重新审视 MIMIC-IV 基准：使用语言模型进行电子健康记录的实验

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G. Moreno and Lynda Tamine
University of Toulouse, IRIT, 31000 Toulouse, France
{ firstname.lastname } @irit.fr

Abstract

The lack of standardized evaluation benchmarks in the medical domain for text inputs can be a barrier to widely adopting and leveraging the potential of natural language models for health-related downstream tasks. This paper revisited an openly available MIMIC-IV benchmark for electronic health records (EHRs) to address this issue. First, we integrate the MIMIC-IV data within the Hugging Face *datasets* library to allow an easy share and use of this collection. Second, we investigate the application of templates to convert EHR tabular data to text. Experiments using fine-tuned and zero-shot LLMs on the mortality of patients task show that fine-tuned text-based models are competitive against robust tabular classifiers. In contrast, zero-shot LLMs struggle to leverage EHR representations. This study underlines the potential of text-based approaches in the medical field and highlights areas for further improvement.

Keywords: Large language models, MIMIC-IV benchmark, Text-based mortality classification

最近，自然语言处理（NLP）和信息检索（IR）任务的进展主要由基于 Transformer 的模型推动，如 BERT 和 RoBERTa。这些模型在最少监督下完成了对原始语言信息的训练。此外，大型语言模型（LLM）的出现，如 ChatGPT 和 Llama 2，通过扩大参数规模和训练数据扩展了这些功能。在医学领域，应用 LLMs 成为患者和医疗从业者的一个新工具。例如，由实验室测量、程序和药物代码等非语言信息组成的电子健康记录（EHR），通过这些模型被翻译成语言报告。然而，目前尚不清楚 EHR 模型表示对非语言任务的实用性。除了隐私方面的担忧，阻碍 LLMs 广泛采用解决这一问题的关键问题是有效地将患者结构化信息从原始 EHR 格式转换为语言非结构化格式，以利用 LLM 的文本表示的潜力。现有用于患者数据的基于 Transformer 的模型，如 TransformEHR 和 BEHRT，已经调整了其架构以考虑表格输入数据。然而，这一过程需要一个昂贵的预训练步骤，无法利用改进的 LLMs 和免费 EHR 基准如 MIMIC IV 的进步。后者提供了与不同下游任务（例如，死亡患者分类）相关的表格形式的大规模重症监护病房（ICU）患者数据。因此，我们认为提高这些资源的可访问性以满足模型的演进对于这一领域至关重要。

在本文中，我们提出了一种简单但有效的方法来标准化 MIMIC-IV 基准，以便使用最新的基于 Transformer 的架构（BERT, DistilBERT (Sanh et al., 2019) 和 RoBERTa）以及大型语言模型（Llama 2, Meditron (Chen et al., 2023)）用于健康相关的预测任务。为此，我们在 ICU 数据上识别出六个主要特征组，并提出基于模板的数据到文本转换。因此，我们能够提供一个文本文档输入，概述患者的 ICU 记录。此外，为了保证可重复性，我们提供了一个 Hugging Face 数据集对象 [1]，该对象可以自动生成所需文本格式的临床队列 [2]。我们的主要贡献如下：1）一个标准的 MIMIC-IV 基准，已集成到 Hugging Face 数据集库中，允许灵活使用电子健康记录在健康相关的下游任务中的表示；2）使用八种不同模型进行全面实验，以评估我们重新审视的 MIMIC-IV 基准在死亡率分类任务上的有效性。

1. 背景和相关工作

1.1. MIMIC 数据集和基准测试

医疗信息市场的重症监护病房（MIMIC）数据集 (Johnson et al., 2023; Johnson et al.) 是最大和最新的电子健康记录数据集之一。它包含超过 250,000 名在波士顿贝丝以色列女执事医院（Beth Israel Deaconess Medical Center）的重症监护病房接受治疗的患者。出于隐私考虑，每位患者在重症监护病房全过程的详细信息都以去识别化的形式提供 [3]。当前版本是 MIMIC-IV 数据集 (Gupta et al., 2022)，收集了 2008 年至 2019 年之间的患者数据，并使用疾病国际分类（ICD）的 ICD-9 和 ICD-10 版本 [4] 来列出诊断并将医疗程序与诊断关联

。在最近的研究中，使用 MIMIC 数据集 (Harutyunyan et al., 2019; Gupta et al., 2022; Wang et al., 2020)，提出了多个医疗领域的基准点 (Johnson et al., 2023, 2016)。它们作为模型可比性和可重复性的主流手段出现。MIMIC-IV 数据管道 (Gupta et al., 2022) 提出用于预处理下游任务的数据。该管道能够将原始数据转换为准备好的患者数据表格形式。此外，它提供了 ICD 的映射以及标准的降维技术。虽然第一步是提出基准点，但我们在这项工作中计划迈出两步，提出将 MIMIC IV 基准整合

[1] https://huggingface.co/docs/datasets/index
[2] 在 https://huggingface.co/datasets/thbndi/Mimic4Dataset 公开可用
[3] 关于 HIPAA 的安全港条款。
[4] https://www.who.int/standards/classifications/classification-of-diseases

到 Hugging Face 数据集 [5]，这是最大的现成数据集中心之一，并且可能使用基于 Transformer 的模型（包括 LLM）进行电子健康记录的预测任务。

1.2. 用于电子健康记录的 Transformer 模型

通用领域的基于 Transformers 的模型，如 BERT，已经通过使用与医学相关的语言集合（如 PubMed）适应于临床领域（BioBERT (Lee et al., 2020) 和 ClinicalBERT (Alsentzer et al., 2019) ）。最近，已经出现了将电子健康记录（EHRs）的非语言信息编码为模型以建模患者数据的努力，如 BEHRT (Li et al., 2020) 、Med-BERT (Rasmy et al., 2021) 和 TransformEHR (Yang et al., 2023) 。这些模型在灵活的架构中编码了不同的健康模式。然而，它们需要在大规模数据集上进行预训练，并且无法从 NLP 文献中 Transformer 模型的显著进展中受益。此外，诸如 ChatGPT (Achiam et al., 2023) 、Llama 2 (Touvron et al., 2023) 及其医学变体 Meditron (Chen et al., 2023) 等大型语言模型在将非语言健康数据（如图像和 EHR 诊断）转化为文本的不同临床任务中表现出色 (Meskó and Topol, 2023; Yeo et al., 2023) 。然而，对这种语言 EHR 表示用于非语言任务的探索，称为 EHR 下游任务，仍然有限。为了弥补这一差距，我们对 EHR 数据进行了实验以探索它们的潜力。

2. MIMIC-IV 基准测试再探

在这里，我们详细介绍了所使用的管道和 EHR 数据，然后描述了将 EHR 表格数据转换为文本输入所提出的模板。

2.1. 管道

我们依赖于 MIMIC-IV 基准来生成文本的标准评估框架。因此，首先，我们在数据集库中集成了推荐的预处理指南，并实现了 MIMIC-IV-Data-Pipeline [6] 所提供的表格式形式的所有功能，如图 1 左侧所示。在预处理步骤之后，我们获得了一个表格表示，其中包括与实验室、药物、程序和生命体征相关的人口统计、当前诊断特征和时间序列特征，如表 1 所示。

请注意，像 CHAR/LAB 这样的特征是以时间间隔的形式给出的，因此必须应用一种缩减/扩展策略来规范表示的大小。数据插补通常通过从固定数量的时间窗口中采样数据，甚至是在一系列时间窗口中对值进行平均来实现。如 2.3 节所示，我们没有发现这些特征的采样或平均之间存在较大差异。

2.2. 拟议的模板

最后，特征 EHR 数据通过模板策略转化为文本，如图 1 右侧所示，具体如下：

[5] 我们的实施通过要求用户提供原始数据来遵守 MIMIC 的访问政策。

[6] https://github.com/healthylaife/MIMIC-IV-Data-Pipeline

Table 1: MIMIC-IV 基准的特征列表。

Name	Description
Demographics (DEMO) :	The list of demographic data is a tiny vector corresponding to the patient's gender, ethnicity, medical insurance, and age category. This data is encoded to obtain a numerical vector.
Diagnosis (COND) :	The list of diagnoses established on a patient's admission is encoded using a one-hot vector of all ICD codes including the patient's identified diseases. Note that this vector could be large w.r.t. other features.
Chart Events/Lab (CHART/LAB) :	Gives the value of the biological item_id performed in time interval t .
Medications (MEDS) :	For each item_id corresponding to a medication the quantity administered in time interval t or zero if not administrated.
Procedures (PROC) :	The list of medical procedures performed is given as a form of a one-hot vector setting to 1 the item_id of procedures performed in time interval t .
Output Events (OUTE) :	The list of biological samples taken is encoded using a one-hot vector of each item_id of the samples performed in time interval t .

“患者 { 民族 } { 性别 }，{ 年龄 } 岁，由 { 保险 } 覆盖，被诊断为 { cond_text }。” 其中 { cond_text } 对应于 ICD10 诊断的文字描述。

“测量的图事件是：{ chart_text }。” 其中 图表_ 文字 是形式为 { mean_val } 的生物测量列表，对 { feat_label } 来说，mean_val 是该集中的 { feat_label } 测量的平均值。

“在事件期间给药的平均量为：{ meds_text }。” 其中，{ meds_text } 是按以下形式表示的药物剂量列表：{ mean_val } 表示药物 { feat_label } 在事件期间的平均数量值。

“执行的程序为：{ proc_text } 。” 其中 { proc_text } 是指在此期间进行的医疗程序列表。

“收集到的输出是：{ out_text }。” 其中 { out_text } 是该事件期间采集的生物学和前生物学样本的列表。

表 2 显示了文本输入的一个示例。

为了确保我们实验的公平再现性，我们开发了一个数据集对象，该对象能够生成表格信息以及基于模板的文本数据。

对于表格数据，我们创建了表示法 1，该表示法遵循 (Gupta et al., 2022) 中使用的默认配置，但我们的实现中还有其他配置可供选择。类似地，表示法 2 是同一数据的聚合表示。主要区别在于最终特征的数量：前者使用 2766 个特征（作为每个窗口表示的连接结果），后者则使用 1110 个特征（因为所有窗口的值被平均）。我们在 (Gupta et al., 2022) 中提供的患者死亡率分类试点下游任

Figure 1: 生成表格格式和文本格式的数据集流水线。

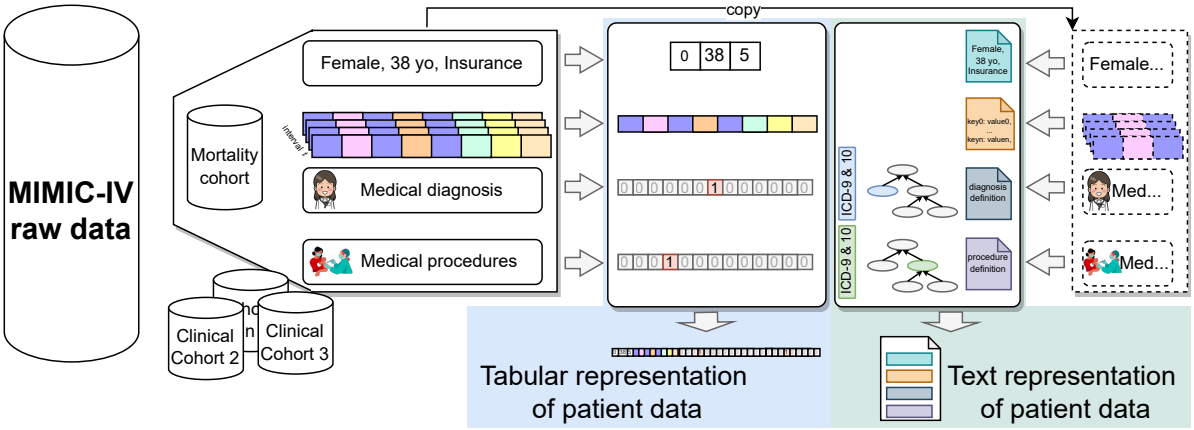

Table 2: 来自 MIMIC-IV 基准数据集的患者文本表示示例。数值已更改以避免泄露示例。

Feature	Example text
DEMO	The patient white male, 55 years old, covered by Other
COND	was diagnosed with Streptococcal sepsis; Acute pancreatitis; resistance to anti-microbial drugs.
CHAR/LAB	The chart events measured were: 73.655 for Heart Rate; 116.859 for Heart rate Alarm - High; ...
MEDS	The mean amounts of medications administered during the episode were: 44.778 of Albumin 5 % ; ...
PROC	The procedures performed were: Dialysis Catheter; 18 Gauge; EKG; ...
OUTE	The outputs collected were: OR EBL; OR Urine; Pre-Admission; ...

务中评估了我们修订的 MIMIC-IV 基准。评估集中在基准的可重复性（参见第 4.2 节），以及使用代表性模型的可行性和有效性（参见第 4.3 节）。模型参数通过 Scikitlearn 库 [7] 中可用的经典机器学习算法的 5 折交叉验证进行选择。我们使用了用于表格数据的算法，例如梯度提升（默认参数）、XGBoost（objective= “binary :logistic”）、随机森林（n_estimators= 300 ，criterion= “gini”）和逻辑回归（默认参数）。

对于文本数据，我们微调了六种不同的基于 Transformer 的模型。我们使用了最佳超参数，包括学习率为 $5e-5$ ，AdamW 优化和 3 个周期。对于我们使用 LLMs 的零样本设置，我们探索了多个提示。下面，我们报告了这几个提示中的两个，它们为任务提供了最高数量的有效响应。我们将输出生成限制为 2 个标记。

我们称提示为 P1：

> Prompt P1: “You are an extremely helpful healthcare assistant. You answer the question using only ’yes’ or ’no’ and considering a patient hospital profile: ‘[textual EHR]’.
> Question: Is the patient dead?.
> Answer (only yes or no): ”

同样地，我们称提示为 P2：

> Prompt P2: “Analyze the provided ICU data for a patient. The data covers the first 48 hours of the ICU stay, including vital statistics, lab test results, and treatments administered. Answer only Yes for a prediction of survival or No for a prediction of mortality. The patient ICU data is: ‘[textual EHR]’. Based on this data, answer.
> Question: Will the patient survive in the next 24 hours?.
> Answer (use only yes or no): ”

我们为微调模型设置了 512 个标记的输入长度限制，以及为零样本模型设置了 1024 个标记的限制。需要注意的是，这种截断只影响了微调模型，并且有时会移除与 MEDS、PROC 和 OUTE 特性相关的相关信息。在第 4.3 节中，我们讨论了一项消融研究，该研究探讨了这些特性的影响。

2.3. 利用表格形式的电子健康记录数据进行评估

我们在表格数据上的结果和来自原始基准 (Gupta et al., 2022) 的参考值在表 3 中展示。注意，我们的结果通过两种不同的聚合策略呈现：表现策略 1 和表现策略 2。在这两种情况下，我们的结果都略高于 (Gupta et al., 2022) 中提出并作为起始点使用的方法。这主要归功于我们对数据进行了细致的预处理。作为一个重要结果，注意到表现策略 2 栏

[7] https://scikit-learn.org/

Table 3: 我们标准化的 MIMIC-IV 与原始基准 (Gupta et al., 2022) 在患者死亡率分类任务上的比较评估。

Algorithm	Representation 1 AU-ROC	Representation 1 AU-PRC	Representation 2 AU-ROC	Representation 2 AU-PRC	(Gupta et al., 2022) AU-ROC	(Gupta et al., 2022) AU-PRC
Gradient Boosting	0.86	0.53	0.86	0.53	0.85	0.48
XGBoost	0.86	0.51	0.85	0.51	0.84	0.47
Random Forest	0.82	0.49	0.84	0.50	0.79	0.39
Logistic Regression	0.77	0.36	0.77	0.37	0.67	0.24

与表现策略 1 的表现相似，但使用的特征显著更少。此外，矢量表示中的 1,110 个值中有 1,034 个是稀疏的，因为它们用于诊断表示。这些结果促使我们继续探索基于文本的表示，因为仅有 66 个来自生物信号的值与文本数据（诊断）结合就足以在表格数据上达到最先进的结果。

2.4. 使用模板化文本输入的评估

利用基于文本的模型进行患者死亡率分类任务的主要结果如表 4 所示。对于微调模型，我们使用了三个通用训练模型，即 DistilBERT（distilbert-base-uncased (Sanh et al., 2019) ）、BERT（bert-base-uncased (Devlin et al., 2018) ）和 RoBERTa（roberta-base (Liu et al., 2019) ）（前三个），以及来自医学领域的另外三个，即 BioClinicalBERT（Bio_ClinicalBERT (Alsentzer et al., 2019) ）、BioBERT（dmis-lab/biobert-v1.1 (Lee et al., 2020) ）和 BioMedNLP（microsoft/BiomedNLP (Gu et al., 2021) ）（后三个）。我们仅报告了过采样 [8] 的结果。结果表明，通用和领域专用模型在 AU-ROC 方面表现相似，所有模型值接近（在 0.87 到 0.88 之间）。然而，AU-PRC 值有所不同，医学领域的模型优于通用模型。虽然观察到通用模型在 AU-PRC 方面有轻微改进，但不足以达到领域专用模型的性能。不出所料，对医学文本的微调有明显的兴趣。然而，通用模型，如 RoBERTa，紧随其后，接近最佳性能。

此外，我们探讨了在零样本设置中使用两个大型语言模型，分别是 Llama2（13b）（meta-llama/Llama-2-7b-hf (Touvron et al., 2023) ）及其医疗变体 Meditron（7b）（epfl-llm/meditron-7b (Chen et al., 2023) ），考虑了名为 P1 和 P2 的两个不同提示。我们普遍观察到，与微调模型相比，零样本部分（如表 4 所示）的表现较差。分析零样本部分后，我们发现提示 P1 比 P2 获得了更好的分数。这些结果表明，模型对该任务的查询格式较为敏感。此外，我们注意到，领域特定的模型例如 Meditron，相较于一般的模型如 Llama 2，在使用两个提示时表现更好，与微调设置相似。这些发现表明，最先进的大型语言模型在探索的提示范围内难以对电子健康记录进行编码和传递到下游任务。一个可能的发展方向是使用大型语言模型处理表格数据，定义更好的转换方法将这一结构化知识整合到语言模型中。此外，这些发现激励我们通过应用替代技术（例如情境学习 (Dong et al., 2022) 或提示微调 (Lester et al., 2021) ）进行进一步的研究和实验。

[8] 我们在没有过采样的情况下发现了类似的结果。

Table 4: 使用病人数据的文本表示，在病人死亡率任务中，通用模型和医疗领域模型的结果。

Models	AU-ROC	AU-PRC
Fine-tuned		
DistilBERT	0.87	0.42
BERT	0.87	0.43
RoBERTa	0.88	0.47
BioClinicalBERT	0.87	0.43
BioBERT	0.88	0.45
BiomedNLP	0.88	0.46
Zero-shot with prompt P1		
Llama 2 (13b)	0.50	0.38
Meditron (7b)	0.61	0.39
Zero-shot with prompt P2		
Llama 2 (13b)	0.50	0.13
Meditron (7b)	0.51	0.23

此外，在这个设置中，除了回答正确或错误之外，我们还考虑未回答的问题。当大型语言模型未能从预期的标记中提供输出（在我们的案例中是“是”或“否”）时，就会出现这样的问题。在我们的实验中，对于表格 4 中报告的结果，我们将“否”视为默认答案。为了提供更多细节，我们在表格 5 中显示了每个模型回答和未回答的问题数量。通过分析，我们发现 Llama 2 模型对数据集中的 3.30% 未作答，而 Meditron 模型使用提示 P1 仅有 0.04% 未作答。相比之下，提示 P2 在 Llama 2 上获得了 69.37% 未作答，而 Meditron 则没有未回答的问题。

通过比较用于描述任务的不同提示，我们可以观察到，Llama 2（通用领域模型）在进行一些修改时难以理解任务。相比之下，Meditron（领域特定模型）在使用任务的不同重构时更加稳定。

我们通过对两个具有代表性的模型 BERT 和 BiomedNLP 进行消融研究来进一步分析特征的累积效应。结果在表 6 中呈现。作为一个主要特征，我们可以轻松识别出 COND 是性能的明显提升者。

Table 5: 大语言模型在零样本设置下回答和未回答样本的数量。

Model	# answered	# unanswered
	With prompt P1	
Llama 2 (13b)	5952 (96.70%)	203 (3.30%)
Meditron (7b)	6152 (99.96%)	3 (0.04%)
	With prompt P2	
Llama 2 (13b)	1885 (30.63%)	4270 (69.37%)
Meditron (7b)	6155 (100.0%)	0 (0.00%)

仅此一个特征就能够达到接近最佳性能的值，表明它是病人档案表示的一个显著信号。然而，其他非专家基础的特征，如 CHAR/LAB，也非常可靠。注意这一结果是令人鼓舞的，因为特征是以聚合的形式给出的。事实上，与最佳模型相比，该模型在 AU-ROC 方面可以表现正确。另外，注意两个模型都在整合所有特征之前达到最佳性能。尤其是，MEDS, PROC 和 OUTE（仅针对 BiomedNLP）没有改善之前的组合。这表明值得研究更为精细的模板以整合这些特征。

在本文中，我们展示了一个可公开获取的 Hugging Face 数据集对象，它提供了一种可重现的方式来使用 MIMIC-IV 基准用于基于文本模型的电子健康记录（EHR）表示和健康相关任务。使用所提出的对象中的 MIMIC IV，我们旨在通过基于文本的模板，促进对其原始表格形式和文本格式的综合公共 EHR 数据集的实验。我们的实验表明，经过微调的基于文本的模型在 AU-ROC 方面表现类似于最强的基于表格的替代方案。相反，零样本设置中的大语言模型（LLM）在编码 EHR 信息时显示出局限性。这一评估为一类新的大语言模型家族提供了起点，以改善当前在健康相关预测任务上的最先进技术（SOTA）。

3. 文献参考

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop , pages 72–78.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079 .

E. Choi, Zhen Xu, Yujia Li, Michael W. Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M. Dai. 2019. Learning the graphical structure of electronic health records with graph convolutional transformer. In AAAI Conference on Artificial Intelligence .

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. arXiv preprint arXiv:2301.00234 .

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) , 3(1):1–23.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. 2022. An Extensive Data Processing Pipeline for MIMIC-IV. In Proceedings of the 2nd Machine Learning for Health symposium , volume 193 of Proceedings of Machine Learning Research , pages 311–325. PMLR.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. Scientific Data , 6(1):96.

Alistair EW Johnson, Lucas Bulgarelli, Tom J Pollard, Steven Horng, L A Celi, and R Mark. MIMIC-IV version 2.2.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng,

Table 6: 使用不同文本特征的消融研究。‘✓’表示该特征在患者表示中使用。CH/LA 代表图表/化验。加粗的结果表示最佳表现。

	COND	✓	✓	✓	✓	✓	✓	
	DEMO		✓	✓	✓	✓	✓	
	CH/LA			✓	✓	✓	✓	✓
	MEDS				✓	✓	✓	
	PROC					✓	✓	
Features	OUTE						✓	
				AU-ROC				
BERT		0.87	0.88	0.86	0.89	0.89	0.88	0.75
BiomedNLP		0.87	0.88	0.88	0.87	0.86	0.88	0.63
				AU-PRC				
BERT		0.41	0.44	0.40	0.46	0.46	0.46	0.24
BiomedNLP		0.45	0.46	0.50	0.42	0.39	0.47	0.13

Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. Scientific data , 10(1):1.

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

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. arXiv preprint arXiv:2005.10433 .

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics , 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 .

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. Scientific reports , 10(1):7155.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified data wrangling with ir_datasets. In SIGIR .

Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. NPJ digital medicine , 6(1):120.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine , 4(1):86.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing , pages 1586–1596.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. ArXiv , abs/1910.01108.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203 .

Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. 2017. Attend and diagnose: Clinical time series analysis using attention models. In Proceedings of the 2018 AAAI Association for the Advancement of Artificial Intelligence .

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

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. Research Square .

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems , 30.

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

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. 2023. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nature Communications , 14(1):7857.

Yee Hui Yeo, Jamil S Samaan, Wee Han Ng, Peng-Sheng Ting, Hirsh Trivedi, Aarshi Vipani, Walid Ayoub, Ju Dong Yang, Omer Liran, Brennan Spiegel, et al. 2023. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. Clinical and molecular hepatology , 29(3):721.