has the same collected time points. The mTAND module introduced in section 5.3.2 is further leveraged to re-represent $\mathbf{e}^{txt}$ into $\boldsymbol{\alpha}$. Specifically, the mTAND^{txt} takes $\boldsymbol{\alpha}$ as queries, $\mathbf{t}^{txt}$ as keys and $\mathbf{e}^{txt}$ as values and outputs $\mathbf{z}^{txt} \in \mathbb{R}^{\alpha \times d_h}$, a set of text interpolation representations at $\boldsymbol{\alpha}$. Thus we have

$$\mathbf{z}^{txt} = \text{mTAND}^{txt}(\boldsymbol{\alpha}, \mathbf{t}^{txt}, \mathbf{e}^{txt}).$$

For mTAND^{ts}, the mTAND module for time series, and mTAND^{txt}, we utilize the same $\{\theta_v(\cdot)\}_{v=1}^{V}$ to encode irregular time points of two modalities to obtain temporal knowledge, because all continuous time points are in the same feature space. However, all of the other components in mTAND^{ts} and mTAND^{txt} are learned separately because the

---

[1] we defer more discussion on computation resource of UTDE to Appendix C.1.

representations of time series and clinical notes are in different hidden spaces. Moreover, since the mTAND^{txt} projects $\mathbf{z}^{txt}$ to the same dimension d_h as the $\mathbf{z}^{ts}$, the dot-products are adoptable in attention modules in the fusion.

5.3.4 Multimodal fusion

Previous works [180, 169, 19, 182] perform fusion strategies on multimodal data omitting irregularity. In our work, we first obtain MISTS and irregular clinical note representations, $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$, by UTDE and mTAND^{txt}, respectively. In addition, we leverage an interleaved attention mechanism [170], which fuses $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$ across temporal steps and integrates irregularity into multimodal representations, as shown in Figure 5.2.

Our multimodal fusion module is composed of a stack of J identical layers. Each layer consists of two self-attention sublayers and two cross-attention sublayers across temporal steps to explore the latent interactions between two modalities. Specifically, for each modality in the j-th layer, we first perform a multi-head self-attention (MH) [170] across temporal steps by taking the output of the corresponding modality from the $j-1$-th layer to obtain contextual embeddings. Formally, we acquire the contextual embeddings of time series and clinical notes, $\hat{\mathbf{z}}_j^{ts}$ and $\hat{\mathbf{z}}_j^{txt}$, by

$$\hat{\mathbf{z}}_j^{ts} = \mathrm{MH}_j^{ts}(\mathbf{z}_{j-1}^{ts}), \quad \hat{\mathbf{z}}_j^{txt} = \mathrm{MH}_j^{txt}(\mathbf{z}_{j-1}^{txt}),$$

where $j = 1 \ldots J$, and $\mathbf{z}_0^{ts} = \mathbf{z}^{ts}$ and $\mathbf{z}_0^{txt} = \mathbf{z}^{txt}$. To capture the cross-modal information between two modalities, two multi-head cross-attentions (CMH) [170, 184] are leveraged to learn knowledge of another modality attended by the current modality and vice versa. Specifically, for a time series branch in the j-th layer, a CMH$_j^{ts}$ transforms $\hat{\mathbf{z}}_j^{txt}$ to keys and values to interact with time series modality, and output $\mathbf{z}_j^{ts}$, the time series representations carrying information passed from clinical notes. For the text branch, the same process is performed but transforming $\hat{\mathbf{z}}_j^{ts}$ to keys and values, to output $\mathbf{z}_j^{txt}$, the clinical note representations integrated with information passed from time series. Formally,

$$\mathbf{z}_j^{ts} = \mathrm{CMH}_j^{ts}(\hat{\mathbf{z}}_j^{ts}, \hat{\mathbf{z}}_j^{txt}), \quad \mathbf{z}_j^{txt} = \mathrm{CMH}_j^{txt}(\hat{\mathbf{z}}_j^{txt}, \hat{\mathbf{z}}_j^{ts}).$$

Upon the CMH output of each modality, a position-wise feedforward sublayer is stacked. We apply pre-layer normalizations and residual connections to every MH, CMH and feedforward sublayer. For simplicity, we only draw MH and CMH in multimodal fusion in Figure 5.2.

In this process, each modality alternately collects temporal knowledge by a MH, and updates its sequence via external information from another modality by a CMH. After $\mathbf{z}^{ts}$ and $\mathbf{z}^{txt}$ are passed through J layers, the output of each modality fully integrates information from another modality. Eventually, the last hidden states of $\mathbf{z}_J^{ts}$ and $\mathbf{z}_J^{txt}$ are extracted and concatenated to pass through a classifier with fully-connected layers to make predictions.

5.4 Experiments

To demonstrate the effectiveness of our methods, we conducted comprehensive experiments and ablation studies on two medical tasks: 48-hour in-hospital mortality prediction (48-IHM) and 24-hour phenotype classification (24-PHE), which are critical in the clinical scenario [185, 186].

5.4.1 Experimental setup

Dataset. MIMIC III is a real-world public EHR of patients admitted to ICUs, including numerical time series and clinical notes [9]. We select the MISTS features and extract clinical notes following [161] and [180], respectively. For each task, the data split of training, validation, and testing sets follows [161], and patients without any clinical notes before the prediction time are removed. We defer additional data preprocessing details to the Appendix C.2. After preprocessing, the number of patients in the training, validation and testing sets for the 48-IHM are 11181, 2473 and 2488; and for the 24-PHE, they are 15561, 3410 and 3379, respectively.

Evaluation metric. The 48-IHM is a binary classification problem with label imbalance with death to discharge ratio of approximately 1:7. The 24-PHE is a multi-label classification problem with 25 acute care conditions, which is more changeling due to earlier prediction time and more prediction classes. We measured the performance of our proposed methods and baselines by the F1 and AUPR on 48-IHM and F1(Macro) and AUROC on 24-PHE, following the previous work [57, 187].

MISTS baselines. We compare UTDE with a classical and 5 SOTA baselines of MISTS: Imputation, IP-Net [163], mTAND [164], GRU-D [172], SeFT [166] and RAINDROP [165]. We utilize Transformer [170] as backbone for UTDE and TDE methods, because Transformer has achieved SOTA results in regular time series modeling [188, 189]. We feed time series embeddings into Transformer and extract the last hidden states of the Transformer output to pass through fully-connected layers to make predictions. Following [165], we added two methods initially designed for forecasting tasks, DGM^2-O [190] and MTGNN [191] in our baselines. Details on MISTS baseline descriptions are in the Appendix C.3.1.

Irregular clinical note baselines. Considering the in-domain knowledge and the length of clinical notes, we utilize Clinical-Longformer [192] with a maximum input sequence length of 1024 as our text encoder, which covers more than 98% of notes in both tasks. Same as time series modality, we feed the text interpolation representations obtained by mTAND^{txt} into Transformer for predictions. We compare our method with two baselines: T-LSTM [193], FT-LSTM [168], and GRU-D [172], which shows strong performance in MISTS modeling. All of these methods model irregularity by acquiring a series of clinical note representations with irregular note-taking time information. To demonstrate our method's effectiveness at tackling irregularity, we further introduce two baselines: Flat [169], utilizing the average of clinical note embeddings of a patient for predictions, and HierTrans [194], utilizing Transformer to model sequential relationships among a series of clinical notes representations without considering irregular note-taking time. We defer additional baseline descriptions to the Appendix C.3.2.

Multimodal fusion baselines. To examine the effectiveness of our fusion method, we consider four baselines for fusion: concatenation [180, 169], Tensor Fusion [195, 196], MAG [181, 197], and MulT [184]. While the first three are asynchronous methods that do not consider temporal information, MulT and our method are synchronous relying on a cross-attention mechanism to integrate information across temporal steps. Additional multimodal fusion baseline details can be found in the Appendix C.3.3.

5.4.2 Main results

In this section, we compare results between our proposed methods and their corresponding baselines in MISTS, irregular clinical notes, and multimodal fusion scenarios, respectively. The data split of each task is fixed across all methods. We conduct 3 different runs for each setting and report the corresponding mean values along with the standard deviations in testing sets, based on the best average performance on validation sets. Details for the hyperparameter selection can be found in the Appendix C.4.[2]

MISTS. Table 5.1 compares the UTDE with other time series baselines. UTDE, which incorporates two different TDE methods, obtains the best performance across two tasks on different evaluation metrics, demonstrating the advantages of our hybrid approach for downstream predictions. Specifically, UTDE relatively outperforms the strongest baseline by 4.4% in terms of AUPR on 48-IHM. Additionally, UTDE shows a 6.5% relative

---

[2] All experiments are conducted on 1 RTX-3090.

Table 5.1: Comparison between UTDE and other MISTS methods. We report average performance on three random seeds, with standard deviation as the subscript. The **Best** and <u>2nd best</u> methods under each setup are bold and underlined, respectively. The performance of 48-IHM is measured on F1 and AUPR, and 24-PHE on F1 (Macro) and AUROC, respectively.

		Imputation	IP-Net	mTAND	GRU-D	SeFT	RAINDROP	DGM²-O	MTGNN	UTDE (Ours)
48-IHM	F1	$39.73_{1.39}$	$37.22_{2.75}$	$\underline{43.87}_{0.54}$	$42.82_{0.57}$	$16.46_{8.61}$	$39.46_{3.70}$	$39.08_{1.53}$	$38.60_{2.50}$	$\mathbf{45.26}_{0.70}$
	AUPR	$44.36_{1.36}$	$39.36_{1.10}$	$\underline{47.54}_{1.28}$	$45.90_{0.40}$	$23.89_{0.46}$	$36.23_{0.37}$	$37.79_{1.54}$	$36.49_{2.10}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$\underline{23.36}_{0.45}$	$17.90_{0.66}$	$19.90_{0.38}$	$18.96_{0.99}$	$6.10_{0.15}$	$21.81_{1.71}$	$18.40_{0.18}$	$14.48_{1.69}$	$\mathbf{24.89}_{0.43}$
	AUROC	$\underline{74.93}_{0.22}$	$73.45_{0.10}$	$73.48_{0.11}$	$73.33_{0.10}$	$65.66_{0.11}$	$73.95_{0.89}$	$71.71_{0.16}$	$70.56_{0.68}$	$\mathbf{75.56}_{0.17}$

improvement in F1 score on the more challenging 24-PHE task compared to the best baseline. Excluding UTDE, mTAND and Imputation are the top performers on 48-IHM and 24-PHE, respectively. However, UTDE, which dynamically incorporates Imputation and mTAND, outperforms its submodules for both tasks across various metrics, showing its ability to integrate knowledge and benefit medical predictions.

Irregular clinical notes. We compare our method with baselines in the clinical notes modality in Table 5.2. All of the methods that model the sequential relationships among clinical notes yield better results than Flat by a large margin, demonstrating that exploiting sequential information of clinical notes can significantly improve the downstream predictions. T-LSTM, FT-LSTM and GRU-D outperform or have comparable result compared to HierTrans on 48-IHM, but do not perform well on the more challenging 24-PHE task, where note sequences are sparser. This highlights the difficulty in modeling irregularity in sparse clinical note sequences. The proposed method, mTAND^{txt}, significantly outperforms HierTrans by relative margins of 7.8% and 5.3% in terms of F1 on the 48-IHM and 24-PHE, respectively. This shows the importance of modeling the irregularity present in clinical notes. Additionally, the results show that mTAND^{txt} surpasses other

Table 5.2: Results comparison in the clinical notes modality.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
Flat	$39.78_{1.14}$	$51.69_{0.79}$	$18.14_{1.36}$	$74.81_{0.22}$
HierTrans	$48.76_{2.44}$	$52.98_{1.69}$	$50.25_{1.21}$	$\underline{84.90}_{0.25}$
T-LSTM	$50.32_{0.89}$	$52.57_{3.25}$	$39.13_{1.35}$	$82.03_{0.07}$
FT-LSTM	$48.51_{1.67}$	$\underline{54.39}_{1.38}$	$38.24_{0.61}$	$81.07_{0.27}$
GRU-D	$\underline{51.01}_{1.50}$	$54.34_{0.75}$	$\underline{51.09}_{1.02}$	$84.19_{0.20}$
mTAND^{txt} (Ours)	$\mathbf{52.57}_{1.30}$	$\mathbf{56.05}_{1.09}$	$\mathbf{52.95}_{0.06}$	$\mathbf{85.43}_{0.07}$

irregularity-modeling methods, particularly achieving a 3.6% relative improvement in terms of F1 on the 24-PHE, demonstrating its strong performance in tickling irregularity in clinical notes.

Multimodal fusion. We first obtain MISTS embeddings by UTDE and irregular clinical note embeddings by mTAND^{txt}, since they have the best results in each modality, and then fuse their representations via various multimodal fusion strategies. The results are shown in Table 5.3. Compared to models that use only one source of available data, most fusion strategies achieve better results, illustrating the effectiveness of multimodal fusion. Our fusion method yields better results than baselines for both tasks, achieving a particularly 4.3% relative improvement in F1 on the 48-IHM, showing the power of the interleaved attention mechanism. Synchronous strategies consistently achieve better results than asynchronous methods by incorporating temporal information in multimodal fusion, resulting in better integration of irregularity and fusion of different modalities. Our method further outperforms the MulT, which separately applies a cross-modal Trans-

Table 5.3: Performance comparison of different fusion strategies. Concat and TF use the concatenation and Tensor Fusion method to fuse the two modalities, respectively.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
TS only	$45.26_{0.70}$	$49.64_{1.00}$	$24.89_{0.43}$	$75.56_{0.17}$
Note only	$52.57_{1.30}$	$56.05_{1.09}$	$52.95_{0.06}$	$85.43_{0.07}$
Concat	$52.77_{0.70}$	$57.13_{0.7}$	$53.30_{0.35}$	$85.94_{0.21}$
TF	$51.44_{0.66}$	$57.07_{0.82}$	$49.84_{0.83}$	$84.74_{0.16}$
MAG	$53.20_{2.13}$	$57.86_{1.07}$	$53.73_{0.37}$	$85.94_{0.07}$
MulT	$\underline{54.13}_{1.20}$	$\underline{58.94}_{1.94}$	$\underline{54.20}_{0.33}$	$\underline{85.96}_{0.07}$
Interleaved (Ours)	$\mathbf{56.45}_{1.30}$	$\mathbf{60.23}_{1.54}$	$\mathbf{54.84}_{0.31}$	$\mathbf{86.06}_{0.06}$

former and a self-attention Transformer for each modality. This result shows that alternately obtaining temporal information and cross-modal knowledge for different modalities is more capable of fusing different modalities and integrating irregularity into multimodal representations than learning these two components separately.

5.4.3 Ablation study

UTDE with different submodules in MISTS. UTDE could have incorporated different TDE methods as submodules to obtain fused time series embeddings. We explored the effectiveness of the gate mechanism in UTDE by substituting mTAND to IP-Net in Table 5.4. The $\text{UTDE}_{\text{IP}-\text{Net}}$ underperforms $\text{UTDE}_{\text{mTAND}}$ but still achieves better results than its submodules, Imputation and IP-Net, on both tasks, demonstrating that UTDE successfully learns from different submodules and achieves optimal performance via the gate mechanism.

Table 5.4: Ablation study on the effects of substituting different submodules in UTDE. $\mathrm{UTDE_{IP-Net}}$ consists of IP-Net and Imputation, and $\mathrm{UTDE_{mTAND}}$ incorporates mTAND and Imputation.

		Imputation	IP-Net	$\mathrm{UTDE_{IP-Net}}$	$\mathrm{UTDE_{mTAND}}$
48-IHM	F1	$39.73_{1.39}$	$37.22_{2.75}$	$\underline{44.88}_{1.96}$	$\mathbf{45.26}_{0.70}$
	AUPR	$44.36_{1.36}$	$39.36_{1.10}$	$\underline{45.49}_{3.45}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$23.36_{0.45}$	$17.90_{0.66}$	$\underline{24.06}_{0.51}$	$\mathbf{24.89}_{0.43}$
	AUROC	$74.93_{0.22}$	$73.45_{0.10}$	$\underline{75.17}_{0.07}$	$\mathbf{75.56}_{0.17}$

UTDE with various backbones in MISTS. To evaluate the effectiveness of UTDE across different backbone encoders, we further leverage CNN [198] and LSTM [199] to encode time series representations obtained from TDE and UTDE methods. The results are shown in Table 5.5. The empirical analysis shows that Imputation and mTAND performance varies across different time series encoders. However, UTDE consistently outperforms them, demonstrating the gains of dynamically integrating different time series embeddings for medical predictions regarding the effectiveness and generalizability across time series backbones.

Does UTDE benefit performance in multimodal fusion? We drop UTDE (w/o UTDE) in our fusion model and perform only Imputation (w Imputation) and mTAND (w mTAND^{ts}) to obtain MISTS embeddings, respectively. Table 5.6 shows results. Consistent with the time series modality, the fusion model with learned mTAND embeddings does not consistently outperform the one with classical imputation embeddings, and vice versa. However, our fusion model with UTDE consistently surpasses those using only one

Table 5.5: Comparison of UTDE and its submodules with different time series backbones.

		CNN			LSTM			Transformer		
		Imputation	mTAND	UTED	Imputation	mTAND	UTED	Imputation	mTAND	UTED
48-IHM	F1	$39.66_{1.72}$	$\underline{41.40}_{1.16}$	$\mathbf{44.45}_{1.41}$	$39.72_{0.70}$	$\underline{43.61}_{0.55}$	$\mathbf{44.58}_{0.18}$	$39.73_{1.39}$	$\underline{43.87}_{0.54}$	$\mathbf{45.26}_{0.70}$
	APUR	$41.84_{0.52}$	$\underline{46.62}_{0.27}$	$\mathbf{48.22}_{0.99}$	$42.52_{0.98}$	$\underline{47.36}_{0.67}$	$\mathbf{48.17}_{0.36}$	$44.36_{1.36}$	$\underline{47.54}_{1.28}$	$\mathbf{49.64}_{1.00}$
24-PHE	F1	$\underline{20.09}_{0.70}$	$19.05_{1.17}$	$\mathbf{20.64}_{0.54}$	$19.21_{1.37}$	$\underline{19.49}_{0.32}$	$\mathbf{21.55}_{0.21}$	$\underline{23.36}_{0.45}$	$19.90_{0.38}$	$\mathbf{24.89}_{0.43}$
	AUROC	$\underline{74.69}_{0.07}$	$72.31_{0.21}$	$\mathbf{74.90}_{0.06}$	$\underline{73.95}_{0.14}$	$71.50_{0.04}$	$\mathbf{75.15}_{0.11}$	$\underline{74.93}_{0.22}$	$73.48_{0.11}$	$\mathbf{75.56}_{0.17}$

Table 5.6: Ablation study of our multimodal fusion model.

	48-IHM		24-PHE	
	F1	AUPR	F1	AUROC
Ours	$\mathbf{56.45}_{1.30}$	$\mathbf{60.23}_{1.54}$	$\mathbf{54.84}_{0.31}$	$\mathbf{86.06}_{0.06}$
:w/o UTDE				
w Imputation	$54.59_{0.91}$	$56.80_{0.54}$	$54.46_{0.17}$	$85.98_{0.02}$
w mTAND^{ts}	$54.89_{1.09}$	$59.11_{1.21}$	$54.07_{0.51}$	$85.92_{0.12}$
:w/o mTAND^{txt}	$51.14_{1.79}$	$57.81_{0.76}$	$53.33_{0.62}$	$85.60_{0.06}$

TDE approach. This result further indicates that UTDE can maintain optimal performance for predictions by integrating MISTS embeddings from different TDE approaches.

Does tackling irregularity in clinical notes improve performance in multimodal fusion? We remove mTAND^{txt} and directly fuse a series of clinical notes representations with UTDE representations. The results are shown in the last row in Table 5.6. Performance drops when the fusion model ignores irregularity in clinical notes, showing the importance of tackling irregularity in clinical notes for medical predictions.

Does the length of clinical notes affect results in multimodal fusion? Clinical notes are often lengthy and contain valuable patient information. A longer encoded

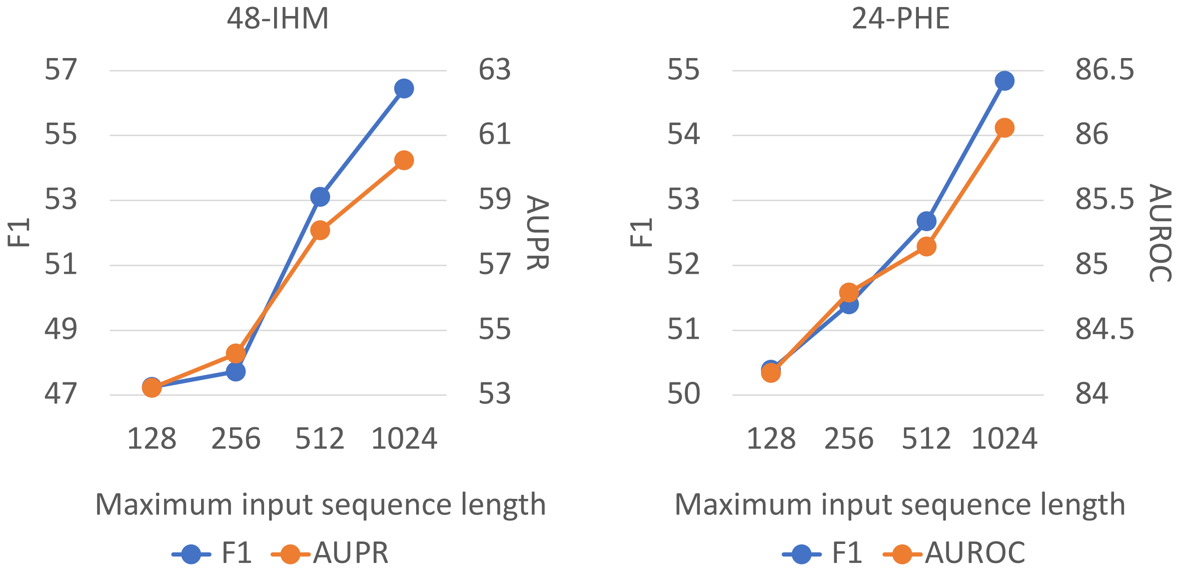


Figure 5.4: Performance of fusion models along with different maximum input sequence lengths.

clinical note brings more expressive power. We adjust our fusion model by encoding clinical notes with Bio-Clinical BERT [67] with maximum input sequence lengths of 128, 256, and 512, and Clinical-Longformer [192], with a maximum input sequence length of 1024, respectively. Figure 5.4 shows improvement in performance as maximum input sequence length increases in both tasks across various evaluation metrics, highlighting the value of clinical notes and the importance of modeling long-term dependency in text in the multimodal fusion scenario.

5.5 Conclusion

In this chapter, we propose a unified system to fully model irregularity in multimodal EHRs for medical predictions. We first tackle irregularity in time series via a gating mechanism and long sequential clinical notes via a time attention mechanism separately, and effectively integrate irregularity into multimodal representations by an interleaved fusion strategy. We hope that our work will encourage further explorations of tackling irregularity in both single modality and multimodal scenarios.

Chapter 6

Conclusion and Future Work

In this thesis, I detail works I conducted during my Ph.D. studies aimed at building deep learning models to enhance diverse medical applications. My contributions spanned numerical measurement modeling, NLP, and multimodal learning. In this concluding chapter, I will summarize the key findings and outline future research directions for each area of focus.

6.1 Numerical Measurement Modeling

In [20], we focus on the training data scarcity setting and propose utilizing domain adaptation for modeling medical clinical measurements in EHRs to improve medical predictions. We comprehensively study three different domain adaptation methods: fine-tuning, adversarial learning, and contrastive learning, and investigate their effectiveness in improving the performance of mortality prediction for trauma patients within this data scarcity setting. The results show that learning domain-invariant information from large public EHRs via different domain adaptation methods can improve the medical prediction performance within small hospitals or sub-medical domains with data scarcity.

In addition, aggregating the distinct features in source and target domains in domain adaptation frameworks significantly improves target domain performances compared to using overlapping features of the two domains. This encourages potential future work to fully utilize data information on such data scarcity settings:

- Application to diverse medical conditions: Explore the application of domain adaptation methodologies beyond trauma patient mortality prediction to other medical conditions and healthcare settings. This aims to validate the adaptability and effectiveness of these techniques across various clinical prediction tasks with more diverse data sources, potentially revolutionizing outcome prediction in different medical domains.
- Advanced encoding techniques: Innovate more sophisticated encoding strategies to address complex feature disparities in healthcare data, including temporal patient data, unstructured clinical notes and, multi-modal data sources, to improve the generalizability of leveraging domain adaptation methods in scenarios of data scarcity with feature disparity.

6.2 NLP

In [21], we propose a novel method to improve downstream tasks under privacy-restricted scenarios by leveraging medical knowledge in the powerful LLMs while alleviating the privacy concerns of private medical data. We assert the importance of the privacy-restricted scenario in the medical domain. Especially, LLMs are becoming more powerful, and their applications are becoming more common through simple API calls. However, there is a concern that these models may not effectively address data privacy issues. We believe the following directions are worth pursuing:

- Building generally applicable medical SLM models for various medical tasks by distilling medical knowledge from strong LLM teachers under privacy-restricted scenarios.
- Although recent studies [7] have shown the medical capacity of LLMs, the hallucination of medical knowledge generation still raises concerns. Improving the quality of medical data generated by LLMs, which is highly correlated with the decision-making ability of SLMs, is worth exploring.

In [22], we introduce a semi-automated pipeline for creating a diverse machine-generated medical instruction fine-tuning (IFT) dataset, named *MedInstruct-52k*, aimed at enhancing the effectiveness of medical LLMs. We offer public access to *MedInstruct-52k*, and a series of medical LLMs trained on the dataset *AlpaCare*. In addition, to better evaluate the medical instruction-following ability of existing medical LLMs, we propose a clinician-curated instruction test set, *MedInstruct-test*, for evaluation. Our work encourages potential future work on :

- Investigating data filtering techniques to secure a higher-quality medical dataset for instruction tuning. [130] underscored the significance of training models with high-quality data over merely increasing the dataset size. The challenge of evaluating data quality within the medical domain remains largely unexplored.
- Integrating the internet with LLM as teacher models to generate more reliable and up-to-date medical instruction-response pairs, improving data quality and diversity.

6.3 Multimodal Learning

In [23], we emphasize the significance of addressing irregularities within multimodal EHRs and propose a method to comprehensively modeling the irregularity for improving medical predictions. We believe our findings can encourage further explorations of tackling irregularity in both single modality and multimodal scenarios:

- Investigating training strategies or designing gate mechanisms to reduce the complexity of training could offer valuable insights. The selection of a model that optimally integrates information across various levels in MISTS modality demands more computational resources due to the introduction of submodules and the employment of the most effective integration method during inference.
- Further exploration of alternative multimodal fusion strategies by both considering the irregularity and improving the effectiveness of model training is encouraged.

Appendix A

Enhancing Small Medical Learners with Privacy-preserving Contextual Prompting

Limitations

While our work enhances SLM performance by using keyword representations of raw data, it only mitigates but does not eliminate privacy concerns. Given that FTC is based on GPT3.5, the medical knowledge it generates may be inaccurate or biased, which can impact SLM performance. Moreover, the inference time for LLMs may be slower than that of SLMs, leading to longer overall inference times compared to models solely reliant on local SLMs. Furthermore, due to the training cut-off time, the medical knowledge in LLM could be outdated, potentially hindering medical decision-making. We aim to integrate LLM with the internet and knowledge graph in future work to generate more reliable medical knowledge for enhancing SLM decision-making. These issues underscore the need for further research on the use of LLMs in privacy-restricted medical scenarios.

A.1 Data and code

Our codes and generated data are public at:https://github.com/XZhang97666/PrivacyBoost-SLM.

A.2 SLM implementation and training details

We implement both SFT and FTC based on huggingface transformers [200], and train on NVIDIA A40-48GB GPUs. For all datasets, we utilize AdamW [201] as optimizer. For MedQA and HEADQA, we set learning rates of 5×10^{-5}, 5×10^{-5}, and 2×10^{-6} for BioLinkBERT-Base, BioLinkBERT-Large, and BioMedLM in both FTC and SFT settings. For MedMCQA, we set learning rates of 2×10^{-5}, 2×10^{-5}, and 2×10^{-6} for BioLinkBERT-Base, BioLinkBERT-Large, and BioMedLM in both FTC and SFT settings. For BioLinkBERT-Base and BioLinkBERT-Large, we limit training to 100 epochs with a 200-step warm-up and apply early stopping after 5 epochs without validation improvement. Batch sizes are 8 for few-shot and full-training scenarios across all datasets. For BioMedLM, we set the training epochs to 10 for all datasets. We run experiments with three random seeds {0, 1, 2} and report mean results and standard deviations.

A.3 Additional experimental results

Development sets results of MedQA and HeadQA. We report development sets results of MedQA and HeadQA in Table A.1 and A.2, and Figure A.1.

Table A.1: Results (%) on development sets of MedQA and HEADQA between LLM, SFT, and FTC under different training sizes.

	MedQA				**HEADQA**			
	100	200	500	full	100	200	500	full
LLM	38.30				47.60			
SFT	$33.28_{0.85}$	$34.01_{1.00}$	$34.20_{0.96}$	$42.21_{0.91}$	$36.54_{0.30}$	$39.00_{0.61}$	$34.88_{2.00}$	$41.48_{0.48}$
FTC	$\mathbf{43.66}_{1.26}$	$\mathbf{45.70}_{0.10}$	$\mathbf{45.62}_{0.70}$	$\mathbf{50.73}_{0.35}$	$\mathbf{55.03}_{1.10}$	$\mathbf{56.18}_{0.76}$	$\mathbf{57.45}_{0.53}$	$\mathbf{60.21}_{1.47}$

Table A.2: Results comparison of FTC and FTCR in the full-training setting on development sets of MedQA and HEADQA.

	MedQA	**HEADQA**
FTCR	$48.56_{0.62}$	$58.17_{0.78}$
FTC	$\mathbf{50.73}_{0.35}$	$\mathbf{61.35}_{0.16}$

A.4 Alternative measurement for privacy budget

We further conducted BPC measurement experiments on the training, validation, and test sets of the MedQA dataset, respectively, to demonstrate the effectiveness of using keywords to represent raw medical data while keeping privacy. To evaluate the BPC of raw data and keywords, we separately input keywords with various proportions and raw data into BioMedLM across different sets. Specifically, for each list of keywords, we form a sentence of keywords by concatenating the list of keywords and separating each pair of keywords with an empty space. Subsequently, we calculate the corresponding BPC values to assess the outcomes. The results are shown in Table A.3.

Figure A.1: Accuracy comparison (%) of ablation studies on development sets of MedQA and HEADQA. The upper part of the table examines the effect of different context components on SLM training, while the lower part investigates the impact of relationships within the context.

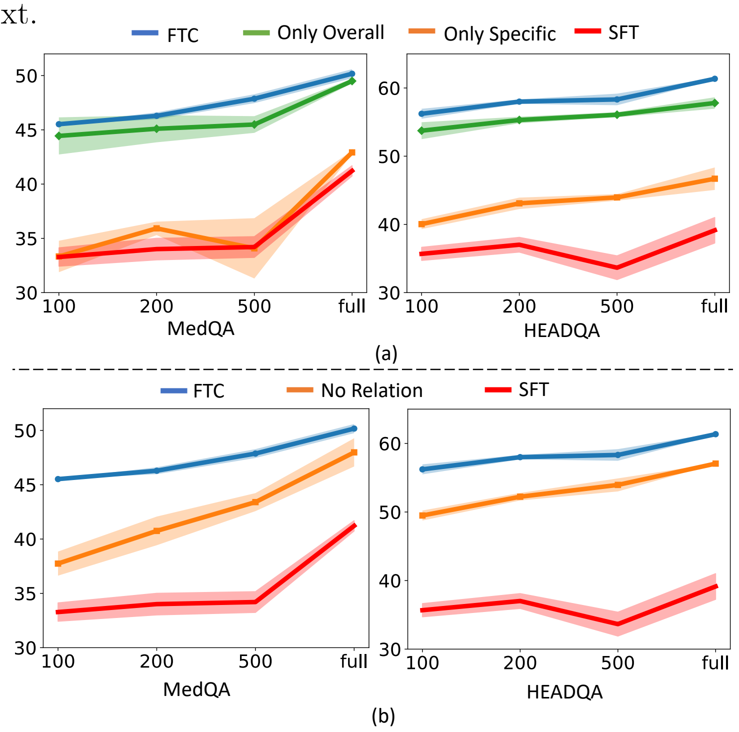


Table A.3: Comparison of BPC values for raw data and concatenated keywords across different data splits of the MedQA dataset

	Raw Data	Keywords			
	-	25%	50%	75%	100%
Training	12.41	10.37	10.04	9.91	9.87
Validation	12.44	10.31	10.01	9.92	9.89
Test	12.44	10.32	10.06	9.94	9.91

The BPC values of keywords consistently exhibit lower values than those of the raw data. A higher BPC is consistently observed across various subsets of the MedQA dataset when compared to the approach of inputting keywords into BioMedLM. This comparison implies that, on average, raw data holds a greater level of uncertainty in contrast to the utilization of keywords. This disparity could be attributed to the fact that raw data encompasses a larger amount of medical-unrelated information, which includes privacy-related data, despite its comprehensive information coverage.

A decrease in the proportion of keywords results in an increase in BPC. Decreasing the number of keywords fed into the BioMedLM leads to an increase in BPC. This indicates that a higher volume of keywords contributes to a more meaningful representation of medical information in raw data, subsequently enhancing the performance of both LLM and FTC.

Keywords representation obtains the lowest BPC compared to random words and random span. We further evaluate the BPC of different raw data representation methods (random span, random words and keywords) while maintaining the privacy budget the same. Specifically, we feed different data representation content into BioMedLM on the training, validation and test sets of the MedQA dataset, and calculate the corresponding BPC values respectively. The results are presented in Table A.4.

Table A.4: BPC values different raw data representation methods.

Representation Methods	Keywords	Random words	Random span
Training	9.87	10.04	12.25
Validation	9.89	10.05	12.22
Test	9.91	10.04	12.22

Keyword representation consistently obtains the lowest BPC compared to the other two data representation methods, providing a more effective representation of medical knowledge within the same privacy budget. Interestingly, the order of BPC performance aligns with the performance of LLM prompting and SLM fine-tuning in Table 7 of our work. The method with a lower BPC value achieves better performance in both LLM prompting and SLM fine-tuning. This further highlights the importance of raw data representation in generating high-quality context and effective SLM training.

A.5 General domain experimental setup

We perform experiments on two commonsense datasets to demonstrate the broad applicability of our approach.

Datasets. 1. CommonsenseQA [124] is a multi-choice question-answering dataset featuring 5 options per question, requiring commonsense reasoning. The dataset is split into 9741/1221/1140 instances for training, development, and test sets, respectively. As the test set is not publicly accessible, we follow previous work [63] and report results on the development set. **2. OpenbookQA** is a 4-option multi-choice question-answering dataset that demands open book facts, broad common knowledge, and multi-hop reasoning [125]. The dataset is split into 4957/500/500 instances for training, development, and test sets, respectively. We report results on the test set.

Context generation from LLM. We utilize the grounded entities from [100] as keywords to query the GPT-3.5 *gpt-3.5-turbo* engine, generating context through a greedy decoding process (by setting the temperature to 0). We employ the same demonstration format as in the medical domain and create a privacy-restricted context in accordance with the in-context learning paradigm. For each dataset, we supply seven-shot handcrafted examples.

SLM training. We employ T5-base [71] as the SLM backbone for both SFT and FTC. In FTC, we use Fusion-in-Decoder [88] to incorporate context information for decision-making. Specifically, each context related to a candidate answer is concatenated with the question and all candidate answers, then processed independently by the encoder. The concatenated representations of all contexts are subsequently fed into the decoder to generate predictions. In both datasets, we utilize AdamW [201] with a learning rate of 5×10^{-5} for both SFT and FTC. We limit training to 100 epochs with a 200-step warm-up and apply early stopping after 5 epochs without validation improvement. The batch size is set to 8 for both datasets.

A.6 Prompt details

In this section, we present examples of prompts for both medical and general domains. The context of each example consists of three parts: (1) An overall context, which provides high-level information derived from the extracted keywords and candidate answers (red); (2) A specific context, which focuses on the knowledge associated with a candidate answer (blue) and its relation to the overall context (green); and (3) A preliminary decision, which draws a conclusion based on the contexts provided earlier (orange).

Prompts for medical datasets. Three clinicians were involved in the prompt design and writing. We held 5 meetings with clinicians to discuss the prompt design and iterate 4 versions. The final version of context needs around 10 minutes to write per context. Our medical prompts on *MedQA* and *MedMCQA* are based on [6], *HEADQA* is based on Wikipedia and written and verified by clinicians. Here we provide the prompts that we used in our experiments.

Table A.5: Prompts for MedQA

Question Keywords: male, marathon runner, office, complaint, right-sided rib pain, Physical examination, normal heart, lung findings, exhalation, dysfunction, ribs 4-5, right, muscles, muscle groups, dysfunction, direct method

Candidate Answers: (a) anterior scalene (b) latissimus dorsi (c) pectoralis minor (d) quadratus lumborum

Continued on next page

Table A.5 – *Continued from previous page*

Context: Normal heart and lung findings on a physical exam coupled with evidence of exhalation dysfunction in ribs 4-5 on the right suggest a musculoskeletal cause of exertional chest pain.
(a): The anterior scalene muscle attaches to the first rib. It is not associated with exhalation dysfunction in ribs 4-5.
(b): The latissimus dorsi muscle attaches to ribs 9 and 10. It is not associated with exhalation dysfunction in ribs 4-5.
(c): The pectoralis minor muscle is attached to ribs 3, 4, and 5. Dysfunction in the fourth and fifth ribs can be caused by issues with the pectoralis minor muscle due to its attachment to these ribs. It is associated with exhalation dysfunction in ribs 4-5.
(d): Quadratus lumborum muscle attaches to ribs 11 and 12. It is not associated with exhalation dysfunction in ribs 4-5.
Therefore, the answer is (c).

Question Keywords: male, office, low back pain, denies, any, trauma, says, truck, day, job, Examination, patient, prone position, deep sacral, left, posterior inferior lateral angle, right, lumbosacral junction, springs, freely, compression, diagnosis
Candidate Answers: (a) left-on-left sacral torsion (b) left-on-right sacral torsion (c) right unilateral sacral flexion (d) right-on-right sacral torsion

Context: The physical exam shows the deep sacral sulcus on the left, a posterior inferior lateral angle on the right and normal spring test.
(a): This condition is characterized by the deep sacral sulcus on the right, a posterior inferior lateral angle on the left and normal spring test. It is not consistent with the findings from the physical exam.
(b): The left-on-right sacral torsion would be indicated by a deep sacral sulcus on the right, a posterior inferior lateral angle on the left, and a positive spring test. It is not consistent with the findings from the physical exam.
(c): This condition is characterized by a posterior inferior lateral angle on the right, a deep sacral sulcus on the right, and an absence of normal spring test. It is not consistent with the findings from the physical exam.
(d): This condition is characterized by a deep sacral sulcus on the left, a posterior inferior lateral angle on the right and normal spring test. It is consistent with the findings from the physical exam.
Therefore, the answer is (d).

Continued on next page

Table A.5 – *Continued from previous page*

Question Keywords: man, comes, office, nonproductive cough, runny nose, frontal headache, headache, morning, nd, ibuprofen, relief, not, shortness of breath, Medical history, no medications, ibuprofen, pain, Vital signs, temperature, 37.4, °, 99.4, °, pulse, 88/min, 18/min, blood pressure, 120/84, Examination, nares, erythematous, mucous membranes, Examination, throat, erythema, follicular lymphoid hyperplasia, posterior oropharynx, no, cervical adenopathy, Lungs, clear, auscultation, patient's, symptoms
Candidate Answers: (a) Allergic rhinitis (b) Epstein-Barr virus (c) Mycoplasma pneumonia (d) Rhinovirus

Context: Sore throats are common symptoms in multiple upper respiratory viruses.
(a): A non-productive cough is a common symptom in upper respiratory viruses but is not present in allergic rhinitis. It is not the cause of the symptoms.
(b): The absence of shortness of breath indicates mycoplasma is less probable. It is not the cause of the symptoms.
(c): Cervical adenopathy is commonly seen in cases of Epstein Barr virus. The absence of cervical adenopathy indicates Epstein Barr virus is less likely. It is not the cause of the symptoms.
(d): Rhinovirus can cause this patient's symptoms, including sore throat, runny nose and a frontal headache. It is the cause of the symptoms.
Therefore, the answer is (d).

Question Keywords: healthy, woman, comes, physician, 8, months, husband, killed, car crash, decreased, difficulty falling asleep, states, sad, cries, frequently, door lock, five, house, five, pieces, toilet paper, perfectionist, urges, rituals, Pharmacotherapy, neurotransmitters
Candidate Answers: (a) Dopamine (b) Glutamate (c) Norepinephrine (d) Serotonin

Context: The woman is exhibiting symptoms of major depressive episodes, such as difficulty falling asleep, frequent crying, and a persistent feeling of sadness.
(a): Dopamine is a neurotransmitter that increases positive emotions. It is implicated in many disease processes, including Parkinson's and ADHD, and is targeted by antipsychotic medications but not used as a sleep aid. It is not a treatment for the patient's symptoms.
(b): Glutamate is a neurotransmitter that is associated with multiple neurological disorders including epilepsy, stroke, and autism. It is not a treatment for the patient's symptoms.
(c): Norepinephrine is a catecholamine with adrenergic properties. It is not a treatment for the patient's symptoms.

Continued on next page

Table A.5 – *Continued from previous page*

(d): Serotonin is a neurotransmitter which is the target for multiple antidepressants, anxiolytics, and antipsychotics. It could be a treatment to address the patient's symptoms of depression and anxiety.
Therefore, the answer is (d).

Question Keywords: man, comes, office, preoperative, evaluation, adrenalectomy, scheduled, 2, weeks, One, month, care, emergency department, pain, right flank, motor vehicle collision, blood pressure, 160/100, mm Hg and CT scan, abdomen, incidental, left adrenal mass, laboratory studies, complete blood count, serum electrolyte concentrations, liver function tests, reference ranges, patient, healthy, elevated blood pressure, no medications, follow-up visit, office 2, weeks, disclosed, elevated, urinary normetanephrine, metanephrine, plasma, concentrations, patient, surgeon, recommended, adrenalectomy, vital signs, temperature, 36.6, 97.9, pulse, 100/min, 14/min, blood pressure, 170/95, Physical examination, no significant, findings, preoperative, preparation, treatment
Candidate Answers: (a) Labetalol (b) A loading dose of potassium chloride (c) Nifedipine (d) Phenoxybenzamine

Context: The patient is being evaluated for adrenalectomy due to a large left adrenal mass, which is likely causing elevated blood pressure as a symptom of pheochromocytoma. Elevated urinary normetanephrines confirm the diagnosis.
(a): This beta-blocker works by blocking the effects of adrenaline and other stress hormones on the heart and blood vessels. It is not a treatment for pheochromocytoma.
(b): The use of a potassium chloride loading dose is a treatment specifically for hypokalemia, which is a condition where there are abnormally low levels of potassium in the blood. It is not a treatment for pheochromocytoma.
(c): This drug is commonly prescribed to treat high blood pressure and angina. It can also help relieve symptoms of Raynaud's phenomenon. It is not a treatment for pheochromocytoma.
(d): This medication is used as a preoperative preparation treatment to block alpha-adrenergic receptors in the body and it effectively treats hypertension caused by pheochromocytoma. It is a treatment for pheochromocytoma.
Therefore, the answer is (d).

Table A.6: Prompts for HEADQA

Question Keywords: autosomal dominant trait

Continued on next page

Table A.6 – *Continued from previous page*

Candidate Answers: (a) The trait appears more frequently in males. (b) The unaffected people do not transmit the trait. (c) The trait tends to skip generations. (d) The affected people have both affected parents. (e) The trait tends to appear in the progeny of related parents.

Context: Autosomal dominant inheritance is a mode of genetic transmission in which a trait or condition can be passed down from parent to child. One copy of a mutated gene from one parent can cause the genetic condition. For example, let 'A' represent the affected allele and 'a' represent the unaffected allele. An affected person may have the genotype AA or Aa, while an unaffected person has the genotype aa. Consequently, an individual with genotype AA has a 100% chance of passing on the affected allele, and someone with genotype Aa has a 50% chance of doing so.
(a): Autosomal dominant inheritance is not influenced by an individual's sex, as it is not sex-dependent. The expression of the trait occurs regardless of gender. It is not a characteristic of autosomal dominant inheritance.
(b): Unaffected individuals do not have the mutated gene and therefore cannot transmit the trait. It is a characteristic of autosomal dominant inheritance.
(c): Autosomal dominant traits can be passed down through multiple generations. Since the affected allele is dominant, an individual will express the trait as long as they inherit the affected gene. It is not a characteristic of autosomal dominant inheritance.
(d): Only one affected parent is needed to transmit on the autosomal dominant trait to their child. It is not a characteristic of autosomal dominant inheritance.
(e):A dominant gene can appear in any progeny, regardless of the parent. It is not a characteristic of autosomal dominant inheritance.
Therefore, the answer is (b).

Question Keywords: caring, patient, supraglottic laryngectomy
Candidate Answers: (a) He has lost the ability to speak by extirpation of the true vocal cords. (b) The tracheostomy they have performed will be permanent. (c) You have a risk of bronchoaspiration due to difficulty swallowing. (d) You may have constipation due to cervical dissection. (e) A portion of the larynx has been removed along with a vocal cord.

Context: Supraglottic laryngectomy or horizontal partial laryngectomy is an operation to remove the epiglottis, false vocal cords, and superior half of the thyroid cartilage.
(a): Supraglottic laryngectomy removes the false vocal cords, but the true vocal cords are not affected, and the patient's ability to speak should not be significantly impacted. It is not typical to lose the ability to speak by extirpation of the true vocal cords.

Continued on next page

Table A.6 – *Continued from previous page*

(b): If a tracheostomy tube is in place after the procedure, it is typically removed within 24-48 hours of surgery. It is not typical to involve A permanent tracheostomy as a part of the supraglottic laryngectomy process.
(c): Supraglottic laryngectomy results in severe disturbance to the swallowing mechanism by removal of protective layers and sensation. There is an increased risk of bronchoaspiration. It is related to care of patients with supraglottic laryngectomy.
(d): Constipation is not a side effect of supraglottic laryngectomy. It is not related to care of patients with supraglottic laryngectomy.
(e): Supraglottic laryngectomy is an operation to remove the epiglottis, false vocal cords, and superior half of the thyroid cartilage. In this procedure, the true vocal cords are not typically affected, preserving the patient's ability to speak as much as possible. It is not common to remove a portion of the larynx along with a true vocal cord during this procedure.
Therefore, the answer is (c).

Question Keywords: estrogenic treatment, adverse effects, NOT, adverse effect, pharmacological action
Candidate Answers: (a) Edema (b) Breast pain (c) Ovarian cancer (d) Sickness (e) Headaches

Context: Estrogen therapy involves supplementing a patient with estrogen, the primary female sex hormone. Potential side effects include breast tenderness or swelling, edema, nausea, leg cramps, endometrial cancer, and more.
(a): Edema is a potential adverse effect of estrogen therapy. Estrogen and aldosterone both originate from cholesterol, and an excessive amount of estrogen in the body can stimulate aldosterone receptors, leading to water retention in nephrons. This water retention can result in edema. It is a non-adverse effect.
(b): Estrogen promotes ductal growth and fat deposition in the breasts. Excessive estrogen levels can lead to mammary duct hyperplasia, which may result in breast pain. It is not a non-adverse effect.
(c): Ovarian cancer is not known to be an adverse effect of estrogenic treatment. It is a possible choice for a non-adverse effect.
(d): Edema is a possible adverse effect of estrogenic treatment, and swelling in body parts may cause the feeling of sickness. It is not a non-adverse effect.
(e): Headache is a possible adverse effect of estrogenic treatment. It is not a non-adverse effect.
Therefore, the answer is (c).

Question Keywords: cardiac valvular prosthesis, biological, mechanical, implanted, patient, aspects, characteristics, patient, prosthesis, INCORRECT, statement

Continued on next page

Table A.6 – *Continued from previous page*

Candidate Answers: (a) Permanent anticoagulation is necessary in mechanical prostheses. (b) In general, biological prostheses are indicated in young patients, with long life expectancy. (c) Biological prostheses would be indicated in cases that present a formal contraindication for anticoagulation. (d) The rate of structural deterioration of a biological prosthesis is inversely proportional to the age of the subject. (e) Biological prostheses do not require permanent anticoagulation.

Context: Cardiac valvular prostheses (biological or mechanical) are artificial cardiac valves implanted into a patient's heart. Mechanical valves may last a lifetime, but they come with an increased risk of blood clots, necessitating the use of blood thinners such as warfarin. In contrast, biological valves, which are made from pig or cow tissue, do not increase the risk of bleeding or clotting but tend to wear out sooner.
(a): Mechanical valves increase the risk of blood clotting. It is not an incorrect statement.
(b): The latest revisions of the ESC/EACTS guidelines suggest that bioprostheses are acceptable in patients aged between 60 and 65 years at the time of surgery. The reoperation rate for structural valve degeneration (SVD) of bioprostheses occurred exclusively among patients younger than 56 years. Young patients are not typically recommended for a biological prosthesis.
(c): Biological prostheses, which are made from pig or cow tissue, do not increase the risk of either bleeding or clotting but will wear out sooner. It is not an incorrect statement.
(d): The disadvantages of biological heart valves are a smaller valve orifice area and the risk of structural valve degeneration, which may necessitate reoperation. Thus, the younger the patient, the higher risk of structural deterioration. It is not an incorrect statement.
(e): Biological prosthesis do not increase the risk of clotting so do not require permanent anticoagulant. It is not an incorrect statement.
Therefore, the answer is (b).

Question Keywords: connection, automatic, emotional responses, control, behaviors, guiding, behavior, manifestation, emotional responses
Candidate Answers: (a) The angular gyrus of the limbic system. (b) The convolution or lobe of the insula. (c) The prefrontal orbitofrontal or ventromedial cortex. (d) The thalamus (e) The cortex of somatosensory association.

Continued on next page

Table A.6 – *Continued from previous page*

Context: The prefrontal orbitofrontal cortex has multiple functions including mediating context specific responding, encoding contingencies in a flexible manner, encoding value, encoding inferred value, inhibiting responses, learning changes in contingency, emotional appraisal, altering behavior through somatic markers, driving social behavior, and representing state spaces. The orbitofrontal cortex thus plays a key role in emotion, by representing the reward value of the goals for action.
(a): The angular gyrus (AG) is a hub of several networks that are involved in various functions, including attention, self-processing, semantic information processing, emotion regulation, and mentalizing. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
(b): The insula is important for gustatory and sensorimotor processing, risk-reward behavior, autonomics, pain pathways, and auditory and vestibular functioning. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
(c): The prefrontal cortex guides behavior by controlling the manifestation of emotional responses through understanding rewards, encoding values, and driving behaviors. It is the potential correct answer.
(d): The thalamus acts as the body's information relay station. All sensory information (except for olfaction) must be processed through the thalamus before being sent to the cerebral cortex for interpretation. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
(e): The somatosensory cortex is responsible for processing all bodily sensations. These sensations originate from receptors located throughout the body that detect temperature, pain, touch, pressure, and proprioception. It is not the area responsible for connecting automatic emotional responses and controlling complex behaviors.
Therefore, the answer is (c).

Table A.7: Prompts for MedMCQA

Question Keywords: Maximum, increase, prolactin level
Candidate Answers: (a) Risperidone (b) Clozapine (c) Olanzapine (d) Aripiprazole

Continued on next page

Table A.7 – *Continued from previous page*

Context: The four drugs in answer choices are all atypical antipsychotics, which are used to treat psychotic conditions like schizophrenia through blockage of dopamine and serotonin receptors. These drugs block dopamine D2 receptors and serotonin 5-HT2 receptors. Maximum increase in prolactin, or hyperprolactinemia, is one of the side effects of atypical antipsychotics, because dopamine tends to inhibit prolactin release from the anterior pituitary. (a): Risperidone is a type of atypical antipsychotics that block dopamine D2 receptor and serotonin 5-HT2 receptor. It is generally used to treat schizophrenia or disorders with concomitant psychosis. Hyperprolactinemia is one of the most common side effects of risperidone. It is the drug to increase prolactin levels.
(b): Clozapine is used to treat schizophrenia or disorders with concomitant psychosis. Clozapine is associated with side effects such as agranulocytosis, seizures, and myocarditis, but it does not appear to elevate prolactin levels. It is not the drug to increase prolactin levels.
(c): Olanzapine is used to treat schizophrenia or disorders with concomitant psychosis. The side effect of olanzapine does not include hyperprolactinemia. It is not the drug to increase prolactin levels.
(d): Aripiprazole is generally used to treat schizophrenia or disorders with concomitant psychosis. The side effect of olanzapine does not include hyperprolactinemia. It is not the drug to increase prolactin levels.
Therefore, the answer is (a).

Question Keywords: male, complains, severe back pain, inability, left lower limb, Radiographic studies, compression, nerve elements, intervertebral, foramen, vertebrae L5, S1, structure, space-occupying lesion
Candidate Answers: (a) Anulus fibrosus (b) Nucleus pulposus (c) Posterior longitudinal ligament (d) Anterior longitudinal ligament

Context: The male is complained of a severe back pain and inability to move, and radiographic evidence shows the compression of a nerve component. This may suggest a herniated intervertebral disk through a tear in the surrounding annulus fibrosus. The soft, gelatinous nucleus pulposus is forced out through a weakened part of the disk, compressing nerve components of the spinal cord and resulting in back pain and nerve root irritation. This impingement is resulting in paralysis, and should be considered a medical emergency.
(a): Annulus fibrosus is a tough, circular exterior of the intervertebral disc, made up of fibrous connective tissue. It surrounds the soft inner core, the nucleus pulposus. It is not the component that is forced out by the tear.
(b): Nucleus pulposus is the inner core of the vertebral disc. The tear in the annulus fibrosus causes it to be forced out. It could result in compression of the nerve components of the vertebrae.

Continued on next page

Table A.7 – *Continued from previous page*

(c): Posterior longitudinal ligament connects and stabilizes the bones of the spinal column. It runs almost the entire length of the spine, from the 2nd vertebra in the cervical spine (neck) all the way down to the sacrum (end of the spine). This ligament is located adjacent to the spinal cord. It is not easily teared or curved.
(d): Anterior longitudinal ligament is a ligament that runs down the anterior surface of the spine. It traverses all of the vertebral bodies and intervertebral discs on their ventral side. It has a high tensile strength and is resistant to tearing or deformation. It is not easily teared or curved.
Therefore, the answer is (b).

Question Keywords: Neuroendocrine cells, lungs
Candidate Answers: (a) Dendritic cells (b) Type I pneumocytes (c) Type II pneumocytes (d) APUD cells

Context: Neuroendocrine cells are part of the neuroendocrine system. The neuroendocrine cells of the lung make hormones that control the flow of air and blood in the lungs.This may suggest a herniated intervertebral disk through a tear in the surrounding annulus fibrosus. The soft, gelatinous nucleus pulposus is forced out through a weakened part of the disk, compressing nerve components of the spinal cord and resulting in back pain and nerve root irritation. This impingement is resulting in paralysis, and should be considered a medical emergency.
(a): Dendritic cells are a type of antigen-presenting cell in the immune system that act as messengers between the innate and adaptive immune systems. It is not a type of neuroendocrine cell.
(b): Type I pneumocytes are alveolar cells that line the alveolar surface of the lungs and are responsible for gas exchange. It is not a type of neuroendocrine cell.
(c): Type II pneumocytes are alveolar cells that secrete surfactant to reduce alveolar surface tension and prevent alveolar collapse. It is not a type of neuroendocrine cell.
(d): APUD cells are a type of neuroendocrine cell that function through amine precursor uptake and decarboxylation. It is accurate to say that they are a type of neuroendocrine cell.
Therefore, the answer is (d).

Question Keywords: Presence, remote, contamination,water
Candidate Answers: (d) Streptococci (b) Staphalococci (c) Clastridium pertringes (d) Vibrio

Context: Infections that can be spread through water contamination are generally transmitted orally or via fecal matter. (a): Streptococci are spread through direct contact with the nose and throat discharges of an infected individual or with infected skin lesions. Water is not a medium for the spread of streptococci. It is not related water contamination.

Continued on next page

Table A.7 – *Continued from previous page*

(b): Staphylococci is spread by skin contact, like a bite or cut. It is not related water contamination.
(c): Clostridium perfringens are one of the most common causes of food poisoning. They are environmentally stable and specific to contamination by sewage. Their spread is a indicator of water contamination.
(d): Vibrio species are gram-negative bacteria that spread through foodborne infection, but they are highly salt tolerant and unable to survive in fresh water. It is not related water contamination.
Therefore, the answer is (c).

Question Keywords: True, Mooren's ulcer, 2007, 2013
Candidate Answers: (a) Painless condition (b) Affects cornea (c) Sudden loss of vision (d) Bilateral in majority of cases

Context: Mooren's ulcer is characterized by painful peripheral corneal ulceration of unknown etiology. The disease generally begins with intense limbal inflammation and swelling in the episclera and conjunctiva. Patients often experience severe pain, photophobia, and tearing along with a red inflamed eye.
(a): Mooren's ulcer is a painful ulceration of the eye. It is not the truth of Mooren's ulcer.
(b): Mooren's ulcer is characterized by painful peripheral corneal ulceration of unknown etiology. It is the truth of Mooren's ulcer.
(c): The symptoms of Mooren's ulcer do not include sudden loss of vision. It is not the truth of Mooren's ulcer.
(d): About one third of Mooren's ulcer cases present bilaterally. The proportion is less than half. It is not the majority of cases.
Therefore, the answer is (b).

Prompts for general domain datasets. Our prompts on *CommonsenseQA* and *OpenbookQA* are based on [63].

Table A.8: Prompts for Commonsense QA

Question Keywords: fountain pen, people, ink, absorb, pen, hand done, extra, use, fountain
Candidate Answers: (a)shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter

Context: Fountain pens need to be filled with ink for writing. Extra ink should be absorbed using special tools.

Continued on next page

Table A.8 – *Continued from previous page*

(a): A fountain pen can be conveniently carried in a shirt pocket. It is not associated with the tool to absorb extra ink from fountain pens.
(b): Calligraphers use fountain pens to create stunning handwriting. It is not associated with the tool to absorb extra ink from fountain pens.
(c): An inkwell serves as a container for the ink used in a fountain pen. It is not associated with the tool to absorb extra ink from fountain pens.
(d): A fountain pen can be kept safely in a desk drawer. It is not associated with the tool to absorb extra ink from fountain pens.
(e): Blotters are designed to absorb excess ink from pens. It is the tool for absorbing extra ink.
Therefore, the answer is (e).

Question Keywords: fox, forest, walk, look, city
Candidate Answers: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook (e) dense forest

Context: Foxes are animals that typically live in forests. They walk from the city to the forest to look for their living place.
(a): Pretty flowers are in forests. It is not a reason for a fox walking into the forest.
(b): Foxes sometimes prey on chickens in hen houses. It is not a reason for a fox to walk into the forest.
(c): Forests are the natural habitat of foxes. Foxes walk from city to forest to look for their natural habitat. (d): Forests and foxes are common subjects in storybooks. It is not a reason for fox walking to the forest.
(e): Dense forest is a type or category of forests characterized by having a high density of trees and vegetation. It is a type of forest.
Therefore, the answer is (c) or (e).

Question Keywords: grape, put, check
Candidate Answers: (a) mouth (b) grocery cart (c) super market (d) fruit basket (e) fruit market

Context: Grapes need to be put into a place for checking out.
(a): Grapes can be eaten by mouth. It is not a place to put grapes for checking out.
(b): Grapes can be brought during grocery shopping and people put groceries into grocery carts before checking out. It could be a potential place to put grape.
(c): Super markets sell grapes. It is not a place to put grapes for checking out.
(d): Fruit markets sell grapes. It is not a place to put grapes for checking out.
(e): Fruit baskets are often used as gifts to hold and present a variety of fresh grapes. It is not a place to put grapes for checking out.
Therefore, the answer is (b).

Continued on next page

Table A.8 – *Continued from previous page*

Question Keywords: drawstring bag, head, woman, bag, drawstring, check, baggage
Candidate Answers: (a) garbage can (b) military (c) jewelry store (d) safe (e) airport
Context: A woman can check baggage such as a drawstring bag at the check-in counter.
(a): A garbage can is a container that is specifically designed to hold and contain trash or waste materials. It is not related to the context.
(b): Military refers to the armed forces of a country, which is responsible for defending the nation and its interests against external threats. It is not a place where a woman can check bags.
(c): Jewelry stores sell jewelry. It is not a typical place to check baggage.
(d): Check baggage could keep the bag safe. A woman can check her drawstring bag to keep the bag safe.
(e): Airport is a place where the woman can check her drawstring bag as baggage at the check-in-counter. It is common to check baggage in airport.
Therefore, the answer is (e).

Question Keywords: cable, entertainment, home, require, equipment
Candidate Answers: (a) radio shack (b) substation (c) television (d) cabinet (e) desk
Context: A cable transmits electricity or information and data to home entertainment equipment that requires electricity.
(a): Radio Shack is a retailer that sells cable. It is not a home entertainment equipment used cable.
(b): Cables are used to transmit electrical energy between substations and other parts of the electrical power system. It is not a home entertainment equipment used cable.
(c): Television is a type of home electric entertainment equipment that requires cable. It is a home entertainment equipment used cable.
(d): Cabinet is a place to store cable. It is not a home entertainment equipment used cable.
(e): Desk with built-in cable management features can help keep cables tidy. It is not a home entertainment equipment used cable.
Therefore, the answer is (c).

Question Keywords: people, populate, might, may, sammy, go
Candidate Answers: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock

Continued on next page

Table A.8 – *Continued from previous page*

Context: People may like to go to places where people populate together. (a): Populated areas are locations where people gather and live in close proximity to each other. It could be a place where people populate together. (b): Deserts are inhospitable environments for people. It is not a place where people populate together. (c): People go to race competitions on the race track. It could be a place where people populate together. (d): Apartments serve as living spaces for people. It is not a place where people populate together. (e): Roadblocks are structures set up to restrict or regulate the movement of people and vehicles. It is not a place where people populate together. Therefore, the answer is (a) or (c).

Question Keywords: highway, maps, replace, street, google, map, highway, gps, service
Candidate Answers: (a) united states (b) mexico (c) countryside (d) atlas (e) oceans

Context: Google Maps and GPS services have replaced traditional physical maps for navigating highways and streets.
(a): People in the United States use Google Maps and GPS services to navigate highways and streets. It is not the tool that GPS replaced with.
(b): People in Mexico use Google Maps and GPS services to navigate highways and streets. It is not the tool that GPS replaced with.
(c): Google Maps and GPS services cover the countryside. It is not the tool that GPS replaced with.
(d): Google Maps and GPS services have replaced traditional physical maps for navigating highways and streets. Atlases are examples of traditional physical maps. It is the tool that GPS replaced with.
(e): Google Maps and GPS services cover the oceans and are commonly used in marine navigation. It is not the tool that GPS replaced with.
Therefore, the answer is (d).

Table A.9: Prompts for Openbook QA

Question Keywords: acid, environment, aquatic, rain, effect, acid rain
Candidate Answers: (a) decrease in plant life (b) increase in fish population (c) increase in plant growth (d) cleaner and clearer water

Context: The acid rain is a type of rain that has an acidic effect due to the presence of acid in the atmosphere. Acid rain is harmful to the environment, especially aquatic life.The acid in the rain can have a negative effect on the water quality of aquatic environments.

Continued on next page

Table A.9 – *Continued from previous page*

(a): Acid rain can have a negative effect on plant life. The acid in the rain can damage plant cells and cause a decrease in plant growth, leading to a decrease in plant life. It is likely to have a decrease in plant life by acid rain.
(b): Acid rain can have a harmful effect on aquatic life, including fish. The acid in the water can make it difficult for fish to breathe and can harm their reproductive systems. It is not likely to have an increase in fish population by acid rain.
(c): As previously mentioned, the acid in the rain can damage plant cells and cause a decrease in plant growth. It is not possible to have an increase in plant growth.
(d): Acid rain can have a harmful effect on water quality, making it more acidic and harmful to aquatic life. It is not possible to have cleaner and clearer water.
Therefore, the answer is (a).

Question Keywords: moon, surface
Candidate Answers: (a) is smooth on the entire surface (b) contains large cavities cause by explosions (c) contains an internal core of cheese (d) is filled with lakes

Context: The moon is a natural satellite that orbits around the Earth. Its surface is covered with dead volcanoes, impact craters, and lava flows, some visible to the unaided stargazer.
(a): The moon has mountains, craters, and other features caused by impacts from meteoroids and asteroids. It is not entirely smooth on the surface.
(b): Impact craters are formed when an asteroid craters, each of which was formed when an asteroid or comet collided with the Moon's surface. The moon's surface contains large cavities caused by explosions from impacts.
(c): The core is largely composed of iron and some nickel. The inner core is a solid mass about 480 km in diameter. It does not contain an internal core of cheese. (d): The moon has lunar maria composed of basalt formed from surface lava flows that later congealed. It is not filled with lakes.
Therefore, the answer is (b).

Question Keywords: car, approach, night
Candidate Answers: (a) the headlights become more intense (b) the headlights recede into the dark (c) the headlights remain at a constant (d) the headlights turn off

Context: Headlights of a car are a source of light. As a car approaches, the source of light becomes closer, and that source will appear brighter.
(a): As the car becomes closer, the distance to the source of light decreases. The headlights become brighter and more intense. This is a possible phenomenon.
(b): If the source does not change and the headlights are closer, the headlights cannot become dimmer. This is not a commonsense relation.

Continued on next page

Table A.9 – *Continued from previous page*

(c): If the distance to the source of light changes, the brightness of headlights will change. It is not able to remain constant.
(d): Turning off the headlights would cause the driver to be driving in complete darkness, which is dangerous and can lead to accidents. It is not a reasonable condition.
Therefore, the answer is (a).

Question Keywords: change, easter, weather change, weather, christmas
Candidate Answers: (a) the air may chill (b) the ground may freeze (c) the plants may die (d) the ground may warm
Context: In the US, Christmas falls in the winter season, while Easter arrives at the beginning of spring.
(a): The air becomes chill as temperature drops. The temperature commonly increases from winter to spring. It is not a likely scenario.
(b): During winter, the ground usually freezes, whereas in spring, it does not. It is not a probable scenario.
(c): Extreme cold or hot weather can cause plants to die. The beginning of spring provides suitable weather conditions for plants to grow. It is not common to have plants die.
(d): As winter transitions into spring, the weather becomes warmer. The temperature of the ground is influenced by the weather.
Therefore, the answer is (d).

Question Keywords: heat, recipe, moisture, good, ocean
Candidate Answers: (a) a violent storm (b) violent sea animals (c) condensation (d) inland storms
Context: The ocean, a vast body of water that covers a large portion of the Earth's surface, serves as a source of heat and moisture.
(a): The heat and moisture present in the ocean can create ideal conditions for a hurricane or typhoon. Hurricane and typhoon are violent storms.
(b): Violent sea animals are not related to heat and moisture in the ocean. It is not a likely choice.
(c): Condensation is the process by which water vapor becomes liquid, which is the reverse of evaporation. This can happen in one of two ways: either the air is cooled to its dew point or it becomes so saturated with water vapor that it cannot hold any more water. It is not likely to occur in hot conditions.
(d): Although heat and moisture can cause inland storms, they are not directly related to the ocean. It is not a likely choice.
Therefore, the answer is (a).

Continued on next page

Table A.9 – *Continued from previous page*

Question Keywords: hummingbird, take
Candidate Answers: (a) bees (b) energy (c) pollen (d) honey

Context: Hummingbirds dip their long bills into flowers to drink nectar to get energy.
(a): Hummingbirds and bees are both attracted to the sweet nectar produced by flowers, but bees extract the nectar from the base of the flowers, while hummingbirds dip their long bills into the flowers to drink the nectar and obtain energy. No relationship can be found.
(b): Hummingbirds obtain energy by getting nectar from flowers through dipping their long bills into the flowers. No relationship can be found.
(c): When hummingbirds drink nectar, they also inadvertently take grains of pollen which stick to their feathers and bills, and get carried to the next flower they visit. No relationship can be found.
(d): Hummingbirds do not produce or consume honey. This fact is unrelated to their method of obtaining energy by drinking nectar from flowers.
Therefore, the answer is None.

Question Keywords: responsible, sun
Candidate Answers: (a) puppies learning new tricks (b) children growing up and getting old (c) flowers wilting in a vase (d) plants sprouting, blooming and wilting

Context: The sun is the source of energy for physical cycles on Earth.
(a): Puppies learning new tricks involves the acquisition and processing of information, which is essential for the puppies to learn and adapt to their environment. It is not directly related to the effect of the sun.
(b): Children grow up and age over time. The sun is not directly responsible for the passage of time itself. It is not directly related to the effect of the sun.
(c): Flowers in a vase become wilting because they are cut from their original source of nutrients and water and are no longer able to receive the essential nourishment they need to stay healthy and vibrant. It is not directly related to the effect of the sun.
(d): Plants need sunlight to photosynthesize and grow, and the sun's heat and light play a crucial role in the process of plant growth and decay. It is the thing that the sun is responsible for.
Therefore, the answer is (d).

Appendix B

AlpaCare: Instruction Fine-tuned Large Language Models for Medical Applications

Limitations

Our approach utilizes 'teacher' LLMs, such as GPT-4 and ChatGPT, to automatically generate medical instruction-response pair datasets, employing these teacher models as medical knowledge bases. However, this could result in hallucinations in the medical knowledge generation. To enhance the generation reliability, we aim to integrate LLMs with the internet and knowledge graphs in future work.

B.1 Medical Task Difficulty Level Scoring System

We introduce a clinician-crafted seed set to generate *MedInstruct-52k* and a free-form medical instruction evaluation set, *MedInstruct-test*. This set spans a medical difficulty scale ranging from 1 to 5, where 1 represents the easiest tasks and 5 indicates the most challenging ones. A clinician assessed the difficulty levels of all instances within both the seed set and *MedInstruct-test* based on the scoring system shown in Table B.1 This system offers a refined dimension for prompting GPT-4 to produce tasks across varied difficulty levels and to evaluate medical proficiency of IFT models.

Table B.1: Scoring system for evaluating the difficulty level of medical tasks.

Score	Description
1	The fact is very basic. The answer becomes apparent immediately after reading the question, or it can be easily found through a direct internet search.
2	The fact is simple but may require a slight application of real-world knowledge, rephrasing, or extending the information to find the answer.
3	This involves facts that require more real-world application, dealing with practical and somewhat complicated situations. It may require more complex paraphrasing and/or communication skills, such as emotional support, psychological evaluations, and ethical considerations. The tested knowledge in this category can be quite challenging.
4	This level involves complicated medical facts. Answering questions at this level may require multi-step thinking processes. The questions might be lengthy and detailed, necessitating simplification for a clearer answer. This category might include most USMLE questions. It may also require a demonstration of enhanced emotional support, psychological evaluations, and ethical considerations. Questions might be based on vague symptom descriptions, making the diagnosis challenging, or involve recent advancements, publications, or current global health issues like pandemics.
5	This category involves complex medical knowledge applied to real-world, intricate situations. The questions are detailed and lengthy, often requiring simplification and multi-step thinking to answer. Some questions might be based on actual medical cases with challenging diagnoses and treatments. The symptom descriptions might be highly vague. Questions could also involve new technologies, recent publications, or current pandemics, requiring decision-making or choosing the best available option. Instructions might also necessitate the demonstration of humane care.

B.2 Prompt details for *MedInstruct-52k* generation

Here we provide prompts that we use for query GPT-4 and ChatGPT for task and response generation.

Table B.2: Task generation prompt

Your objective is to generate diverse medical-related tasks.

Here are the requirements:
1. Ensure that all tasks are related to the medical domain.
2. Craft tasks that encompass varied points of view, e.g. experts, students and patients, etc.
3. Maximize the range of task topics, e.g. diseases, treatment, diagnoses, epidemiology, pharmacology, pathophysiology, anatomy, genetics, medical education, etc.
4. Introduce different task formats, e.g. text generation, open Q&A, chat, rewrites, summarizations, classifications, USMLE style Q&A, multiple-choice Q&A, single-hop reasoning and multiple-hop reasoning etc.
5. All the formats specified in point 4 MUST be represented in the task you generate.
6. Create tasks with medical difficulty levels from 1 to 5, with 1 being the easiest and 5 the hardest.
7. Use diverse language in the instructions. For instance, combine questions with imperative forms.
8. Some instructions might require specific inputs. If an input is not necessary, such as with general instructions like ”What are the side effects of COVID-19?”, use ”¡noinput¿” in the input field.
9. When provided, inputs must range between 50 to 200 words and offer detailed medical context , e.g. symptom descriptions, radiology reports, clinical notes, and exam questions, etc.
10. Generate a detailed and comprehensive input instead ask user-provided input.
11. Ensure USMLE style Q&A and multiple-choice Q&A tasks have both questions and choices in input, and the question context should be detailed.
12. The USMLE-style question length must exceed 50 words.
13. Match instruction and input to the task’s perspective. Patient perspectives should be simple and in the first person, while clinician views should have professional terminology.
14. Ensure the lengths of inputs for different tasks are notably distinct.
15. Each task should adhere to the following structure: ’Type: \n, Topic: \n, View: \n, Difficulty: \n, Instruction: \n, Input: ’. Start each new task with ’###’.

List of 15 tasks:
Seed task 1
Seed Task 2
Seed Task 3

B.3 Training hyperparameter details

We report the hyperparameter setup for model tuning of *AlpaCare* 7B and 13B models. The details are shown in Table B.4.

Table B.3: Output generation prompt

You are a medical expert tasked with answering various medical questions. You MUST generate your response based on the requirements.

Here are the requirements:
1. For multiple-choice, calculation, and classification problems, you can generate intermediate thinking steps if necessary; otherwise, provide the final answer directly.
2. All the intermediate thinking steps must be generated before final answer.
3. For multiple-choice questions, you MUST generate the answer choice in the following format: 'The answer is (your choice).' For example:
'Choose the correct answer. Where in your body will you find the tibia bone? A) Arm B) Foot C) Skull D) Leg
The tibia bone is one of the two bones in the lower leg, the other being the fibula. The answer is D) Leg.'
4. For other types of questions, except multiple-choice, do not use the format mentioned in point 3.

task instruction
task input (if exist)

Table B.4: ***AlpaCare*** **hyperparameter setup.**

Model Size	Data Size	GPUs	Epoch	LR	Batch Size
7B	52k	4 40G A100	3	2e-5	128
13B	52k	4 80G A100	5	1e-5	128

B.4 Additional experimental results

B.4.1 General domain free-form instruction evaluation

We show the detailed score of each reference model for general domain free-form instruction evaluation in Table B.5.

B.4.2 More analysis in general domain performance

Compared to Alpaca [127], AlpaCare achieves better results in the general domain. This improvement is likely due to the intensive knowledge and reasoning embedded in the medical dataset [8, 202]. For example, in the BBH results, the top three categories where AlpaCare outperforms Alpaca are 'dyck_languages', 'movie_recommendation', and 'navigate', which requires strong knowledge and reasoning abilities. To further support these findings, we conducted a knowledge-intensive commonsense evaluation using StrategyQA [203] and an additional reasoning benchmark evaluation using DROP [204] to compare

Table B.5: **Comparison on general domain free-form instruction evaluation.** A performance comparison between *AlpaCare* and IT baselines on AlpacaFarm on 4 distinct reference models: Text-davinci-003, GPT-3.5-turbo, GPT-4 and Claude-2. 'AVG' represents the mean performance score across all referenced models.

	AlpacaFarm				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	38.7	20.6	14.5	16.9	22.7
ChatDoctor	37.4	20.3	13.1	14.0	21.2
Medalpaca	38.2	24.4	20.6	20.1	25.8
PMC	15.8	2.6	13.3	1.6	8.3
Baize-H	29.9	16.9	12.7	13.7	18.3
AlpaCare	**56.4**	**38.6**	**34.2**	**33.7**	**40.7**

Alpaca and AlpaCare with LLaMA [16]-7B as backbone, following the methodologies in [150] and [155], respectively. The Table B.6 below presents the results.

Model	**StrategyQA**	**DROP**
Alpaca	57.80	23.68
AlpaCare	58.02	24.96

Table B.6: Performance of Alpaca and AlpaCare on StrategyQA and DROP datasets.

These results reinforce that AlpaCare's enhanced performance is not limited to the medical domain but also extends to broader general domain tasks, thereby confirming its superior generalizability.

B.4.3 Ablation study

We show the detailed score of 4 reference models for medical free-form instruction evaluation on 13B instruction-tuned models in Table B.7.

Table B.7: **Result comparison of 4 reference models on 13B instruction-tuned models.**

	iCliniq				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	46.7	37.0	19.6	21.7	31.3
Medalpaca	8.1	4.4	1.0	2.0	3.9
PMC	40.6	29.0	14.3	17.5	25.4
AlpaCare	**66.7**	**51.2**	**48.2**	**50.2**	**54.4**
	MedInstruct				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	39.8	22.5	27.1	18.1	26.9
Medalpaca	0.2	0	0	0	0.1
PMC	44.9	31.9	32.8	29.2	34.7
AlpaCare	**71.3**	**49.1**	**49.8**	**47.7**	**54.5**

We show the detailed score of 4 reference models for medical free-form instruction evaluation on different backbones in Table B.8.

Table B.8: **Results on different LLM backbone across 4 reference models by using gpt-3.5-tubro as the judge.** Comparing the performance of AlpaCare and Alpaca using different LLM backbones, with 4 distinct reference models.

		iCliniq				
		Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
LLaMA	Alpaca	38.8	30.4	12.8	15.6	24.4
	AlpaCare	**66.6**	**50.6**	**47.4**	**49.7**	**53.6**
LLaMA-2	Alpaca	45.8	36.3	18.2	20.8	30.3
	AlpaCare	**66.5**	**50.4**	**47.8**	**50**	**53.7**
LLaMA-3	Alpaca	42.3	28.6	26.4	10.0	26.8
	AlpaCare	**77.6**	**53.9**	**46.3**	**49.7**	**56.9**
		MedInstruct				
		Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
LLaMA	Alpaca	35.0	20.6	21.5	15.6	24.4
	AlpaCare	**67.6**	**48.8**	**47.4**	**49.7**	**53.5**
LLaMA-2	Alpaca	39.6	22.7	26.4	18.5	26.8
	AlpaCare	**70.6**	**48.8**	**50.0**	**48.4**	**54.2**
LLaMA-3	Alpaca	38.4	16.9	14.6	13.0	20.7
	AlpaCare	**78.5**	**50.0**	**51.4**	**46.5**	**56.6**

We show the detailed score of 4 reference models for medical free-form instruction evaluation by using Claude-2 as the judge in TableB.9.

Table B.9: **Medical free-from instruction evaluation results by using Claude-2 as judge.**

	iCliniq				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	40.7	33.4	18.5	14.2	26.7
ChatDoctor	24.5	23.8	11.3	9.9	17.4
Medalpaca	38.3	37.1	15.8	15.4	26.7
PMC	0	0	1.9	3.1	1.3
Baize-H	46.9	35.6	12.1	7.5	25.5
AlpaCare	**64.5**	**46.8**	**26.9**	**17.1**	**38.8**
	MedInstruct				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	44.4	19.9	19.2	13.4	23.5
ChatDoctor	34.8	17.5	14.1	10.3	21.7
Medalpaca	41.4	19.3	18.0	12.5	23.1
PMC	2.8	1.5	0.1	0.7	1.8
Baize-H	40.1	25.1	21.8	15.2	19.8
AlpaCare	**76.4**	**42.6**	**42.9**	**31.6**	**31.5**

Appendix C

Improving Medical Predictions by Irregular Multimodal Electronic Health Records Modeling

C.1 Computation resource of UTDE

We set the integration level of UTDE as a hyperparameter and use validation sets to search the level on which to operate, which requires more computation resources than a model with only a single TDE method. Specifically, each time series experiment run takes less than 10 minutes with a 1 RTX-3090. The integrating operation is a hyperparameter with three levels. In this case, the total running time of UTDE will be less than 30 minutes across different integrating levels, which is affordable.

C.2 Data prepossessing

Table C.1: Links for data generation and preprocessing used in experiments

	Links
MIMIC III	https://mimic.physionet.org/
Time series features selection and extraction	https://github.com/YerevaNN/mimic3-benchmarks
clinical notes extraction	https://github.com/kaggarwal/ClinicalNotesICU

The dataset link, and time series and clinical notes extraction used in the experiments are listed in Table C.1. For time series, we follow [161] to select numerical time series features and extract time series within 48/24 hours and split the training, validation and test sets for each task. We rescale each numerical feature to be between 0 and 1. We also rescale the time to be in [0, 1] for all tasks. The clinical notes within 48/24 hours are extracted by following [180]. For patients with more than 5 clinical notes, we utilize the last 5 clinical notes preceding the prediction time, due to computational resource limitations. We hypothesize that a note is taken closer to prediction time, the more influential it is.

Note that our early-stage phenotype classification is a brand new task compared to phenotype classification in [161], which uses the whole time series of an ICU stay. Our belief is that acute care conditions should occur during the ICU stay, and the earlier they can be predicted, the more valuable they become. Therefore, we focus on extracting the first 24 hours of data for phenotype classification, rather than using the entire admission data. This approach is also supported by [19] in their research on early-stage diagnoses prediction.

C.3 Baselines

C.3.1 MISTS baselines

Imputation: Discretizes MISTS to hourly intervals and obtains imputation embeedings, as described in Section 5.3.2.
IP-Net [163]: Employs a semi-parametric RBF interpolation network to obtain interpolation representations and a prediction network for prediction. We utilize a Transformer encoder as the prediction network.
mTAND [164]: Presents a multi-time attention module to obtain an interpolation representation, as described in Section 5.3.2. We adopt a Transformer as the time series encoder to predict downstream tasks.
GRU-D [172]: Extends the GRU model to include a learnable decay term, such that the last observation is decayed to the empirical mean of time series.
SeFT [166] : Uses differentiable set function learning, such that all of the observations are first modeled individually and then pooled together via an attention based approach.
RAINDROP [165]: Assumes that each variable of MISTS acts as a separate sensor and leverages graph neural networks to learn the dependencies between different variables.
DGM^2-O [190]: A model initially designed for forecasting tasks, that utilizes a kernel-based approach to interpolate irregular time series.
MTGNN [191]: A graph neural network initially designed for forecasting tasks, in which the inter-variate relationships are constructed by connecting each node with its top k nearest neighbors in a defined metric space.
The implementations of IP-Net [163] and mTAND [164] follow the original paper[1] [2]. We directly adopt the implementations of GRU-D [172], SeFT [166], RAINDROP [165], DGM^2-O [190] and MTGNN [191] provided by [165] [3].
Following [165], predictions with forecasting models are designed as single-step forecasting problems.

---

[1] https://github.com/mlds-lab/interp-net

[2] https://github.com/reml-lab/mTAN

[3] https://github.com/mims-harvard/Raindrop

C.3.2 Irregular clinical notes baselines

Time-Aware LSTM (T-LSTM) [193]: A variant of LSTM taking the elapsed time between notes into account with a decreasing function.
Flexible Time-aware LSTM (FT-LSTM) [168]: Encodes the temporal information of clinical notes by utilizing time-aware trainable parameters in an LSTM cell.
We utilize Clinical-Longformer with a maximum sequence length of 1024 [192] as the text encoder by using the pre-trained weights provided in HuggingFace [200][4]. We directly adopt the implementations of T-LSTM and FT-LSTM provided by [168]. and GRU-D [172] provided by [165]. We leverage the same implementation of mTAND as MISTS baseline.

C.3.3 Multimodal fusion baselines

Multimodal Adaptation Gate (MAG) [205, 181]:Adjusts the representation of one modality with a displacement vector derived from the other modalities.
Tensor Fusion (TF) [195, 196]: Performs an outer product on representations of different modalities.
Multimodal Transformer (MulT) [184]: Uses a cross-modal Transformer followed by a self-attention Transformer to obtain multimodal representations across time steps for each modality.
We utilize the implementations of MAG and TF provided by [19] [5], and MulT [184] provided by the original paper[6]. We perform Concat, MAG and TF as late fusion by first applying a Transformer on every modality to acquire representations of different modalities, and then integrating the last hidden state of every single modality with different fusion strategies to obtain multimodal representations for downstream tasks.

C.4 Hyperparameters and training details

We use a batch size of 32 and learning rate for pre-trained language models (PLMs) of 2×10^{-5} and others of 0.0004. We use the Adam algorithm for gradient-based optimization [59]. We store the parameters that obtain the highest F1 and Macro-F1 in the validation set, and use it to make predictions for testing samples for 48-IHM and 24-PHE, respectively. The chosen hyperparameters are the same across tasks (48-IHM and 24-PHE) and models (both baselines and our methods) based on MISTS, irregular clinical note and multimodal fusion settings.

---

[4] https://huggingface.co/yikuan8/Clinical-Longformer

[5] https://github.com/emnlp-mimic/mimic

[6] https://github.com/yaohungt/Multimodal-Transformer

C.4.1 MISTS

For all MISTS models, we run the models for 20 epochs. We search for hidden units of Imputation, mTAND, IP-Net, GRU-D and SeFT, over the range {64,128}. For Imputation, we set the kernel size of 1D Convolution as 1. For mTAND we search for hidden size of time embeddings over the range {64,128} and take the the number of time embeddings, V, to be 8. We utilize a 3-layer Transformer as the backbone encoder for Imputation, mTAND and IP-Net. For UTDE, we search the hyperparameters of submodules Imputation and mTAND over the same range as the model with only a single method, and use a 3-layer Transformer as backbone encoder. We search for the gate integration level in {"patient", "temporal", "hidden space" }.

C.4.2 Irregular clinical notes

In our primary study, we empirically found that all models in the clinical note modality converge within 6 epochs, so that we train all the models for 6 epochs. In addition, we found that fine-tuning the PLM in the first 3 epochs and regarding the PLM as a feature extractor in later epochs achieved better results than fine-tuning the PLM in the whole training. We search for hidden units of T-LSTM, FT-LSTM, GRU-D and mTAND^{txt} over the range {64,128}. For mTAND^{txt}, time embeddings hidden size is searched over the range {64,128} and the number of embeddings V is equal to 8.

C.4.3 Multimodal fusion

Same as the clinical note modality, we run all fusion models for 6 epochs, and fine-tune the PLM in the first 3 epochs. We utilize 3-layer Transformer encoders to encode each modality for Concat, MAG and TF. For MulT, we perform 3 layer cross-modal Transformer followed by a 3 layer self-attention Transformer for each modality. We learn a 3 layer interleaved Transformer for our multimodal fusion strategy (J=3). We search for the hyperparameters of UTDE and mTAND^{txt} over the same range in each single modality setting. We search for the hidden size of Transformers over the range {64,128}.

Bibliography

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language models are few-shot learners*, in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. B. Rao, P. Barnes, Y. Tay, N. M. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. C. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. García, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. O. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. S. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, *Palm: Scaling language modeling with pathways*, *ArXiv* **abs/2204.02311** (2022).

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, *Training language models to follow instructions with human feedback*, 2022.

[4] OpenAI, *Introducing chatgpt*, 2022. Accessed: 2023-05-11.

[5] OpenAI, *Gpt-4 technical report*, 2023.

[6] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et. al.*, *Large language models encode clinical knowledge*, *arXiv preprint arXiv:2212.13138* (2022).

[7] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, *Capabilities of gpt-4 on medical challenge problems*, 2023.

[8] V. Liévin, C. E. Hother, and O. Winther, *Can large language models reason about medical questions?*, 2023.

[9] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, *Mimic-iii, a freely accessible critical care database*, *Scientific data* **3** (2016), no. 1 1–9.

[10] Anthropic, *Claude 2*, 2023.

[11] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts, *The flan collection: Designing data and methods for effective instruction tuning*, 2023.

[12] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, *Finetuned language models are zero-shot learners*, in *International Conference on Learning Representations*, 2022.

[13] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, *Medalpaca–an open-source collection of medical conversational ai models and training data*, *arXiv preprint arXiv:2304.08247* (2023).

[14] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, *Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge*, *Cureus* **15** (2023), no. 6.

[15] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, *Pmc-llama: Towards building open-source language models for medicine*, 2023.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, *Llama: Open and efficient foundation language models*, 2023.

[17] S. N. Golmaei and X. Luo, *Deepnote-gnn: predicting hospital readmission using clinical notes and patient network*, in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–9, 2021.

[18] M. Mahbub, S. Srinivasan, I. Danciu, A. Peluso, E. Begoli, S. Tamang, and G. D. Peterson, *Unstructured clinical notes within the 24 hours since admission predict short, mid & long-term mortality in adult icu patients*, *Plos one* **17** (2022), no. 1 e0262182.

[19] H. Yang, L. Kuang, and F. Xia, *Multimodal temporal-clinical note network for mortality prediction*, *Journal of Biomedical Semantics* **12** (2021), no. 1 1–14.

[20] X. Zhang, S. Li, Z. Cheng, R. Callcut, and L. Petzold, *Domain adaptation for trauma mortality prediction in ehrs with feature disparity*, in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1145–1152, 2021.

[21] X. Zhang, S. Li, X. Yang, C. Tian, Y. Qin, and L. R. Petzold, *Enhancing small medical learners with privacy-preserving contextual prompting*, 2023.

[22] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, and L. R. Petzold, *Alpacare:instruction-tuned large language models for medical application*, 2023.

[23] X. Zhang, S. Li, Z. Chen, X. Yan, and L. Petzold, *Improving medical predictions by irregular multimodal electronic health records modeling*, 2022.

[24] W. H. Organization *et. al.*, *Injuries and violence: the facts 2014*, .

[25] O. Blow, L. Magliore, J. A. Claridge, K. Butler, and J. S. Young, *The golden hour and the silver day: detection and correction of occult hypoperfusion within 24 hours improves outcome from major trauma*, *Journal of Trauma and Acute Care Surgery* **47** (1999), no. 5 964.

[26] M. J. Otero-López, P. Alonso-Hernández, J. A. Maderuelo-Fernández, B. Garrido-Corro, A. Domínguez-Gil, and A. Sánchez-Rodríguez, *Preventable adverse drug events in hospitalized patients*, *Medicina clinica* **126** (2006), no. 3 81–87.

[27] C. Subbe, A. Slater, D. Menon, and L. Gemmell, *Validation of physiological scoring systems in the accident and emergency department*, *Emergency Medicine Journal* **23** (2006), no. 11 841–845.

[28] T. Desautels, J. Calvert, J. Hoffman, Q. Mao, M. Jay, G. Fletcher, C. Barton, U. Chettipally, Y. Kerem, and R. Das, *Using transfer learning for improved mortality prediction in a data-scarce hospital setting*, *Biomedical informatics insights* **9** (2017) 1178222617712994.

[29] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, *Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis*, *IEEE journal of biomedical and health informatics* **22** (2017), no. 5 1589–1604.

[30] C. Xiao, E. Choi, and J. Sun, *Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review*, *Journal of the American Medical Informatics Association* **25** (2018), no. 10 1419–1428.

[31] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, *Self-training with noisy student improves imagenet classification*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.

[32] Z. Sun, S. Peng, Y. Yang, X. Wang, and F. Li, *A general fine-tuned transfer learning model for predicting clinical task acrossing diverse ehrs datasets*, in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 490–495, IEEE, 2019.

[33] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, *A theory of learning from different domains*, *Machine learning* **79** (2010), no. 1 151–175.

[34] G. Csurka, *Domain adaptation for visual applications: A comprehensive survey*, *arXiv preprint arXiv:1702.05374* (2017).

[35] Y. Ziser and R. Reichart, *Neural structural correspondence learning for domain adaptation*, *arXiv preprint arXiv:1610.01588* (2016).

[36] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, *Variational recurrent adversarial deep domain adaptation*, .

[37] F. Khoshnevisan and M. Chi, *An adversarial domain separation framework for septic shock early prediction across ehr systems*, in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 64–73, IEEE, 2020.

[38] Y. Zhang, T. B. Wu, B. J. Daigle, M. Cohen, and L. Petzold, *Identification of disease states associated with coagulopathy in trauma*, *BMC medical informatics and decision making* **16** (2016), no. 1 1–9.

[39] J. B. MacLeod, M. Lynn, M. G. McKenney, S. M. Cohn, and M. Murtha, *Early coagulopathy predicts mortality in trauma*, *Journal of Trauma and Acute Care Surgery* **55** (2003), no. 1 39–44.

[40] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Doretto, *Few-shot adversarial domain adaptation*, *arXiv preprint arXiv:1711.02536* (2017).

[41] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, *Unified deep supervised domain adaptation and generalization*, in *Proceedings of the IEEE international conference on computer vision*, pp. 5715–5725, 2017.

[42] L. van der Maaten and G. Hinton, *Visualizing data using t-sne*, *Journal of Machine Learning Research* **9** (2008), no. 86 2579–2605.

[43] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, *The eicu collaborative research database, a freely available multi-center database for critical care research*, *Scientific data* **5** (2018), no. 1 1–13.

[44] S. Purushotham, C. Meng, Z. Che, and Y. Liu, *Benchmarking deep learning models on large healthcare datasets*, *Journal of biomedical informatics* **83** (2018) 112–134.

[45] J. Fang, J. Zhu, and X. Zhang, *Prediction of length of stay on the intensive care unit based on bayesian neural network*, in *Journal of Physics: Conference Series*, vol. 1631, p. 012089, IOP Publishing, 2020.

[46] S. Kornblith, J. Shlens, and Q. V. Le, *Do better imagenet models transfer better?*, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019) 2656–2666.

[47] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, *Convolutional neural networks for medical image analysis: Full training or fine tuning?*, *IEEE transactions on medical imaging* **35** (2016), no. 5 1299–1312.

[48] Z. Li and D. Hoiem, *Learning without forgetting*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40** (2018) 2935–2947.

[49] G. Koch, R. Zemel, R. Salakhutdinov, *et. al.*, *Siamese neural networks for one-shot image recognition*, in *ICML deep learning workshop*, vol. 2, Lille, 2015.

[50] H. Shimodaira, *Improving predictive inference under covariate shift by weighting the log-likelihood function*, *Journal of statistical planning and inference* **90** (2000), no. 2 227–244.

[51] L. Torrey and J. Shavlik, *Transfer learning*, in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI global, 2010.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, *Advances in neural information processing systems* **27** (2014).

[53] R. Hadsell, S. Chopra, and Y. LeCun, *Dimensionality reduction by learning an invariant mapping*, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[54] C. Zhang, V. Maroufy, B. Chen, and H. Wu, *Missing data issues in ehr*, in *Statistics and Machine Learning Methods for EHR Data*, pp. 149–173. Chapman and Hall/CRC, 2020.

[55] S. v. Buuren and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in r*, *Journal of statistical software* (2010) 1–68.

[56] S. Wang, M. B. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. 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.

[57] K. Lin, Y. Hu, and G. Kong, *Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model*, *International journal of medical informatics* **125** (2019) 55–61.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et. al.*, *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in neural information processing systems*, pp. 8026–8037, 2019.

[59] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, *arXiv preprint arXiv:1412.6980* (2014).

[60] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training gans*, *Advances in neural information processing systems* **29** (2016) 2234–2242.

[61] L. Van der Maaten and G. Hinton, *Visualizing data using t-sne.*, *Journal of machine learning research* **9** (2008), no. 11.

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et. al.*, *Scikit-learn: Machine learning in python*, *the Journal of machine Learning research* **12** (2011) 2825–2830.

[63] S. Li, J. Chen, Y. Shen, Z. Chen, X. Zhang, Z. Li, H. Wang, J. Qian, B. Peng, Y. Mao, W. Chen, and X. Yan, *Explanations from large language models make small reasoners better*, 2022.

[64] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, *Domain-specific language model pretraining for biomedical natural language processing*, *ACM Transactions on Computing for Healthcare (HEALTH)* **3** (2021), no. 1 1–23.

[65] M. Yasunaga, J. Leskovec, and P. Liang, *Linkbert: Pretraining language models with document links*, *arXiv preprint arXiv:2203.15827* (2022).

[66] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, *Biobert: a pre-trained biomedical language representation model for biomedical text mining*, *Bioinformatics* **36** (2020), no. 4 1234–1240.

[67] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, *Publicly available clinical bert embeddings*, *arXiv preprint arXiv:1904.03323* (2019).

[68] E. Bolton, D. Hall, M. Yasunaga, T. Lee, C. Manning, and P. Liang, "Pubmed gpt." `https://crfm.stanford.edu/2022/12/15/pubmedgpt.html`, 2022. Accessed: 2023-03-29.

[69] E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, and E. Alsentzer, *Do we still need clinical language models?*, 2023.

[70] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, 2019.

[71] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, *Journal of Machine Learning Research* **21** (2020), no. 140 1–67.

[72] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.

[73] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, *Chain-of-thought prompting elicits reasoning in large language models*, 2023.

[74] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, *Reframing human-ai collaboration for generating free-text explanations*, *arXiv preprint arXiv:2112.08674* (2021).

[75] Y. Fu, H.-C. Peng, L. Ou, A. Sabharwal, and T. Khot, *Specializing smaller language models towards multi-step reasoning*, *ArXiv* **abs/2301.12726** (2023).

[76] N. Ho, L. Schmid, and S.-Y. Yun, *Large language models are reasoning teachers*, *ArXiv* **abs/2212.10071** (2022).

[77] K. Shridhar, A. Stolfo, and M. Sachan, *Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions*, *ArXiv* **abs/2212.00193** (2022).

[78] M. Neumann, D. King, I. Beltagy, and W. Ammar, *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing*, in *Proceedings of the 18th BioNLP Workshop and Shared Task*, (Florence, Italy), pp. 319–327, Association for Computational Linguistics, Aug., 2019. 1902.0766.

[79] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et. al.*, *Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models*, *PLoS digital health* **2** (2023), no. 2 e0000198.

[80] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi,

J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, *Scaling instruction-finetuned language models*, 2022.

[81] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, *Don't stop pretraining: Adapt language models to domains and tasks*, 2020.

[82] G. Hinton, O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network*, 2015.

[83] P. Wang, A. Chan, F. Ilievski, M. Chen, and X. Ren, *Pinto: Faithful language reasoning using prompt-generated rationales*, 2022.

[84] J. Bardhan, A. Colas, K. Roberts, and D. Z. Wang, *Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries*, *arXiv preprint arXiv:2205.01290* (2022).

[85] C. Shivade *et. al.*, *Mednli-a natural language inference dataset for the clinical domain*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics*, pp. 1586–1596, 2019.

[86] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, and D. Jurafsky, *Improving factual completeness and consistency of image-to-text radiology report generation*, *arXiv preprint arXiv:2010.10042* (2020).

[87] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, *Generate rather than retrieve: Large language models are strong context generators*, 2023.

[88] G. Izacard and E. Grave, *Leveraging passage retrieval with generative models for open domain question answering*, 2020.

[89] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, *Qa-gnn: Reasoning with language models and knowledge graphs for question answering*, 2022.

[90] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. Liang, and J. Leskovec, *Deep bidirectional language-knowledge graph pretraining*, in *Neural Information Processing Systems (NeurIPS)*, 2022.

[91] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, *REALM: retrieval-augmented language model pre-training*, *CoRR* **abs/2002.08909** (2020) [arXiv:2002.0890].

[92] K. Lee, M. Chang, and K. Toutanova, *Latent retrieval for weakly supervised open domain question answering*, *CoRR* **abs/1906.00300** (2019) [arXiv:1906.0030].

[93] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, *Retrieval-augmented generation for knowledge-intensive NLP tasks*, *CoRR* **abs/2005.11401** (2020) [arXiv:2005.1140].

[94] D. Chen, A. Fisch, J. Weston, and A. Bordes, *Reading wikipedia to answer open-domain questions*, *CoRR* **abs/1704.00051** (2017) [arXiv:1704.0005].

[95] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, *Okapi at trec-3*, in *Text Retrieval Conference*, 1994.

[96] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, *Dense passage retrieval for open-domain question answering*, *CoRR* **abs/2004.04906** (2020) [arXiv:2004.0490].

[97] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, *RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 5835–5847, Association for Computational Linguistics, June, 2021.

[98] D. S. Sachan, S. Reddy, W. L. Hamilton, C. Dyer, and D. Yogatama, *End-to-end training of multi-document reader and retriever for open-domain question answering*, *CoRR* **abs/2106.05346** (2021) [arXiv:2106.0534].

[99] K. Lee, M. Chang, and K. Toutanova, *Latent retrieval for weakly supervised open domain question answering*, *CoRR* **abs/1906.00300** (2019) [arXiv:1906.0030].

[100] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec, *Greaselm: Graph reasoning enhanced language models for question answering*, 2022.

[101] F. Petroni, T. Rocktäschel, P. S. H. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, *Language models as knowledge bases?*, *CoRR* **abs/1909.01066** (2019) [arXiv:1909.0106].

[102] A. Roberts, C. Raffel, and N. Shazeer, *How much knowledge can you pack into the parameters of a language model?*, *CoRR* **abs/2002.08910** (2020) [arXiv:2002.0891].

[103] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. L. Bras, Y. Choi, and H. Hajishirzi, *Generated knowledge prompting for commonsense reasoning*, *CoRR* **abs/2110.08387** (2021) [arXiv:2110.0838].

[104] S. Sousa and R. Kern, *How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing*, *Artificial Intelligence Review* **56** (2023), no. 2 1427–1492.

[105] A. Act, *Health insurance portability and accountability act of 1996*, *Public law* **104** (1996) 191.

[106] R. Chevrier, V. Foufi, C. Gaudet-Blavignac, A. Robert, and C. Lovis, *Use and understanding of anonymization and de-identification in the biomedical literature: scoping review*, *Journal of medical Internet research* **21** (2019), no. 5 e13484.

[107] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, *Survey on rnn and crf models for de-identification of medical free text*, *Journal of Big Data* **7** (2020) 1–22.

[108] Z. Liu, B. Tang, X. Wang, and Q. Chen, *De-identification of clinical notes via recurrent neural network and conditional random field*, *Journal of biomedical informatics* **75** (2017) S34–S42.

[109] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, *De-identification of patient notes with recurrent neural networks*, *Journal of the American Medical Informatics Association* **24** (2017), no. 3 596–606.

[110] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, *Mimic-iv*, *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)* (2020).

[111] C. Dwork, *Differential privacy: A survey of results*, in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5*, pp. 1–19, Springer, 2008.

[112] N. Fernandes, M. Dras, and A. McIver, *Generalised differential privacy for text document processing*, in *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*, pp. 123–148, Springer International Publishing, 2019.

[113] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, *Deep learning with differential privacy*, in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[114] O. Melamud and C. Shivade, *Towards automatic generation of shareable synthetic clinical notes using neural language models*, in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, (Minneapolis, Minnesota, USA), pp. 35–45, Association for Computational Linguistics, June, 2019.

[115] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, *What disease does this patient have? a large-scale open domain question answering dataset from medical exams*, 2020.

[116] D. Vilares and C. Gómez-Rodríguez, *Head-qa: A healthcare dataset for complex reasoning*, 2019.

[117] A. Pal, L. K. Umapathi, and M. Sankarasubbu, *Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering*, 2022.

[118] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, *Measuring massive multitask language understanding*, 2021.

[119] D. Dai, W.-J. Jiang, J. Zhang, W. Peng, Y. Lyu, Z. Sui, B. Chang, and Y. Zhu, *Mixture of experts for biomedical question answering*, *ArXiv* **abs/2204.07469** (2022).

[120] J. Mao, J. Zhang, Z. Zeng, W. Peng, W.-J. Jiang, X. Wang, H. Liu, and Y. Lyu, *Hierarchical representation-based dynamic reasoning network for biomedical question answering*, in *International Conference on Computational Linguistics*, 2022.

[121] V. Liévin, A. G. Motzfeldt, I. R. Jensen, and O. Winther, *Variational open-domain question answering*, 2022.

[122] Y. Liu, S. Chowdhury, C. Zhang, C. Caragea, and P. S. Yu, *Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering*, 2020.

[123] Y. Ge, Y. Guo, Y.-C. Yang, M. A. Al-Garadi, and A. Sarker, *Few-shot learning for medical text: A systematic review*, 2022.

[124] A. Talmor, J. Herzig, N. Lourie, and J. Berant, *Commonsenseqa: A question answering challenge targeting commonsense knowledge*, *CoRR* **abs/1811.00937** (2018) [arXiv:1811.0093].

[125] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, *Can a suit of armor conduct electricity? a new dataset for open book question answering*, in *EMNLP*, 2018.

[126] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, *Self-instruct: Aligning language models with self-generated instructions*, 2023.

[127] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model." `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[128] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang, *Wizardlm: Empowering large language models to follow complex instructions*, 2023.

[129] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*, March, 2023.

[130] L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, and H. Jin, *Alpagasus: Training a better alpaca with fewer data*, 2023.

[131] A. Bukharin and T. Zhao, *Data diversity matters for robust instruction tuning*, 2024.

[132] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz, *Can generalist foundation models outcompete special-purpose tuning? case study in medicine*, 2023.

[133] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, *Towards expert-level medical question answering with large language models*, 2023.

[134] C. Xu, D. Guo, N. Duan, and J. McAuley, *Baize: An open-source chat model with parameter-efficient tuning on self-chat data*, 2023.

[135] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023.

[136] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, *What disease does this patient have? a large-scale open domain question answering dataset from medical exams*, *Applied Sciences* **11** (2021), no. 14 6421.

[137] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, *Pubmedqa: A dataset for biomedical research question answering*, 2019.

[138] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, *Multitask prompted training enables zero-shot task generalization*, in *International Conference on Learning Representations*, 2022.

[139] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language models are few-shot learners*, 2020.

[140] B. Peng, C. Li, P. He, M. Galley, and J. Gao, *Instruction tuning with gpt-4*, *arXiv preprint arXiv:2304.03277* (2023).

[141] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, J. Kernion, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, and J. Kaplan, *A general language assistant as a laboratory for alignment*, 2021.

[142] Sharegpt, *Sharegpt*, 2023.

[143] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. K. Keloth, X. Zhou, H. He, L. Ohno-Machido, Y. Wu, H. Xu, and J. Bian, *Me llama: Foundation large language models for medical applications*, *ArXiv* **abs/2402.12749** (2024).

[144] H. Tran, Z. Yang, Z. Yao, and H. Yu, *Bioinstruct: Instruction tuning of large language models for biomedical natural language processing*, *ArXiv* **abs/2310.19975** (2023).

[145] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, *Judging llm-as-a-judge with mt-bench and chatbot arena*, 2023.

[146] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, *Large language models are not fair evaluators*, 2023.

[147] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpacaeval: An automatic evaluator of instruction-following models." `https://github.com/tatsu-lab/alpaca_eval`, 2023.

[148] D. Vilares and C. Gómez-Rodríguez, *HEAD-QA: A healthcare dataset for complex reasoning*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 960–966, Association for Computational Linguistics, July, 2019.

[149] A. Ben Abacha and D. Demner-Fushman, *On the summarization of consumer health questions*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 2228–2234, Association for Computational Linguistics, July, 2019.

[150] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, *A framework for few-shot language model evaluation*, Sept., 2021.

[151] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, *Lima: Less is more for alignment*, 2023.

[152] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto, *Alpacafarm: A simulation framework for methods that learn from human feedback*, 2023.

[153] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei, *Challenging big-bench tasks and whether chain-of-thought can solve them*, 2022.

[154] S. Lin, J. Hilton, and O. Evans, *Truthfulqa: Measuring how models mimic human falsehoods*, 2022.

[155] Y. K. Chia, P. Hong, L. Bing, and S. Poria, *Instructeval: Towards holistic evaluation of instruction-tuned large language models*, 2023.

[156] S. A. Tisherman and D. M. Stein, *Icu management of trauma patients*, *Critical Care Medicine* **46** (2018), no. 12 1991–1997.

[157] C. Alberti, C. Brun-Buisson, H. Burchardi, C. Martin, S. Goodman, A. Artigas, A. Sicignano, M. Palazzo, R. Moreno, R. Boulmé, *et. al.*, *Epidemiology of sepsis and infection in icu patients from an international multicentre cohort study*, *Intensive care medicine* **28** (2002), no. 2 108–121.

[158] B. Afessa, O. Gajic, and M. T. Keegan, *Severity of illness and organ failure assessment in adult intensive care units*, *Critical care clinics* **23** (2007), no. 3 639–658.

[159] J. Adler-Milstein, C. M. DesRoches, P. Kralovec, G. Foster, C. Worzala, D. Charles, T. Searcy, and A. K. Jha, *Electronic health record adoption in us hospitals: progress continues, but challenges persist*, *Health affairs* **34** (2015), no. 12 2174–2180.

[160] X. Zhang, S. Li, Z. Cheng, R. Callcut, and L. Petzold, *Domain adaptation for trauma mortality prediction in ehrs with feature disparity*, in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1145–1152, 2021.

[161] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, *Multitask learning and benchmarking with clinical time series data*, *Scientific Data* **6** (2019), no. 1 96.

[162] Z. C. Lipton, D. Kale, and R. Wetzel, *Directly modeling missing data in sequences with rnns: Improved classification of clinical time series*, in *Machine learning for healthcare conference*, pp. 253–270, PMLR, 2016.

[163] S. N. Shukla and B. Marlin, *Interpolation-prediction networks for irregularly sampled time series*, in *International Conference on Learning Representations*, 2019.

[164] S. N. Shukla and B. M. Marlin, *Multi-time attention networks for irregularly sampled time series*, *arXiv preprint arXiv:2101.10318* (2021).

[165] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, *Graph-guided network for irregularly sampled multivariate time series*, *arXiv preprint arXiv:2110.05357* (2021).

[166] M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt, *Set functions for time series*, in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 4353–4363, PMLR, 13–18 Jul, 2020.

[167] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, *Latent ordinary differential equations for irregularly-sampled time series*, *Advances in neural information processing systems* **32** (2019).

[168] D. Zhang, J. Thadajarassiri, C. Sen, and E. Rundensteiner, *Time-aware transformer-based network for clinical notes series prediction*, in *Machine Learning for Healthcare Conference*, pp. 566–588, PMLR, 2020.

[169] I. Deznabi, M. Iyyer, and M. Fiterau, *Predicting in-hospital mortality by combining clinical notes with time-series data*, in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4026–4031, 2021.

[170] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, *Advances in neural information processing systems* **30** (2017).

[171] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, *A transformer-based framework for multivariate time series representation learning*, in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2114–2124, 2021.

[172] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, *Recurrent neural networks for multivariate time series with missing values*, *Scientific reports* **8** (2018), no. 1 1–12.

[173] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, *Advances in neural information processing systems* **31** (2018).

[174] M. McDermott, B. Nestor, E. Kim, W. Zhang, A. Goldenberg, P. Szolovits, and M. Ghassemi, *A comprehensive ehr timeseries pre-training benchmark*, in *Proceedings of the Conference on Health, Inference, and Learning*, pp. 257–278, 2021.

[175] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, *Outrageously large neural networks: The sparsely-gated mixture-of-experts layer*, *arXiv preprint arXiv:1701.06538* (2017).

[176] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, *Adaptive mixtures of local experts*, *Neural computation* **3** (1991), no. 1 79–87.

[177] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, *arXiv preprint arXiv:1810.04805* (2018).

[178] K. Huang, J. Altosaar, and R. Ranganath, *Clinicalbert: Modeling clinical notes and predicting hospital readmission*, *arXiv preprint arXiv:1904.05342* (2019).

[179] Z. Liu, J. Zhang, Y. Hou, X. Zhang, G. Li, and Y. Xiang, *Machine learning for multimodal electronic health records-based research: Challenges and perspectives*, *arXiv preprint arXiv:2111.04898* (2021).

[180] S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava, *Using clinical notes with time series data for icu management*, *arXiv preprint arXiv:1909.09702* (2019).

[181] B. Yang and L. Wu, *How to leverage multimodal ehr data for better medical predictions?*, *arXiv preprint arXiv:2110.15763* (2021).

[182] Z. Xu, D. R. So, and A. M. Dai, *Mufasa: Multimodal fusion architecture search for electronic health records*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10532–10540, 2021.

[183] S. M. Kazemi, R. Goel, S. Eghbali, J. Ramanan, J. Sahota, S. Thakur, S. Wu, C. Smyth, P. Poupart, and M. Brubaker, *Time2vec: Learning a vector representation of time*, *arXiv preprint arXiv:1907.05321* (2019).

[184] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, *Multimodal transformer for unaligned multimodal language sequences*, in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, p. 6558, NIH Public Access, 2019.

[185] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, *Doctor ai: Predicting clinical events via recurrent neural networks*, in *Machine learning for healthcare conference*, pp. 301–318, PMLR, 2016.

[186] P. Gupta, P. Malhotra, L. Vig, and G. Shroff, *Transfer learning for clinical time series analysis using recurrent neural networks*, *arXiv preprint arXiv:1807.01705* (2018).

[187] A. Arbabi, D. R. Adams, S. Fidler, M. Brudno, *et. al.*, *Identifying clinical terms in medical text using ontology-guided machine learning*, *JMIR medical informatics* **7** (2019), no. 2 e12596.

[188] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, *Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting*, in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[189] B. Lim and S. Zohren, *Time-series forecasting with deep learning: a survey*, *Philosophical Transactions of the Royal Society A* **379** (2021), no. 2194 20200209.

[190] Y. Wu, J. Ni, W. Cheng, B. Zong, D. Song, Z. Chen, Y. Liu, X. Zhang, H. Chen, and S. B. Davidson, *Dynamic gaussian mixture based deep generative model for robust forecasting on sparse multivariate time series*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 651–659, 2021.

[191] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, *Connecting the dots: Multivariate time series forecasting with graph neural networks*, in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 753–763, 2020.

[192] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, *Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences*, *arXiv preprint arXiv:2201.11838* (2022).

[193] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, *Patient subtyping via time-aware lstm networks*, in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74, 2017.

[194] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, *Hierarchical transformers for long document classification*, in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844, IEEE, 2019.

[195] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, *Tensor fusion network for multimodal sentiment analysis*, *arXiv preprint arXiv:1707.07250* (2017).

[196] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, *Efficient low-rank multimodal fusion with modality-specific factors*, *arXiv preprint arXiv:1806.00064* (2018).

[197] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, *Integrating multimodal information in large pretrained transformers*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 2359–2369, Association for Computational Linguistics, July, 2020.

[198] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998), no. 11 2278–2324.

[199] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, *Neural computation* **9** (1997), no. 8 1735–1780.

[200] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, *Transformers: State-of-the-art natural language processing*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct., 2020.

[201] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019.

[202] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, and W. Y. Wang, *A survey on large language models for critical societal domains: Finance, healthcare, and law*, 2024.

[203] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, *Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies*, *Transactions of the Association for Computational Linguistics (TACL)* (2021).

[204] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, *Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs*, 2019.

[205] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, *Integrating multimodal information in large pretrained transformers*, in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, p. 2359, NIH Public Access, 2020.