BioClinical ModernBERT：用于生物医学和临床自然语言处理的最先进长上下文编码器

Thomas Sounack[1] **Joshua Davis**[1,2] **Brigitte Durieux**[1,3]

Antoine Chaffin[4] **Tom J. Pollard**[5] **Eric Lehman**[6] **Alistair E. W. Johnson**[7]

Matthew McDermott [8] **Tristan Naumann**[9] **Charlotta Lindvall**[1,8]

[1] Dana-Farber Cancer Institute [2] Albany Medical College [3] McGill University
[4] LightOn [5] Massachusetts Institute of Technology [6] OpenEvidence
[7] Microsoft [8] Harvard Medical School [9] Microsoft Research

Correspondence: thomas_sounack@dfci.harvard.edu

Abstract

基于编码器的 transformer 模型在生物医学和临床自然语言处理（NLP）中起着核心作用，因为它们的双向自注意力机制使它们非常适合通过判别任务从非结构化文本中有效地提取结构化信息。然而，与解码器模型相比，编码器的发展较为缓慢，导致生物医学和临床环境中的领域适应性有限。我们介绍了 BioClinical ModernBERT，这是一种领域适应的编码器，它基于现代 BERT 的最新版本，结合了长上下文处理，并在速度和性能方面对生物医学和临床 NLP 做出了重大改进。BioClinical ModernBERT 通过在迄今为止最大的生物医学和临床语料库上继续预训练而开发，该语料库包含超过 535 亿个标记，并解决了之前临床编码器的关键限制，即依赖于来自单一来源的数据，而是利用来自不同机构、领域和地理区域的 20 个数据集。在涵盖广泛使用案例的四个下游任务中，它优于现有的生物医学和临床编码器。我们发布了 BioClinical ModernBERT 的基础版（150M 参数）和大版（396M 参数），以及支持进一步研究的训练检查点。

1 引言

随着临床信息数据库变得更加全面，自然语言处理（NLP）方法在支持下游医疗保健任务中的应用激增，这些任务从临床决策支持和护理服务分析到临床试验的患者群体选择（Cho et al., 2024; Abdel-Jaber et al., 2022）。这一趋势正值深度学习模型的快速演变，特别是基于 transformer 的架构，它们在生成和判别 NLP 任务中都表现出色（Min et al., 2024）。

基于编码器的架构通常使用双向注意力处理整个输入序列，因其能够捕捉丰富的上下文表示，传统上更适用于非生成任务，如分类和范围提取。相比之下，基于解码器的模型依赖于自回归解码，每次生成一个标记，同时只关注之前生成的标记。此结构使其非常适合开放式文本生成（Nielsen et al., 2025）。虽然基于编码器的方法曾经占主导地位，但最近的研究越来越集中在基于解码器的模型上，基于编码器的架构相对受到的关注较少（Shool et al., 2025）。这导致解码器被用于适合编码器的非生成任务，采用结构化输出生成等变通方法（Geng et al., 2025）。社区对基于解码器模型的偏好至少部分可归因于它们支持较长输入长度以及具有更好的泛化能力和零样本能力；这些优势主要驱动于更大的工程投资和扩展努力。尽管如此，编码器在许多应用中仍然是必要的，并与大型语言模型结合使用，例如在检索增强生成（RAG）框架中（Lewis et al., 2020）。

ModernBERT（Warner et al., 2024）的发布代表了编码器架构的一个重大进步，引入了更长的上下文窗口、提高的效率、扩展的词汇表以及大幅增加的训练数据量。这些改进在临床领域尤为显著。ModernBERT 具有最长可达 8,192 词元的上下文窗口，并可以通过旋转位置嵌入（RoPE）（Su et al., 2023）扩展，能够在一次处理过程中处理完整的临床记录和文档，消除了将其分割成更小块的必要性。与 BERT 的 30,000 词相比，50,368 的扩展词汇量支持学习更精确的词元嵌入，这对捕获临床和生物医学术语的多样性和复杂性尤其有益。

https://github.com/lindvalllab/BioClinical-ModernBERT

在这项工作中，我们对 ModernBERT 模型进行两步连续的预训练，采用生物医学和临床语料库，提供基础版和大版两个版本，并公开发布训练好的模型及其训练检查点。我们展示了显著的改进，与现有的编码器相比，在多个基准测试中实现了最先进的性能，同时支持长上下文输入，并在各种输入分布中保持最高的整体效率。我们的方法利用了有史以来最大的用于生物医学或临床编码器的训练数据集，并结合了多种临床文本来源，以补充传统依赖的 MIMIC 数据 (Johnson et al., 2016, 2023) 。值得注意的是，当 MIMIC 用通用去识别标签替换受保护的健康信息（PHI）时，限制了模型学习实体特定上下文的有意义表示的能力，我们纳入的几个其他数据集使用现实的替代标识符，这使得 BioClinical ModernBERT 能够在这些实体周围学习更自然的表示。

2 相关工作

虽然仅有解码器的模型已获得显著的流行程度，但仅有编码器的架构由于其在性能和效率之间的良好权衡，尤其是在信息检索（IR）和 RAG 管道中，仍被广泛使用。它们较低的推理成本和高效处理文档的能力使其在大规模应用中具有吸引力。尽管许多现有的管道仍依赖于像 BERT 这样的旧模型，这些模型受限于短上下文窗口、过时的训练语料库和不太高效的设计选择，最近的努力如 MosaicBERT、CrammingBERT 和 GTE-en-MLM 已引入了更新的架构和更长的上下文支持。然而，这些模型主要专注于检索任务或高效训练，而非更广泛的下游性能。ModernBERT 通过将训练数据扩展到两万亿个标记、引入改进的架构选择，并在更广泛的任务范围内提供具有竞争力的性能，解决了这些限制中的许多问题。

最近在自然语言处理（NLP）方面的进展已经促使开发出适合领域的 Transformer 模型，这些模型显著地增强了临床文本的提取和解释。在这些模型中，BioBERT 和 Clinical BERT 分别成为生物医学和临床信息学的关键工具。BioBERT 通过在大规模生物医学文献上的预训练扩展了原始 BERT 架构，使其能够更准确地理解领域特定语言的语义。Clinical BERT 在此基础上，通过在 MIMIC-III 数据集的临床笔记上的额外预训练，使其更好地捕捉临床文档的细微差别。依赖于 RoBERTa 架构的 BioMed-RoBERTa 采取了一种不同的方法，通过在 Semantic Scholar 上的训练，作为其生物医学文本的来源。这些模型在概念提取、时态关系识别和结果预测等任务中表现优于通用领域编码器，使它们成为临床 NLP 流水线中的重要组成部分。

长上下文 长文本理解在临床自然语言处理领域尤其重要，因为临床文档的结构、术语和长度通常变化很大，常常超过许多基于编码器的模型的标准 512-令牌输入限制 (Rule et al., 2021) 。处理较长序列的能力使模型能够捕获和整合分散在整个临床笔记甚至多个文档中的相关信息。这种扩展的上下文有助于识别短上下文模型通常错过的模式和关系。因此，长文本编码器在一系列临床自然语言处理任务上表现出色，例如表型分析、队列选择和医学实体识别。 Li et al. (2022) 引入的 Clinical-Longformer 和 Clinical-BigBird 就是这种情况的例证，目前它们是唯一可公开使用的能够处理长序列长度的临床模型。

3 方法

3.1 预训练

3.1.1 生物医学数据集

我们使用 PubMed 摘要和 PMC 全文文章作为生物医学文本的来源。通过各自的 API，我们收集了总计 50.7B 的生物医学文本 token 用于预训练。

3.1.2 临床数据集

自从 Clinical BERT、Clinical Longformer 和 Clinical BigBird 发布以来，随着 MIMIC-IV (Johnson et al., 2023) 的引入，可用的临床文本领域已显著发展，这增加了模型训练可用的临床标记的数量。我们将 MIMIC-IV 纳入到我们的临床预训练语料中，并通过从各种机构、国家和临床背景中收集额外的数据集来进一步多样化它。这解决了现有临床编码器的一个关键限制，即往往依赖于单一机构作为其临床文本的来源。更广泛和多样化的数据集基础应能改善推广性，并更好地捕捉跨机构护理实践的多样性，以及个人临床医生之间的风格和结构差异。根据最近由 Wu et al. 的评论，我们精心挑选了一组 20 个公开可用的、非合成的临床文本数据集。由于 ModernBERT 专门在英语文档上训练，我们将选择范围限制在包含英语文本的数据集。进一步的细节，包括数据集的原产国、临床来源和临床背景，详见 Table 1 。这些临床数据集共计 28 亿标记。

正如 Alsentzer et al. (2019) 所指出，目前临床编码器的一个已知限制是在去识别任务中的表现较差，这主要是由于在 MIMIC-III 中使用了占位符标识 PHI。为了解决这个问题，我们引入了诸如 CheXpert Plus (Chambon et al., 2024) 之类的数据集，这些数据集不使用遮蔽标识，而是以合成替代品替换 PHI。

最后，尽管 MIMIC-III 和 MIMIC-IV 有少量重叠，我们不预计这会对预训练产生不利影响。我们采用高达 30 % 的掩码语言模型（MLM）概率，遵循在 Warner et al. (2024) 中的实现，并在整个预训练策略中对临床数据集训练 6 个周期。作为比较，RoBERTa (Liu et al., 2019) 使用 15 % 的 MLM 概率训练了 40 个周期。因此，我们不认为训练会受到这种冗余的影响，也不寻求排除重叠样本。

3.1.3 训练计划

ModernBERT 使用了一个名为“升温-稳定-衰减（WSD）”的学习率调度器进行训练，该调度器由三个阶段组成：一个逐步增加学习率的升温阶段，一个将学习率维持在最高值以加速训练的稳定阶段，以及一个降低学习率以确保训练数据收敛的衰减阶段。这个调度器允许在稳定阶段的任何检查点上继续对新数据进行预训练，而不会出现冷重启问题，并允许在不需要新的升温阶段的情况下重用稳定阶段的学习率，最终通过衰减阶段在新的目标域实现最佳性能。

这种训练范式特别适合通过持续预训练进行领域适应：从稳定阶段的检查点恢复训练可以高效地在新的领域特定数据上进行训练而不会重新引入预热不稳定性，而衰减阶段则有助于针对下游领域的专门化。我们在 BioClinical ModernBERT 的开发中应用了这一策略，通过受 Clinical BERT 背后方法启发的两阶段训练程序。在这种情况下，模型从 BioBERT 初始化，该模型是在 PubMed 和 PMC 上训练的，并使用 MIMIC-III 适应临床语言。同样，我们旨在通过首先从 ModernBERT 的衰减前检查点恢复稳定阶段的训练，在生物医学和临床语料的混合上训练，再通过衰减阶段专门化于临床数据，以平衡广泛的生物医学知识和临床特异性。

第一阶段涉及在生物医学和临床数据上进行联合预训练，具体来说包括 PubMed、PMC 和 20 个精心整理的临床数据集，共计 1605 亿个标记。在这个阶段我们纳入临床数据，以减轻在第二阶段训练时生物医学知识的灾难性遗忘。

因此，我们从 ModernBERT 基础模型和大型模型的最终稳定阶段检查点初始化 BioClinical ModernBERT，继续用各自的学习率进行训练，基础模型的学习率为 3e-4，大型模型为 5e-5。为了保持与原始训练设置的一致性，我们还采用相同的批处理大小：基础模型为 72，大型模型为 77。

在第二阶段中，我们仅对 20 个临床数据集进一步优化模型。对于这个阶段，我们将 MLM 概率降低到 15 %，这基于来自 Ankner et al. 的研究结果，并通过我们自己的实验确认，相较于 30 % 的掩码比例，这一设置提升了下游性能。

在这一阶段，BioClinical ModernBERT 在临床数据上被训练了三个周期。我们尝试了几种学习率方案，发现基础模型在三个周期中应用了 1− 的平方根衰减时，能获得最佳下游结果。对于大型模型，其最佳性能通过在前两个周期使用恒定学习率，在最后一个周期应用 1− 的平方根衰减而实现。

我们在第一阶段之后还训练了一个单独的模型，仅使用生物医学数据额外训练了 50.7B 个 tokens，采用 1− 平方根衰减学习率调度。然而，与在临床数据上训练的变体相比，这个模型在我们的下游评估中 consistently 表现不佳。虽然我们在论文中不再详细讨论这个模型，但我们在 Hugging Face 上公开发布了该模型，命

Name	Country	Clinical Source	Clinical Context	Samples	Tokens (M)
ACI-BENCH (Yim et al., 2023)	US	Clinical Notes	Not Reported	207	0.1
ADE Corpus (Gurulingappa et al., 2012)	Several	Clinical Notes	Not Reported	20,896	0.5
Brain MRI Stroke (Kim et al., 2019)	Korea	Radiology Reports	Neurology	2,603	0.2
CheXpert Plus (Chambon et al., 2024)	US	Radiology Reports	Pulmonology	223,460	60.6
CHIFIR (Rozova et al., 2023)	Australia	Pathology Reports	Hematology / Oncology	283	0.1
CORAL (Sushil et al., 2024)	US	Progress Notes	Hematology / Oncology	240	0.7
Eye Gaze CXR (Karargyris et al., 2021)	US	Radiology Reports	Pulmonology	892	0.03
Gout Chief Complaints (Osborne et al., 2020)	US	Chief Complaint	Internal Medicine	8,429	0.2
ID-68 (Anazi et al., 2017)	UK	Clinical Notes	Psychology	78	0.02
Inspect (Huang et al., 2023)	US	Radiology Reports	Pulmonology	22,259	2.8
MedNLI (Romanov and Shivade, 2018)	US	Clinical Notes	Internal Medicine	14,047	0.5
MedQA (Jin et al., 2020)	US	National Medical Board Examination	Not Reported	14,366	2.0
MIMIC-III (Johnson et al., 2016)	US	Clinical Notes	Internal Medicine	2,021,411	1,047.7
MIMIC-IV Note (Johnson et al., 2023)	US	Clinical Notes	Internal Medicine	2,631,243	1,765.7
MTSamples (MTSamples, 2018)	Not Reported	Clinical Notes	Internal Medicine	2,358	1.7
Negex (Chapman et al., 2013)	US	Discharge Summaries	Not Reported	2,056	0.1
PriMock57 (Korfiatis et al., 2022)	UK	Simulated Patient Care	Internal Medicine	57	0.01
Q-Pain (Logé et al., 2021)	US	Clinical Vignettes	Palliative Care	51	0.01
REFLACX (Bigolin Lanfredi et al., 2022)	US	Radiology Reports	Pulmonology	2,543	0.1
Simulated Resp. Interviews (Fareez et al., 2022)	Canada	Simulated Patient Care	Pulmonology	272	0.6

Table 1: 用于预训练的临床数据集，标记数量以百万计。

名为 Bio ModernBERT 基底 和 大的 。

3.1.4 计算资源

BioClinical ModernBERT 是在配备 8 个 NVIDIA H100 SXM5 GPU 的服务器上训练的。持续的预训练过程大约花费了四天用于基础模型，八天用于大型模型，不包括数据预处理和消融研究所花费的时间。

3.2 下游评估

3.2.1 模型

我们将我们的模型与流行的临床和生物医学编码器进行了比较。我们提供了关于每个模型所用训练数据的总结，以及在 Appendix A 中的相应标记计数。

临床-Longformer 和临床-BigBird 和 (Li et al., 2022) 是唯一可用的长上下文临床编码器，序列长度为 4096 个标记。它们分别基于 Longformer (Beltagy et al., 2020) 和 BigBird (Zaheer et al., 2021) 架构，并在 MIMIC-III 的临床笔记上进一步训练。

BioBERT (Lee et al., 2019) 在 BERT-base (Devlin et al., 2019) 上初始化，并在 PubMed 和 PMC 数据上进行训练。我们使用托管在 Hugging Face 上的 biobert-v1.1 模型 。

BioMed-RoBERTa (Gururangan et al., 2020) 基于 RoBERTa-base (Liu et al., 2019)，并在来自 Semantic Scholar (Kinney et al., 2025) 的科学论文上进行了训练。

临床 BERT (Alsentzer et al., 2019) 引入了几个模型。我们使用由 Hugging Face 托管的 Bio_ClinicalBERT 模型 ，对应于在 MIMIC-III 的临床记录上使用 BioBERT 训练的版本。

在撰写本文时，

临床现代 BERT 已经可以在 Lee et al. (2025) 中获得。作者在 PubMed 摘要、MIMIC-IV 临床笔记以及 ICD 代码与其描述的对上微调了 ModernBERT base。临床 ModernBERT 是在 ModernBERT 的后衰减版本上用 128 标记序列训练的，这可能限制了其捕捉长程依赖的能力。相比之下，我们使用 8192 标记输入来训练基础版和大型版，使用了更加庞大和多样化的语料库，并根据 ModernBERT 使用的 WSD 计划继续从衰减前检查点开始预训练。

我们还将我们的模型与 ModernBERT-base 和 ModernBERT-large 进行比较，以展示持续预训练所带来的附加价值。

3.2.2 任务

我们在五个数据集上评估模型，一个是多标签分类任务，一个是单标签分类任务，三个是命名实体识别（NER）任务。我们通过使用五个不同的种子微调模型来评估它们的性能，并报告在测试集上的中位数分数。这些模型在早停条件下训练 10 个周期，批量大小为 16，权重衰减为 $1e^{-5}$ 。根据标准做法，我们通过对每个下游任务进行网格搜索来确定基础模型和大型模型的学习率 (Lee et al., 2019; Alsentzer et al., 2019; Li et al., 2022) 。使用的数值可以在 Appendix B 中找到。微调和评估的实现可以在我们的 GitHub 仓库 中找到。

对于分类任务，我们使用加权 F1 分数来衡量性能。对于命名实体识别任务，我们使用由 seqeval 框架提供的 F1 分数 (Nakayama, 2018) 。

分类任务

ChemProt (Krallinger et al., 2017) 是一个数据集，包含来自 PubMed 的 1,820 个化学-蛋白质相互作用的摘要。我们从 BLUE 基准代码库 (Peng et al., 2019) 中收集了数据。根据他们的实现，相关任务是一个针对六个类别的单标签分类问题：CPR:3、CPR:4、CPR:5、CPR:6、CPR:9 和无。

表型 (Moseley et al., 2020) 为 MIMIC-III 的 2,270 份患者笔记提供了关于 15 种表型（包括无和不确定）的存在或不存在的注释。每个笔记由两位专家人工注释者进行注释。由于“无”类别对应于所有表型指标均不存在的情况（即，一个零向量），因此该任务被视为针对剩余 14 种表型的多标签分类问题。

命名实体识别任务

状态变化 (COS) 或胸部 X 光片句子中状态改变的临床事件 (Klassen et al., 2014) 是一个由 1,344 份 UW Harborview Medical Center 去身份化的胸部 X 光片记录中提取出的 1008 个句子的语料库。该数据集为定义的临床事件提供了 4 个实体的注释：解剖位置、位置属性、属性的可能值以及属性的状态改变。这些注释是由两位人类专家生成的。相关任务对应于这四种临床事件的标记级别的多标签分类。

社会历史 或社会历史部分的生活方式和环境因素 (Yetisgen and Vanderwende, 2017) 是一个包含来自 MTSamples 的 364 份临床笔记中社会历史部分的语料库。这些部分中的每一个都由人工标注者注释了三种类型的物质滥用：烟草、酒精和药物，针对每个事件有 7 种实体类

型：状态、类型、方法、数量、频率、暴露历史和戒断史。此外，该数据集包含针对以下因素的注释：职业、婚姻状况、家庭信息、居住、生活状况、环境暴露、身体活动、体重管理、性历史、传染病史。相关任务对应于这些社会历史因素的词级多标签分类。

去标识化医学文本（DEID） (Neamatullah et al., 2008) 是一个由 MIMIC-II 数据库中的 2,434 篇护理笔记构成的语料库 (Saeed et al., 2011)，这些笔记由三位或更多专家注释了 PHI 实例。该数据集提供两个版本：一个版本中，PHI 实例被替换为真实的替代数据，另一个版本中，PHI 实例被替换为对应的标签。使用这些标签，我们将 PHI 实例分类为以下 PHI 之一：年龄、日期、地点、姓名、电话号码或其他（相应的脚本可以在我们的库中找到）。模型随后会接收到包含真实替代数据的文本。相关任务对应于这些类型的 PHI 的标记级别多标签分类。

我们预处理每个 NER 数据集，以 BIO 格式来构建对应的任务 (Ramshaw and Marcus, 1995)。实现的细节可以在我们的 GitHub 仓库 中找到。每个数据集的统计信息可以在 Appendix C 中找到。

3.2.3 推理速度

我们采用了在 Warner et al. (2024) 中描述的推理速度评估方法。为了模拟实际使用情况，我们构建了六个合成数据集，每个数据集中包含 8,192 个文档。前三个数据集由固定长度的样本组成，每个文档分别包含 512、4,096 和 8,192 个标记，对应于所评估模型支持的三种上下文长度。为了评估去填充的影响，我们另外生成了三个变长数据集，其中每个样本的标记数遵循正态分布，中心值为最大序列长度的一半：分别为 256、2,048 和 4,096 个标记。每个数据集的推理速度取 10 次运行的平均值。

3.2.4 计算资源

在下游基准数据集上进行微调是在单个 H100 PCIe GPU 上完成的，而推理速度评估是在配备单个 A100 40GB GPU 的机器上进行的。

4 结果

4.1 性能

Table 2 展示了各种生物医学和临床编码器在前一节描述的任务中的性能。在五个任务中的四个上，BioClinical ModernBERT base 和 large 的表现优于其他模型。

在分类任务中，BioClinical ModernBERT large 在 ChemProt 上以 90.8 % 的 F1 得分以及在 Phenotype 上以 60.8 % 的得分达到了最先进的结果。基础模型也优于所有其他基础模型，在 ChemProt 上取得 89.9 % 的成绩，在 Phenotype 上取得 58.1 % 的成绩。在命名实体识别中，基础模型在社会历史上实现了最先进的性能，并在 DEID 上优于其他基础模型。large 版本在 DEID 上实现了最先进的结果，而两个 BioClinical ModernBERT 模型在 COS 上保持竞争力，取得 95.1 % 的成绩。

为了评估临床专业化后的生物医学知识保留情况，我们在 ChemProt 上测试了第一阶段的检查点。ChemProt 是一个包含 PubMed 摘要的数据集，用作生物医学领域知识的代理。基础模型取得了 90.2 % 的 F1 分数，而大模型得分为 90.5 %。在第二阶段后，基础模型得分为 89.9 %，而大模型提高到 90.8 %，超过了其第一阶段的分数。相比之下，BioBERT 在 ChemProt

	Model	Context length	Classification		Named Entity Recognition		
			ChemProt	Phenotype	COS	Social History	DEID
Base	BioBERT (Lee et al., 2019)	512	89.5	26.6	94.9	55.8	74.3
	Clinical BERT (Alsentzer et al., 2019)	512	88.3	25.8	95.0	55.2	74.2
	BioMed-RoBERTa (Gururangan et al., 2020)	512	89.0	36.8	94.9	55.2	81.1
	Clinical-BigBird (Li et al., 2022)	4096	87.4	26.5	94.0	53.3	71.2
	Clinical-Longformer (Li et al., 2022)	4096	74.2	46.4	<u>95.2</u>	56.8	82.3
	Clinical ModernBERT (Lee et al., 2025)	8192	86.9	54.9	93.7	53.8	44.4
	ModernBERT - base (Warner et al., 2024)	8192	89.5	48.4	94.0	53.1	78.3
	BioClinical ModernBERT - base (ours)	8192	<u>89.9</u>	<u>58.1</u>	95.1	<u>58.5</u>	<u>82.7</u>
Large	ModernBERT - large (Warner et al., 2024)	8192	90.2	58.3	94.4	54.8	82.1
	BioClinical ModernBERT - large (ours)	8192	90.8	60.8	95.1	57.1	83.8

Table 2: 模型在分类和命名实体识别任务上，基于 5 次随机种子的中位数性能。整体最佳得分用粗体表示，基础模型的最佳得分用下划线标出。

上的得分为 89.5 %，而 Clinical BERT 得分为 88.3 %。

4.2 推理速度

推断速度测试结果在 Table 3 中展示。在评估的基础模型中，ModernBERT 表现出最高的整体效率。虽然像 BERT 和 RoBERTa 这样的短上下文编码器在固定输入大小为 512 个标记的数据集上表现良好，分别实现了每秒 98.1 和 97.4 千标记，ModernBERT 仍然提供了具有竞争力的吞吐量，达到 76.2 kTok/s，并在同一任务上超越了其他长上下文模型，如 BigBird（71.5 kTok/s）和 Longformer（55.5 kTok/s）。在所有其他数据集类型中，ModernBERT 始终提供最佳性能，保持在所有输入长度上的处理速度相对稳定：在固定数据集中，512 个标记时为 76.2 kTok/s，4096 个标记时为 73.8 kTok/s，8192 个标记时为 71.1 kTok/s。相反，随着输入长度增加，BigBird 的吞吐量在固定数据集中显著下降，从 71.5 kTok/s 下降到 50.1 kTok/s。此外，ModernBERT 在具有可变长度序列的数据集上表现出显著更强的性能，512 标记的数据集达到 75.1 kTok/s，8192 标记的数据集达到 73.7 kTok/s。

尽管由 Lee et al. 提出的模型在 ModernBERT 基础上初始化，但在我们的推理速度测试中表现出的行为与 ModernBERT 架构不一致。对于包含固定长度序列的数据集，在 512 个令牌的输入时速度更快。然而，当输入长度增加时吞吐量显著下降，其性能在中等和长序列上比 ModernBERT 差。对于包含可变长度序列的数据集，表现较固定长度数据集更差，建议没有使用去填充。进一步检查模型的配置文件后，似乎它遵循的是一个上下文长度为 8192 个令牌的 BERT 架构，这解释了我们观察到的效率结果。

5 讨论

BioClinical ModernBERT base 和 large 在分类和命名实体识别任务中表现出强大而一致的性能，优于现有的临床和生物医学编码器。在 Phenotype 上观察到的显著性能提升可以归因于模型对临床语言的改进适应能力以及其利用该数据集中存在的显著更长输入序列的能力（参见 Appendix C ）。

值得注意的是，我们的结果表明，即使在第二阶段针对临床数据进行专业化之后，BioClinical ModernBERT 仍然保留了大量的生物医学知识。这可以通过在阶段 1 和阶段 2 之间 ChemProt 性能的变化来证明：基础模型略下降 0.3 分，而大型模型则提高了 0.3 分。这些发现表明，在预训练的第一阶段引入临床数据有助于防止灾难性遗忘，使 BioClinical ModernBERT 能够在后续的临床适应过程中保持生物医学知识。相比之下，Clinical BERT 在 ChemProt 上的得分比其仅有生物医学数据的前身 BioBERT 低 1.2 分，这表明在其第二阶段的临床训练后，其生物医学性能有所下降。

我们也发现我们的方法解决了 Clinical BERT 作者之前指出的一个挑战：临床嵌入模型在去标识化任务上的性能下降，这通常归因于数据集中使用了通用的 PHI 掩码标记，如 MIMIC。我们在 DEID 任务上的结果支持这样的假设：在拥有多种 PHI 处理策略的多样化临

	Model	Short		Medium		Long	
		Fixed	Variable	Fixed	Variable	Fixed	Variable
Base	BioBERT (Lee et al., 2019)	98.1	49.3	–	–	–	–
	Clinical BERT (Alsentzer et al., 2019)	98.1	49.3	–	–	–	–
	BioMed-RoBERTa (Gururangan et al., 2020)	97.4	48.6	–	–	–	–
	Clinical-BigBird (Li et al., 2022)	71.5	35.7	50.1	25.0	–	–
	Clinical-Longformer (Li et al., 2022)	55.5	27.8	53.2	26.6	–	–
	Clinical ModernBERT (Lee et al., 2025)	86.2	43.1	44.4	19.8	28.8	12.3
	ModernBERT - base (Warner et al., 2024)	76.2	75.0	73.8	74.9	71.1	73.7
	BioClinical ModernBERT - base (ours)	76.2	75.1	73.8	75.0	71.1	73.7
Large	ModernBERT - large (Warner et al., 2024)	25.8	26.5	25.3	25.6	24.8	25.2
	BioClinical ModernBERT - large (ours)	25.8	26.5	25.3	25.6	24.8	25.2

Table 3: 在每秒处理的千字标记数（kTok/s）中进行推理速度测试，并取 10 次运行的平均值。每个类别的最佳成绩以粗体显示。破折号表示不支持的配置。

床数据集上进行训练，可以增强模型识别和泛化受保护健康信息的能力。特别是，我们语料库中的一些数据集使用合成但逼真的代理来替代 PHI，而不是简单地使用掩码标记，这可能有助于保留语义结构。这种 PHI 表示的变化似乎在微调过程中提高了模型对 PHI 的理解，从而在去标识化任务上获得更强的下游性能。

除了在任务上表现出色之外，我们的结果还显示，BioClinical ModernBERT 是唯一一个在不同长度的序列和输入分布中始终保持高计算效率的临床编码器。虽然短上下文编码器在具有短固定长度输入的数据集上运行时间更快，这部分是由于它们的参数数量较少（BERT 基准为 110M，RoBERTa 基准为 125M，而 ModernBERT 基准为 150M），但在这种场景中，BioClinical ModernBERT 仍然具有竞争力，并且在输入长度增加时表现出明显的优势。这得益于其交替注意力机制，其中三分之二的注意力层仅限于局部滑动窗口，而不是全序列的全局注意力，从而大大减少了复杂性。

对变长输入效率的进一步提升是通过 ModernBERT 的去填充机制实现的，该机制在推理过程中动态排除填充标记。与对统一填充序列进行注意力计算的标准变压器实现不同，去填充通过忽略填充标记来确保仅对实际内容标记执行计算，从而减少内存使用并加快推理过程。该属性在输入长度随着记录类型和患者接触情况显著变化的临床 NLP 中尤其有价值，这在使用静态填充时会导致相当大的低效。

局限性和未来工作 我们的工作有几个限制。首先，尽管我们试图通过使用各种临床数据集为 BioClinical ModernBERT 提供良好的临床泛化能力，但 MIMIC-III 和 MIMIC-IV 仍然构成了临床训练语料库的 95 % 以上；这限制了我们能够实现的真实临床数据多样性。其次，由于 ModernBERT 仅在英语上训练，我们决定仅使用英文临床笔记来训练 BioClinical ModernBERT 。因此，我们的工作不能直接应用于其他语言。最后，关于基准数据集，i2b2 数据集 (Uzuner et al., 2007, 2011; Sun et al., 2013; Stubbs et al., 2015; Stubbs and Uzuner, 2015) 通常用于基准测试临床编码器。尽管我们计划在这些数据集上对 BioClinical ModernBERT 进行基准测试以与文献一致，但在本项目期间这些数据集不可用。我们计划在这些数据集再次可用时立即发布 BioClinical ModernBERT 的基准测试。

6 结论

在这项工作中，我们介绍了 BioClinical ModernBERT ，这是一种专为临床和生物医学文本设计的长上下文编码器。基于 ModernBERT 的基础并利用其训练计划，我们采用了一个由两个阶段组成的持续预训练策略，将大规模生物医学语料库与多样且广泛的公共临床数据集合相结合，这些数据涵盖了来自多个国家、机构和临床环境的 20 个数据集。据我们所知，这构成了迄今为止用于训练编码器的最大生物医学和临床预训练语料库。

我们发布了 BioClinical ModernBERT 的基础版和大型版，使研究人员和实践者能够根据自己的使用场景在计算效率和性能之间进行选择。我们的预训练设计解决了之前临床模型的一个关键限制，即几乎完全依赖于 MIMIC-III，从而在临床自然语言处理任务中实现了更好的泛化效果。通过在训练早期整合生物医学和临床数据，并利用长上下文建模能力，BioClinical ModernBERT 在一系列广泛的临床基准测试中实现了最先进的性能。

此外，BioClinical ModernBERT 支持长上下文输入，最长可达 8,192 个标记。这在医学 NLP 中尤为关键，因为重要信息常常分布在长序列中，而临床推理依赖于捕捉长距离依赖和文档级上下文。尽管有这种扩展的上下文，模型依然保持高度高效，即使在长输入长度下也能提供快速推理。

我们发布了代码、模型和训练检查点，以支持进一步的研究。

7 致谢

我们由衷感谢斯坦福医学与成像人工智能中心以及 PhysioNet (Goldberger et al., 2000) ，感谢他们提供本研究中使用的多个临床数据集的访问权限。我们还感谢 Orion Weller 在解决与标记化过程相关的问题时给予的宝贵帮助。

8 代码可用性声明

我们在 Hugging Face 上提供了 BioClinical ModernBERT 模型
（基底 ，大 ）和 训练检查点 ，并在 GitHub 上提供了 源代码 以实现可重复性。

References

Hussein Abdel-Jaber, Disha Devassy, Azhar Al Salam, Lamya Hidaytallah, and Malak EL-Amir. 2022. A Review of Deep Learning Algorithms and Their Applications in Healthcare. *Algorithms*, 15(2):71.

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. Association for Computational Linguistics.

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T. Asi, Saud Alsahli, Amal Alhashem, Hanan E. Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, Firdous M. Abdulwahab, Mais Hashem, Nadia Alhashmi, Fathiya Al Murshedi, Adila Al Kindy, Ahmad Alshaer, Ahmed Rumayyan, Saeed Al Tala, Wesam Kurdi, Abdulaziz Alsaman, Ali Alasmari, Selina Banu, Tipu Sultan, Mohammed M. Saleh, Hisham Alkuraya, Mustafa A. Salih, Hesham Aldhalaan, Tawfeg Ben-Omran, Fatima Al Musafri, Rehab Ali, Jehan Suleiman, Brahim Tabarki, Ayman W. El-Hattab, Caleb Bupp, Majid Alfadhel, Nada Al Tassan, Dorota Monies, Stefan T. Arold, Mohamed Abouelhoda, Tammaryn Lashley, Henry Houlden, Eissa Faqeih, and Fowzan S. Alkuraya. 2017. Expanding the genetic heterogeneity of intellectual disability. 136(11):1419–1429.

Zachary Ankner, Naomi Saphra, Davis Blalock, Jonathan Frankle, and Matthew Leavitt. 2024. Dynamic Masking Rate Schedules for MLM Pretraining. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 477–487, St. Julian's, Malta. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. 2022. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. 9(1):350.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats.

Wendy W. Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E. Chapman, Mike Conway, Melissa Tharp, Danielle L. Mowery, and Louise Deleger. 2013. Extending the NegEx Lexicon for Multiple Languages. In *MEDINFO 2013*, pages 677–681. IOS Press.

Ha Na Cho, Tae Joon Jun, Young-Hak Kim, Heejun Kang, Imjin Ahn, Hansle Gwon, Yunha Kim, Jiahn Seo, Heejung Choi, Minkyoung Kim, Jiye Han, Gaeun Kee, Seohyun Park, and Soyoung Ko. 2024. Task-Specific Transformer-Based Language Models in Health Care: Scoping Review. *JMIR Medical Informatics*, 12(1):e49724.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. 9(1):313.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Planem Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. 45(5):885–892.

Suchin Gururangan, Ana Marasovi, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, and Jason A. Fries. 2023. INSPECT: A Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. 10(1):1.

Alistair E. W. 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. 3(1):160035.

Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T. Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A. Krupinski, and Mehdi Moradi. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. 8(1):92.

Chulho Kim, Vivienne Zhu, Jihad Obeid, and Leslie Lenert. 2019. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. 14(2):e0212778.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2025. The Semantic Scholar Open Data Platform.

Prescott Klassen, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen. 2014. Annotating Clinical Events in Text Snippets for Phenotype Detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2753–2757, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations.

Martin Krallinger, Obdulia Rabal, and Anália Lourenço. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop,*, 141-146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. 2025. Clinical ModernBERT: An efficient and long context encoder for biomedical text.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences.

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.

Cécile Logé, Emily Ross, David Yaw Amoah Dadey, Saahil Jain, Adriel Saporta, Andrew Y. Ng, and Pranav Rajpurkar. 2021. Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2024. Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Computing Surveys*, 56(2):1–40.

Edward T. Moseley, Joy T. Wu, Jonathan Welt, John Foote, Patrick D. Tyler, David W. Grant, Eric T. Carlson, Sebastian Gehrmann, Franck Dernoncourt, and Leo Anthony Celi. 2020. A Corpus for Detecting High-Context Medical Conditions in Intensive Care Patient Notes Focusing on Frequently Readmitted Patients. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1362–1367. European Language Resources Association.

MTSamples. 2018. Medical Transcriptions.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Ishna Neamatullah, Margaret M. Douglass, Li-wei H. Lehman, Andrew Reisner, Mauricio Villarroel, William J. Long, Peter Szolovits, George B. Moody, Roger G. Mark, and Gari D. Clifford. 2008. Automated de-identification of free-text medical records. 8(1):32.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2025. Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks.

John David Osborne, Tobias O'Leary, Amy Mudano, James Booth, Giovanna Rosas, Gurusai Sujitha Peramsetty, Anthony Knighton, Jeff Foster, Ken Saag, and Maria Ioana Danila. 2020. Gout Emergency Department Chief Complaint Corpora.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain.

Vlada Rozova, Anna Khanina, Jasmine C. Teng, Joanne S. K. Teh, Leon J. Worth, Monica A. Slavin, Karin A. Thursky, and Karin Verspoor. 2023. Detecting evidence of invasive fungal infections in cytology and histopathology reports enriched with concept-level annotations. 139:104293.

Adam Rule, Steven Bedrick, Michael F. Chiang, and Michelle R. Hribar. 2021. Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center. *JAMA Network Open*, 4(7):e2115334.

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical care medicine*, 39(5):952–960.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, 58 Suppl(Suppl):S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58 Suppl(Suppl):S20–S29.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. RoFormer: Enhanced Transformer with Rotary Position Embedding.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46 Suppl(0):S5–S12.

Madhumita Sushil, Vanessa E. Kennedy, Divneet Mandair, Brenda Y. Miao, Travis Zack, and Atul J. Butte. 2024. CORAL: Expert-Curated medical Oncology Reports to Advance Language Model Inference. 1(4).

Ozlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association: JAMIA*, 14(5):550–563.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Jiageng Wu, Xiaocong Liu, Minghui Li, Wanxin Li, Zichang Su, Shixu Lin, Lucas Garay, Zhiyun Zhang, Yujie Zhang, Qingcheng Zeng, Jie Shen, Changzheng Yuan, and Jie Yang. 2024. Clinical text datasets for medical artificial intelligence and large language models —a systematic review. *NEJM AI*, 1(6):AIra2400012.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic Identification of Substance Abuse from Social History in Clinical Text. In *Artificial Intelligence in Medicine*, volume 10259, pages 171–181. Springer International Publishing.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. 10(1):586.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences.

A 编码器训练语料库

Table 4 总结了每个模型的训练数据构成，涵盖临床、生物医学以及其他来源，并在可获得的情况下提供了标记计数。这个细分显示了用于训练 BioClinical ModernBERT （我们的方法）的数据的多样性和规模，与之前的模型相比。

Model	Clinical		Biomedical		Other		Total
	Source	Tokens	Source	Tokens	Source	Tokens	(B tokens)
BioBERT	—	—	PubMed + PMC	18.0	Wikipedia + BookCorpus	3.3	21.3
Clinical BERT	MIMIC-III	1.0	PubMed + PMC	18.0	Wikipedia + BookCorpus	3.3	22.3
BioMed-RoBERTa	—	—	S2ORC papers	7.6	—	—	7.6
Clinical-Longformer	MIMIC-III	1.0	—	—	—	—	1.0
Clinical-BigBird	MIMIC-III	1.0	—	—	—	—	1.0
Clinical ModernBERT	MIMIC-IV	1.7	PubMed	(N/A)	Medical codes & description	(N/A)	13.0
BioClinical ModernBERT (ours)	20 clinical datasets	2.8	PubMed + PMC	50.7	—	—	53.5

Table 4: 按数据类型划分的训练数据组成（单位：亿标记）。

B 微调学习率

Table 5 报告了在每个下游任务中微调编码器的选定学习率。基础和大型模型变体显示了不同的值。

Task	Base	Large
ChemProt	5e-5	2e-5
Phenotype	8e-5	5e-5
COS	1e-4	1.5e-4
Social History	1.5e-4	2e-4
DEID	7e-5	7e-5

Table 5: 为每个任务和模型大小选择的学习率。

C 下游评估数据集描述

Table 6 总结了数据集的统计信息，包括用于分类数据集的类别数量或用于 NER 数据集的 BIO 标记实体类型的数量、每个分割的样本数量，以及使用 ModernBERT 分词器计算的每个样本的平均标记数。

	Dataset	# Classes / Entities	Samples per Split			Avg. # Tokens		
			Train	Val	Test	Train	Val	Test
Classif.	ChemProt	6	19,460	11,820	16,943	68.2	67.7	75.2
	Phenotype	14	1,589	227	454	3,112.2	3,118.7	3,175.8
NER	COS	11	705	151	152	24.5	24.0	23.4
	Social History	51	254	55	55	55.9	57.0	44.9
	DEID	13	1,703	365	366	276.6	270.0	281.5

Table 6: 数据集特征：类别或实体类型的数量，每个分割的总样本数，以及每个样本的平均标记数。