arXiv:2304.00958v2 [cs.CL] 4 May 2023

DrBERT：面向生物医学和临床领域的法语鲁棒预训练模型

YanisLabrak[*1,4] **Adrien Bazoge**[*2,3] **Richard Dufour**[2] **Mickael Rouvier**[1]**Emmanuel Morin**[2] **BéatriceDaille**[2] **Pierre-AntoineGourraud**[3]1LIA，阿维尼翁大学 2LS2N，法国国家科学研究中心联合研究单位 6004，南特大学3数据诊所，南特大学附属南特大学医院 4Zenidoc {firstname.lastname}@univ-avignon.fr {firstname.lastname}@univ-nantes.fr

摘要

近年来，预训练语言模型（PLMs）在广泛的自然语言处理（NLP）任务中取得了最佳性能。早期模型是在通用领域数据上训练的，而后续出现了专门模型，能够更有效地处理特定领域。本文对法语医学领域的预训练语言模型展开了原创性研究。我们首次比较了基于网络公开数据和医疗机构私有数据训练的预训练语言模型的性能。我们还在一组生物医学任务上评估了不同的学习策略。特别是，我们表明，可以利用现有的外语生物医学预训练语言模型，在目标数据上进一步进行预训练。最后，我们发布了首批面向法语生物医学领域的专用预训练语言模型，称为DrBERT；同时还发布了用于训练这些模型、规模最大的自由许可证医疗数据语料库。

1 引言

近年来，预训练语言模型（BERT）已被证明能显著提升许多自然语言处理（NLP）任务的性能。近年来，BERT（Devlin 等人，2018年）和 RoBERTa（Liu 等人，2019年）等模型越来越多地借助近期的无监督方法，利用海量未标注数据，例如基于 Transformer 架构的掩码语言模型（Vaswani 等人，2017年）。这些预训练语言模型大多常在通用领域语料库上进行预训练，例如新闻文章、书籍或百科全书。还可以通过额外的微调步骤，将这些预训练语言模型应用于特定任务（Devlin 等人，2018年）。

尽管这些通用模型应用于各种场景，但近期研究表明，要实现最优性能（例如金融（Yang 等人，2020年）、医疗（Yang 等人，2021年）或旅游（Zhu 等人，2021年）等专业领域），只有使用针对目标条件进行适配的预训练语言模型才能达到这一效果。

语言模型适配特定领域通常有两种策略。第一种是从头开始训练新模型，仅使用目标专业领域的文本数据。第二种方法称为持续预训练（Howard 和 Ruder，2018年），即继续训练已经过预训练的模型，使其从通用模型转变为专用模型。尽管研究表明，第一种策略通常能带来更好的性能（Lee 等人，2019年），但第二种策略所需资源的限制要严格得多（Chalkidis 等人，2020年；El Boukkouri 等人，2022年），无论是计算资源还是数据量方面均是如此。

然而，领域专用数据通常难以获取，因此可用的专业预训练语言模型寥寥无几。对于英语以外的语言，这一难题尤为突出。在医学领域，临床护理过程中产生的数据蕴含着医学推理的细微之处，且数量最为庞大。然而，受患者隐私限制，这些数据很少能够被获取。

本文介绍并免费发布 DrBERT，这是首个专为法语生物医学领域打造的、基于 RoBERTa 的预训练语言模型，以及用于训练这些模型的语料库。我们还提出了一项原创研究，评估医学领域不同的语言模型预训练策略，并将其与基于临床私有数据构建的模型 ChuBERT 进行比较。在实验中，使用公开可用的生物医学数据训练的预训练语言模型，其表现可与使用医院报告收集的高度专业化私有数据训练的模型相当，甚至更好；也可优于仅使用通用数据、规模更大的语料库。我们的贡献可概括如下：

[*]贡献相同。

- 我们建立了一个新的基准，汇集了一组法语医学领域的自然语言处理任务，从而能够从句法和语义层面评估语言模型（多标签分类、词性标注、命名实体识别等）。

- 一个从多个生物医学在线来源爬取的大型文本数据集，名为 NACHOS。

- 基于 RoBERTa架构构建并评估首批面向生物医学领域的法语开源预训练语言模型 DrBERT，其中包括对不同预训练策略的分析。

- 一组模型使用公开和私有数据进行训练，且训练数据规模相当。随后，我们通过评估这些模型在广泛的公开和私有任务上的表现，对其进行了比较。

- 在开源 MIT许可证下免费发布公开预训练语言模型，并在 CC0 1.0许可证下发布 NACHOS语料库[1]。

2 相关工作

BERT（Devlin 等人，2018年）是一种基于掩码语言模型概念、使用双向Transformer预训练的上下文化词表示模型（Vaswani 等人，2017年）。自发布以来，它在几乎所有自然语言处理任务上都取得了最先进的（SOTA）性能，同时只需进行极少的任务专用架构修改。

然而，由于每项训练目标的复杂性以及所需数据量庞大，此类模型在计算方面的训练成本很高。因此，新的方法不断涌现，提出了更有效的预训练方式。其中一种方法是 RoBERTa (Liu 等人，2019年)。为了改进最初的 BERT 模型，作者对其训练流程做了一些简单的设计调整。他们修改了掩码语言建模策略，采用动态掩码；移除了下一句预测任务；大幅增加批次大小；并在更长的训练周期中使用了显著更多的数据。如今，RoBERTa 是许多自然语言处理任务和语言领域的标准模型，包括法语领域的 CamemBERT 模型 (Martin 等人，2020年)。

近年来，研究人员通过对 Transformer 架构进行无监督预训练，开发了多个面向生物医学与临床领域的语言模型，主要针对英语。其中最早的模型之一是 BioBERT（Lee 等人，2019年），它基于最初预训练的 BERT 模型，并通过持续预训练，进一步利用生物医学专用数据进行预训练。其他模型，如 BlueBERT（Peng 等人，2019年）和 ClinicalBERT（Huang 等人，2019年），也采用了这一方法，并使用了多种数据来源。若有足够的领域内数据，另一种方法是从头开始预训练模型（如 SciBERT（Beltagy 等人，2019年）、PubMed-BERT（Gu等人，2021年）等）。请注意，SciBERT 使用来自生物医学和计算机科学领域的混合领域数据进行训练，而 PubMed-BERT 仅使用生物医学数据进行训练。Gu 等人（2021 年）根据在 BLURB 基准测试任务上取得的结果，对使用混合领域数据进行预训练的益处提出了质疑。

除英语以外的其他语言中，基于 BERT 的模型要少得多，且主要依赖持续预训练。例如德语（Shrestha，2021年）、葡萄牙语（Schneider 等人，2020年）和瑞典语（Vakili等人，2022年）。只有西班牙语（Carrino等人，2021年）和土耳其语（Türk-men等人，2022年）模型使用来自多个来源的生物医学和临床数据从头开始训练。据我们所知，目前没有专为生物医学领域构建的法语公开模型。

3 预训练数据集

在生物医学领域，关于预训练语言模型的先前研究（Gu等人，2021年）强调了将训练所用的数据来源与目标下游任务相匹配的重要性。由于医疗数据涉及敏感信息（例如保护用户数据和患者的受保护健康信息），因此极难获取。大规模收集与该领域相关的网络数据，似乎是一种能够弥补这一不足的解决方案。然而，这些网络文档的质量参差不齐。此前尚未有人比较过基于网络中特定领域数据训练的预训练语言模型，与基于临床数据仓库中的私人文档训练的预训练语言模型；后者的质量可以得到控制。

[1] https://drbert.univ-avignon.fr/

我们为法语提取了两个不同的医疗数据集。第一个数据集汇集了从各种免费在线来源爬取的数据，第二个数据集则包含来自南特大学医院的私人住院报告。

表1 概述了收集的两个语料库。第3.1节详细介绍的公开网络数据构成了一个名为 NACHOS 的语料库$_{large}$，其中包含 7.4 吉字节数据。名为 NBDW 的私有数据集$_{small}$ 将在第 3.2 节中介绍，包含 4 吉字节数据。为了进行可比实验，我们从 NACHOS 中提取了一个与私有数据规模相同的子语料库（NACHOS$_{small}$）。最后，第3.3 节介绍了对两个数据集所做的预处理。

语料库	Size	#词	#句子
NACHOS$_{large}$ (公开)	7.4 吉字节	11 亿	5420 万
NACHOS$_{small}$ （公开）	4 GB	6.46 亿	2530 万
NBDW$_{small}$ （私有）	4 GB	655 M	43.1 M
NBDW$_{mixed}$ （两者）	4+4 吉字节	1.3 B	68.4 M

表1：公开（NACHOS）和私有（NBDW）采集数据集概览。

3.1 公开语料库 - NACHOS

我们介绍法语开放式医疗保健网络爬取语料库（NACHOS），这是一个通过爬取各种医学主题相关文本来源编纂而成的法语医学开源数据集。它包含超过十亿词，取自24个法语地区的优质网站。该语料库涵盖丰富多样的医学信息：疾病与病症的描述、治疗方法与药物信息、一般健康建议、官方科学会议报告、匿名临床病例、科学文献、学位论文、法语翻译对、大学健康课程，以及大量从原始文本来源、网页抓取和光学字符识别（OCR）获得的数据。表2 总结了 NACHOS 的不同数据来源。

我们采用启发式方法将文本切分为句子，并严格过滤掉简短或质量较低的句子，例如通过 OCR 获得的句子。最后，我们使用基于多语言 Opus EMEA（Tiedemann 和 Nygaard，2004）和 MAS- SIVE（FitzGerald 等，2022年）语料库训练的自有分类器，将这些句子分类为不同语言，仅保留法语句子。

对于 4 吉字节版本的 NACHOS（NACHOS$_{small}$），我们打乱了整个语料库，并随机选取了 2530 万个句子，以最大限度地提高数据源同质性。完整的 NACHOS语料库现已免费线上获取[2]。

资源名称	# 词
HAL	638,508,261
法国国家卫生管理局（HAS）	113,394,539
药品说明书	74,770,229
医疗网站抓取	60,561,495
向法国国家食品、环境及职业健康安全署提交的申请	51,372,932
公共药品数据库 (BDPM)	48,302,695
ISTEX	44,124,422
CRTT	26,210,756
WMT-16	10,282,494
EMEA-V3	6,601,617
维基百科 生命科学 法语	4,671,944
Cochrane 摘要	3,932,526
ANSES RCP	2,953,045
Cerimes	1,717,552
Vidal	410,313
LiSSa	235,838
DEFT-2020	231,396
CLEAR	225,898
CNEDiMTS	175,416
QUAERO 法语医学语料库	72,031
ANSM临床研究注册库	47,678
ECDC	44,482
QualiScope	12,718
WMT-18-Medline	7,673
总计	**1,088,867,950**

表 2：NACHOS语料库的来源。

3.2 私有语料库 - NBDW

该私有语料库称为南特生物医学数据仓库（NBDW），是利用南特大学医院的数据仓库构建的。该数据仓库包含与患者相关的不同维度数据：社会人口统计信息、药物处方，以及与会诊或住院记录相关的其他信息（诊断、生物学检查、影像学检查等）。2018年，CNIL（法国国家信息与自由委员会）批准了 NBDW 数据集的实施和使用；该委员会是负责执行国家和欧洲数据隐私保护法律的法国独立监管机构，授权编号为 N°2129203。

本研究从数据仓库中随机选取并提取了 170 万份去标识化住院报告。如表3所示，这些报告来自不同的医院科室，其中急诊医学、妇科和门诊护理最为常见。

每份报告都被拆分为词元序列，每个序列平均包含 15.26 个词。随后，

[2]https://drbert.univ-avignon.fr/

将所有报告中的词元序列打乱后构建语料库。该语料库包含 6.55 亿个词，来自 4310 万个句子，总大小约为 4 吉字节。

医学专科	# 文档	# 词
其他	474,588	192,832,792
急诊医学	235,579	90,807,406
门诊护理	119,149	50,975,472
会诊	95,135	38,335,804
妇科	132,983	38,204,495
心脏病学	29,633	22,654,583
肿瘤内科	45,603	22,587,869
胃肠病学	46,600	21,340,794
骨科	82,084	18,983,791
血液学	41,776	18,285,983
重症医学	20,819	16,472,785
耳鼻喉科学	69,343	16,131,214
皮肤病学	51,804	15,035,412
风湿病学	31,527	14,647,543
泌尿外科学	51,535	14,272,231
结直肠外科	45,987	13,334,550
内科学	23,904	13,282,253
精神病学	26,628	12,496,503
神经外科学	34,481	10,360,533
肾脏病学	19,171	9,548,533
眼科学	19,700	4,464,515
总计	**1,698,029**	**655,055,061**

表 3：NBDW语料库的来源。

3.3 预处理步骤

提供的文本数据已使用SentencePiece（Kudo和Richardson，2018年）扩展方法拆分为子词单元。该方法结合了字节对编码（BPE）（Sennrich等人，2016年）和WordPiece（Wu等人，2016年），无需预分词（在单词或词元级别进行），从而避免了使用特定语言分词器的要求。我们采用了包含32k个子词词元的词汇表。对于每个从头开始预训练的模型（见第4.2节），分词器均使用预训练数据集中的所有句子构建。

4 模型预训练

本节从两个角度介绍所研究模型的预训练方式：1）所用数据的影响（规模和性质）；2）模型的预训练策略。这两个方面分别在第4.1 节和4.2节中详细说明。第4.3 节则介绍现有的最先进水平预训练模型，并将其用于比较。

4.1 数据的影响

一个问题是确定创建一个性能良好、能够与通用领域训练的模型相竞争的模型所需的数据量。近期研究，例如Zhang 等人(2020年) 和Martin 等人(2020年)，探讨了预训练数据规模对模型性能的影响。根据这些研究，有些任务在数据较少时表现更好，而另一些任务（如常识知识和推理任务）则会随着预训练数据的增加而持续提升性能。

在医学领域，尚无研究比较预训练期间领域特定数据量变化的影响，也没有评估数据质量是否会因其采集来源不同而有所差异。

因此，我们提出在第 $_{small}3_{small}$ 节所述的 NACHOS 或 NBDW 语料库上，对若干语言模型进行预训练并开展评估。此外，我们提出一个在 NACHOS^{large} 上预训练的模型，以探究数据量几乎翻倍是否能提升模型性能。最$_{small}$后，我们探索结合公开的 NACHOS$_{small}$ 和 NBDW 数据源，总计 8 GB（NBDW$_{mixed}$），以验证在低资源领域中结合私有数据与公开数据是否可行。

4.2 预训练策略

除了分析数据规模和来源外，我们还旨在评估医学领域预训练语言模型的三种训练策略：

- 从头开始训练完整模型，包括子词分词器。
- 在我们的医学专用数据上，继续对面向法语的先进语言模型 Camem-BERT 进行预训练，同时保留初始分词器。
- 在我们的法语数据上，继续对面向医学领域的先进语言模型 PubMedBERT 进行预训练；该模型使用英语，但此处保留初始分词器。

关于最后一种策略，我们的目标是比较在我们的法语医疗数据上进一步预训练的英语医学模型，与基于通用法语模型构建的另一个模型的性能。事实上，不同语言的医学领域共享许多术语，因此混合使用两种语言的资源是可行的。

表4 总结了本文评估的所有配置，涵盖了对数据规模和预训练策略的研究。

模型名称	策略	语料库
DrBERT	从头开始	NACHOS$_{large}$
DrBERT	从头开始	NACHOS$_{small}$
ChuBERT	从头开始	NBDW$_{small}$
ChuBERT	从头开始	NBDW$_{mixed}$
CamemBERT	持续预训练	NACHOS$_{small}$
PubMedBERT	持续预训练	NACHOS$_{small}$
CamemBERT	持续预训练	NBDW$_{small}$

表4：所研究的预训练模型配置列表。

模型架构 所有从头开始预训练的模型均采用 CamemBERT 基础配置，该配置与 RoBERTa 基础架构相同（12 层、768 个隐藏维度、12 个注意力头、1.1 亿个参数）。由于资源有限，我们没有训练模型的大型版本。

语言建模 我们使用 HuggingFace 库（Wolf 等人，2019年）在掩码语言建模（MLM）任务上训练模型。该任务会随机将序列中的一部分词元替换为特殊词元，并要求模型使用交叉熵损失对其进行预测。在 BERT 和 RoBERTa 模型（包括 CamemBERT ）中，15% 的词元会被随机选中。在这些被选中的词元中，80% 会被替换为 <掩码>词元，10% 保持不变，另有 10% 会被随机替换为词汇表中的一个词元。训练我们的模型时，我们将这一掩码概率保持在 15%。

优化与预训练 我们对模型进行了 8 万步优化，批次大小为 4,096 个序列，每个序列包含 512 个词元，每步可处理 210 万个词元。学习率先线性预热 1 万步，从零逐渐升至初始 5×10^{-5} 学习率。模型在 Jean Zay 超级计算机上使用 128 张 Nvidia V100 32 吉字节 GPU 训练 20 小时。我们采用混合精度训练（FP16）（Micikevi- cius 等人，2017年）以减少内存占用，从而将每张 GPU 上的批次大小增至 32 个序列。

4.3 基线模型

我们介绍了比较研究中用作基线的若干现有预训练模型。

CamemBERT （Martin 等人，2020年) 是一款基于 RoBERTa 的模型，完全从法语版 OSCAR 语料库（138 吉字节）从零开始预训练。在我们的研究中，该模型是比较实验结果的主要基准，因为它是目前法语领域的最先进模型。我们还使用 CamemBERT 的 4 吉字节模型变体，以比较数据性质和数量的影响。

PubMedBERT （Gu等人，2021年）是一种基于 BERT 、专用于生物医学领域的模型，完全从头开始在包含 31 亿个词的 PubMed 语料库（21 GB）上进行预训练。

ClinicalBERT （Huang等人，2019年）是一种临床专用模型，基于 BERT 分词器和权重，并使用 MIMIC 语料库中的 5 亿个词（3.7 吉字节）进行了进一步预训练。

BioBERTv1.1 （Lee等人，2019年）是一种生物医学专用模型，基于 BERT 分词器和权重，并使用 PubMed 语料库中的 45 亿个词进行了进一步预训练。

5 下游评测任务

为了评估我们模型的不同预训练配置，需要一组医学领域的任务。虽然英语中已有这一自然语言处理领域专用基准（BLURB（Gu等人，2021年）），但法语中尚无此类基准。本节介绍一个原创基准，汇总于表5，其中整合了各种法语自然语言处理医学任务。其中一些任务来自公开数据集（第5.1节），可用于复现我们的实验。其他任务来自私有数据集（第5.2节），无法共享。不过，这些任务有助于更准确地评估我们的模型。

5.1 公开可用的任务

ESSAIS / CAS：法语临床病例语料库 ESSAIS（Dalloux 等人，2021年）和 CAS（Grabar 等人，2018年）语料库分别包含 13,848 例和 7,580 例法语临床病例。部分临床病例附有讨论。全部病例中的一个子集经过了形态句法（词性标注、词形还原）和语义（UMLS 概念、否定、不确定性）标注。在本研究中，我们只关注词性标注任务。

FrenchMedMCQA FrenchMedMCQA语料库（Labrak等，2022年）是一个公开可用的法语医学领域多项选择题问答（MCQA）数据集，包含3,105道来自真实考试的题目，涵盖

主题 / 语料库名称	Task	评估指标	训练集	Dev	Test
公共语料库					
ESSAIS（Dalloux 等人，2021年）	词性标注	宏平均 F1	9,693	2,077	2,078
CAS：法语临床病例语料库（Grabar 等人，2018年）	词性标注	宏平均 F1	5,306	1,137	1,137
MUSCA-DET - 健康社会决定因素抽取（任务1）	嵌套命名实体识别	宏平均 F1	19,861	2,207	5,518
MUSCA-DET - 健康社会决定因素抽取（任务2）	多标签分类	宏平均 F1	19,861	2,207	5,518
QUAERO French Medical Corpus -EMEA(Névéol 等人，2014年）	嵌套命名实体识别	加权 F1	11	12	15
QUAERO French Medical Corpus - MEDLINE（Névéol 等人，2014年）	嵌套命名实体识别	加权 F1	833	832	833
FrenchMedMCQA（Labrak 等人，2022年）	MCQA	完全匹配率 / 汉明评分	2,171	312	622
私有语料库					
医疗报告急性心力衰竭结构化	命名实体识别	宏平均 F1	2,527	281	703
急性心力衰竭 (aHF) 分类	二分类	宏平均 F1	1,179	132	328
技术专科分类	多分类	宏平均 F1	4,413	1,470	1,473
医疗报告处方结构化	命名实体识别	宏平均 F1	61	15	26

表 5：用于评估医学专用模型的语料库、任务和评估指标综述。

法语药学专业文凭考试，包含单选题和多选题。

QUAERO 法语医学语料库QUAERO 法语医学语料库（Névéol 等人，2014年）引入了一个大型生物医学文档语料库，这些文档在实体和概念层级上进行了标注，可用于命名实体识别和分类任务。该语料库涵盖三种文本类型，共计 103,056 个词，来源于 EMEA 或 MEDLINE。人工标注员使用自动预标注结果进行验证，共标注了对应 UMLS 语义组（Bodenreider，2004年）的十类实体。总计有 26,409 个实体标注映射到 5,797 个不同的 UMLS 概念。为简化评估流程，我们按字母顺序排列嵌套标签，并将其合并为一个标签，以便将任务转换为适用于基于 BERT 架构的词元分类格式。

MUSCA-DET MUSCA-DET 是一个法语语料库，包含从南特大学医院生物医学数据仓库临床笔记中“生活方式”部分提取的句子。该语料库包含 27,000 个经过假名化处理的句子，并标注了 26 个与健康的社会决定因素相关的实体（居住状况、婚姻状况、住房、子女、就业、饮酒、吸烟、药物滥用、身体活动）。该语料库包含两项任务：嵌套命名实体识别（NER）和多标签分类。

5.2 私有任务

技术专科分类这项分类任务需要根据医疗报告的转录文本，为一份医学报告分配专科类别。该数据集包含7,356份法语医疗报告，这些报告均由人工标注，并按相同比例抽样——分布于6个专科：精神病学、泌尿外科学、内分泌学、心脏病学、糖尿病学和传染病学。

医疗报告处方结构化（命名实体识别）该任务旨在从一组由法语语音转录文本生成的 100 份长篇医学报告中识别命名实体。命名实体采用 BIO 格式标注，共分为 12 类：*O*、年龄、城市、日期、电子邮件、医院、电话号码、剂量、持续时间、剂型、药物和用药方案。

急性心力衰竭医疗报告结构化（命名实体识别）该语料库包含来自南特大学医院的 350 份住院报告（共分为 3,511 个句子）。这些报告标注了 46 种实体类型，涉及以下临床信息：慢性心力衰竭的病因、急性心力衰竭的诱发因素、糖尿病、吸烟状况、心率、血压、体重、身高、药物治疗、高血压和左心室射血分数。该语料库共包含 6,116 个临床实体。

急性心力衰竭（aHF）分类此任务是根据是否存在急性心力衰竭诊断，对住院报告进行分类。该语料库包含来自南特大学医院的 1,639 份住院报告，这些报告均被标记为急性心力衰竭阳性或阴性。

6 结果与讨论

如前所述，我们在一组与医学领域相关的公开和私有自然语言处理下游任务上，评估了我们针对生物医学领域提出的预训练语言模型的性能。我们首先根据所采用的不同预训练策略分析结果（第6.1节），然后聚焦于其影响

	aHF 命名实体识别			aHF 分类			命名实体识别 医疗报告			专科分类		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 吉字节	40.89	35.22	35.13	81.90	79.12	80.13	87.98	91.66	89.35	99.32	99.09	99.20
CamemBERT OSCAR 4 吉字节	46.32	43.17	42.66	81.49	81.42	81.41	87.79	90.74	88.78	99.53	99.69	99.61
CamemBERT CCNET 4 GB	47.25	42.2	43.11	82.02	79.30	79.98	87.61	92.28	89.34	99.54	99.55	99.55
PubMedBERT	52.61	46.30	47.22	78.17	76.18	76.86	87.07	92.61	89.20	99.25	99.51	99.37
ClinicalBERT	50.11	44.15	44.70	80.13	75.92	77.12	87.04	92.14	88.77	98.58	98.62	98.58
BioBERT v1.1	49.37	47.25	46.01	79.69	78.51	79.00	**88.17**	91.80	89.38	98.59	99.03	98.80
DrBERT NACHOS$_{large}$	55.29	46.66	48.22	81.33	81.25	81.25	87.99	**92.80**	**89.83**	99.82	**99.90**	**99.86**
DrBERT NACHOS$_{small}$	54.55	43.39	45.93	79.85	80.10	79.87	87.57	92.76	89.44	**99.85**	99.85	99.85
ChuBERT NBDW$_{small}$	**56.92**	47.46	49.01	81.03	**82.67**	81.56	87.76	92.63	89.58	99.76	**99.90**	99.83
ChuBERT NBDW$_{mixed}$	54.62	47.81	**49.14**	82.23	81.71	**81.98**	87.42	92.36	89.30	99.81	99.82	99.81
CamemBERT NACHOS$_{small}$	22.02	16.67	16.08	74.86	69.82	69.80	65.72	68.49	66.74	99.44	99.67	99.54
PubMedBERT NACHOS$_{small}$	53.44	**48.21**	48.72	**83.06**	80.39	81.40	87.35	92.69	89.36	99.52	99.58	99.55
CamemBERT NBDW$_{small}$	25.44	19.33	19.12	79.50	74.74	76.02	68.80	71.23	69.64	99.60	99.57	99.58

表6：Pe 我们在私有生物医学下游任务上的表现。最佳模型用粗体表示，次佳模型加下划线 .

预训练数据的规模或性质（第6.2节）。最后，我们通过在通用领域的自然语言处理任务上应用并比较这些领域专用模型，考察其泛化能力（第6.3节）。

请注意，所有预训练语言模型均以相同方式针对所有下游任务进行了微调，且所有报告结果均为四次运行所得分数的平均值。生物医学下游任务的表现分别列于表 6 和 7中，分别对应私有任务和公开任务。为便于阅读，每张表的第一部分列出现有基线模型的结果，第二部分列出我们从头开始训练的专门化模型，最后一部分列出采用持续预训练的模型。

6.1 预训练策略的影响

从表6 和7可以看出，完全从头开始训练的模型（DrBERT NACHOS 和 ChuBERT NBDW）在两类数据来源和任务（即 私有和公开任务）上往往都能取得最佳结果。具体而言，就 F1 分数而言，这些模型在所有私有任务以及几乎所有公开任务中（7 项任务中的 5 项）都取得了最佳结果。其余两项公开任务（MUSCA-DET T2 和 QUAERO-MEDLINE）则由 PubMedBERT NACHOS$_{small}$表现更好；该模型已先在特定领域数据（英语生物医学数据）上进行预训练，随后又使用我们的法语医学数据（NACHOS$_{small}$）进一步预训练。

我们还观察到，从领域通用模型（CamemBERT NACHOS$_{small}$ 或 CamemBERT NBDW$_{small}$）进行持续预训练，无法达到其他特定模型的性能；这两个模型在任何任务中的表现都未能跻身第一或第二。

最后，在通用数据上训练的基线模型（CamemBERT OSCAR）以及在英语生物医学数据上训练的模型（PubMedBERT、ClinicalBERT 和 BioBERT），在少数公开的生物医学任务（CAS POS、FrenchMCQA 或 MUSCA-DET T2）上仍具竞争力；但在私有任务中，它们均未跻身第一或第二。这似乎凸显了使用不匹配数据时私有任务的难度。

6.2 数据的影响

关于用于模型预训练的数据量（$small$ 与 $large$ 或 $mixed$），结果表明，无论预训练策略或数据来源（私有或公开）如何，数据量越大，模型表现越好。然而，对于大多数任务，差异很小；尽管 $small$系统的数据量只有大型模型的一半，但排名通常仅次于大型模型。

我们注意到，在公共任务上，基于网络来源进行预训练的模型明显占据优势，具体包括 OS- CAR 和 NACHOS。事实上，仅依赖私有 NBDW 数据的模型，只有在 MUSCA-DET T1 任务上取得了最佳表现（以 F1分数衡量）。在私有任务上，这一趋势并不明显：基于 NBDW 的模型与公共生物医学数据混合训练后，表现更为可观，甚至更好（ChuBERT NBDW$_{mixed}$），如表 6所示。我们认为，这种差异主要源于所处理数据的性质不同。

最后，我们观察到，基于英语的模型表现与基于法语的 CamemBERT 模型相近。这表明，在特定领域数据上进行预训练是有益的。例如，使用我们的法语专业数据对 PubMedBERT 模型进行持续预训练后，效果更好（PubMedBERT NACHOS$_{small}$），这印证了我们关于跨语言知识迁移有效性的假设，

	MUSCA-DET T1			MUSCA-DET T2			ESSAI POS			CAS POS			FrenchMedMCQA		QUAERO-EMEA			QUAERO-MEDLINE		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	值	汉明损失	完全匹配率	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 吉字节	89.04	88.59	88.54	89.87	87.12	88.20	81.57	81.01	81.10	96.37	94.53	95.22		36.24	**16.55**	90.57	91.06	90.71	76.58	78.67	77.41
CamemBERT OSCAR 4 GB	86.09	85.45	85.43	92.68	90.34	91.27	84.01	83.51	83.69	98.15	95.34	96.42		35.75	15.37	90.75	91.16	90.83	78.55	79.33	78.76
CamemBERT CCNET 4 GB	91.12	89.91	90.33	93.10	90.42	91.38	85.60	85.63	85.42	**98.19**	**96.75**	**97.33**		34.71	14.41	90.31	90.59	90.33	78.06	78.11	77.61
PubMedBERT	93.04	91.45	91.99	84.41	80.60	81.97	88.43	87.93	87.78	97.40	94.86	95.90		33.98	14.14	86.89	87.33	86.79	77.33	77.28	77.09
ClinicalBERT	91.79	89.44	90.36	85.43	81.23	82.95	89.09	88.78	88.24	97.94	95.88	96.73		32.78	14.19	84.91	85.47	84.79	75.56	74.85	75.05
BioBERT 1.1	91.82	89.82	90.46	85.52	80.14	81.91	86.76	84.90	85.18	98.10	96.39	97.12		36.19	15.43	84.55	85.03	84.29	72.62	73.30	72.68
DrBERT NACHOS$_{large}$	92.10	90.27	91.04	**94.97**	90.41	92.24	**90.96**	**89.19**	**89.75**	97.37	94.49	95.65		36.66	15.32	**91.93**	**92.52**	**92.09**	77.85	78.54	77.88
DrBERT NACHOS$_{small}$	93.35	90.62	91.77	91.31	86.60	88.57	90.12	88.37	88.76	97.04	94.88	95.70		**37.37**	13.34	91.54	92.00	91.66	77.91	79.34	78.18
ChuBERT NBDW$_{small}$	**94.88**	90.79	92.23	94.77	90.27	92.17	88.53	87.73	87.71	97.00	94.65	95.61		35.16	14.79	88.11	88.78	88.15	75.05	76.57	74.94
ChuBERT NBDW$_{mixed}$	94.39	**91.93**	**92.73**	94.22	90.02	91.71	86.36	85.50	85.73	97.77	95.30	96.35		34.58	12.21	90.36	90.94	90.52	78.61	79.32	78.63
CamemBERT NACHOS$_{small}$	81.44	81.39	80.96	79.74	78.08	78.70	80.59	79.88	80.04	95.64	91.57	92.46		32.87	13.76	67.56	77.48	71.10	55.45	62.34	57.43
PubMedBERTNACHOS$_{small}$	92.51	91.49	91.53	94.95	**92.55**	**93.62**	84.73	83.80	83.85	97.82	96.12	96.81		35.88	15.21	90.97	91.27	91.03	**82.03**	**81.71**	**81.73**
CamemBERT NBDW$_{small}$	82.35	81.59	81.57	78.14	76.38	77.12	79.44	79.79	79.25	95.98	92.11	93.18		27.73	11.89	53.44	73.11	61.75	48.71	61.33	53.05

表7：公共生物医学下游任务的表现。最佳模型以粗体标示，第二名以下划线标示。

也进一步印证了我们关于跨语言知识迁移有效性的假设。

6.3 通用领域任务上的表现

表8 展示了所有预训练语言模型在通用领域下游任务上的结果。这些任务来自Martin 等人（2020年），他们用这些任务评估 CamemBERT 模型。前四项是词性标注任务（GSD、SEQUOIA、SPOKEN 和 PARTUT），最后一项是自然语言推理任务（XNLI）。

我们的模型在所有任务上的表现均有所下降。降幅最大的是自然语言推理任务：ChuBERT NBDW$_{small}$ 的表现比 CamemBERT 138 GB 低近 13%。我们还观察到，英语专业模型的效率与法语生物医学模型相当。从上述观察可以清楚看出，专门化模型难以泛化到其他任务，但一种语言中捕获的专业信息可以迁移到另一种语言。

7 结论

在这项工作中，我们提出了首个基于 Transformer、采用 RoBERTa架构的法语生物医学和临床语言模型。我们在汇总的多样化私有和公开医学任务集合上，对这些专用模型进行了全面评估。与法语通用模型（CamemBERT）和英语医学模型（BioBERT、PubMedBERT 和 Clinical-BERT）相比，我们开源的 DrBERT 模型在所有医学任务上均提升了最先进水平。此外，我们还证明，利用有限资源（4 吉字节）对从网络爬取的医学文本进行预训练，可以与使用医学报告专用数据训练的模型相媲美，甚至经常超越它们。

结果还表明，在针对法语生物医学下游任务时，在现有的领域特定英语模型（此处为 PubMedBERT）上进行持续预训练，比在法语领域通用模型上进行持续预训练更可行。我们还需要使用更多数据进一步研究这一方法的性能，类似于我们使用 DrBERT NACHOS^{large} 所做的工作。

NACHOS语料库以及预训练的模型和所有预训练脚本[3] 均已在线公开发布，采用 CC0 1.0 开放

[3] https://drbert.univ-avignon.fr/

	GSD	SEQUOIA	SPOKEN	PARTUT	XNLI
CamemBERT OSCAR 138 GB	**98.28**	98.68	97.26	97.70	**81.94**
CamemBERT OSCAR 4 GB	98.14	**99.18**	**97.57**	97.86	81.76
CamemBERT CCNET 4 GB	98.18	98.92	97.20	**97.92**	81.26
PubMedBERT	96.48	96.49	90.00	93.97	73.79
ClinicalBERT	96.49	96.31	89.60	93.17	70.57
BioBERT v1.1	97.32	96.54	91.81	94.52	71.54
DrBERT NACHOS$_{large}$	96.94	98.05	95.92	96.54	72.18
DrBERT NACHOS$_{small}$	97.17	98.21	96.38	96.45	72.86
ChuBERT NBDW$_{small}$	96.45	97.38	94.90	95.83	69.00
ChuBERT NBDW$_{mixed}$	97.18	98.10	96.43	96.33	72.32
CamemBERT NACHOS$_{small}$	97.63	96.90	91.12	94.00	71.26
PubMedBERT NACHOS$_{small}$	97.41	98.71	95.54	97.01	77.35
CamemBERT NBDW$_{small}$	97.55	96.26	89.17	91.34	72.73

表8：在公开领域通用下游任务上的性能。最佳模型以粗体标示，第二佳模型加下划线。

源许可证。

8 伦理考量

关于风险和偏见，所有基于 NACHOS 预训练且可免费获取的模型，可能都会受到Bender等人(2021年) 和Sheng 等人(2021年)的研究中所指出的某些问题的影响，因为 NACHOS 的一些子语料库质量可能低于预期，尤其是非政府来源的子语料库。使用基于 BERT 的生物医学语言模型时，可能会遇到偏见问题，包括公平性、性别化语言、代表性不足和时间准确性。

9 局限性

需要指出我们这项工作存在的一些局限性。首先，评估分词器对模型性能的影响是明智之举，以确保这并非观察到性能提升的主要原因。

此外，本研究无法断定，利用 PubMedBERT 进行持续预训练所观察到的从英语到法语的医学领域迁移能否推广到其他语言或其他领域。

最后，如果使用更多样且数量更大的私人临床数据训练 ChuBERT 模型，或许能显著提升其在私人任务上的性能。

本研究使用了大量计算资源：为训练此处介绍的 7 个模型，耗费了约 18,000 小时的图形处理器计算时间；此外，由于模型配置相关的技术问题和性能不佳，调试又耗费了约 7,500 小时的图形处理器计算时间，总计 25,500 小时。根据 Jean Zay 超级计算机的文档，环境总成本 [4] 相当于 6,604,500 Wh，或 376.45 千克二氧化碳当量；该数值基于 BLOOM 环境成本研究中提及的电网碳强度，该研究同样在 Jean Zay 超级计算机上开展（Luccioni 等人，2022年）。因此，在物质资源有限的情况下，本研究难以复现，也难以迁移至其他语言。

[4]http://www.idris.fr/media/jean-zay/jean-zay-conso-heure-calcul.pdf

致谢

本研究使用了 GENCI-IDRIS（资助编号 2022-AD011013061R1 和 2022-AD011013715）以及 CCIPL（卢瓦尔河地区计算中心）的高性能计算资源。本研究获得 ANR AIBy4（ANR-20-THIA-0011）和 Zenidoc 的资金支持。

参考文献

Iz Beltagy、Kyle Lo 和 Arman Cohan。2019年。Scibert：面向科学文本的预训练语言模型。载于*EMNLP*。

Emily M. Bender、Timnit Gebru、Angelina McMillan- Major 和 Shmargaret Shmitchell。2021年。论随机鹦鹉的危险：语言模型会不会变得太大？载于*2021*年*ACM*公平性、问责制与透明度会议论文集，FAccT '21，第610–623页，美国纽约州纽约。计算机协会。

Olivier Bodenreider。2004。统一医学语言系统（UMLS）：整合生物医学术语。《核酸研究》，32：D267–70。

Casimiro Pio Carrino、Jordi Armengol-Estapé、Asier Gutiérrez-Fandiño、Joan Llop-Palao、Marc Pàmies、Aitor Gonzalez-Agirre 和 Marta Villegas。2021年。面向西班牙语的生物医学与临床语言模型：中等资源场景下领域特定预训练的优势。

IliasChalkidis、ManosFergadiotis、ProdromosMalakasiotis、Nikolaos Aletras 和 Ion Androutsopoulos。2020年。LEGAL-BERT：刚走出法学院的布偶。*CoRR*，abs/2010.02559。

Clément Dalloux、Vincent Claveau、Natalia Grabar、Lucas Emanuel Silva Oliveira、Claudia Maria Cabral Moro、Yohan Bonescki Gumiel 和 Deborah Ribeiro Carvalho。2021年。法语和巴西葡萄牙语生物医学语料库中否定及其作用域检测的监督学习。《自然语言工程》，27(2):181–201。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向Transformer预训练。*CoRR*，abs/1810.04805。

Hicham El Boukkouri、Olivier Ferret、Thomas Lavergne 和 Pierre Zweigenbaum。2022年。重新训练还是从头开始训练？医学领域中BERT预训练策略的比较。载于第*13*届语言资源与评测会议（*LREC*）论文集，第2626–2633页，法国马赛。欧洲语言资源协会。

Jack FitzGerald、Christopher Hench、Charith Peris、Scott Mackie、Kay Rottmann、AnaSanchez、Aaron

Nash、LiamUrbach、VisheshKakarala、RichaSingh、Swetha Ranganath、Laurie Crist、Misha Britan、Wouter Leeuwis、Gokhan Tur 和 Prem Natara- jan。2022年。Massive：一个包含 100 万个示例、涵盖 51 种类型学多样语言的多语言自然语言理解数据集。

NataliaGrabar、VincentClaveau 和 Clément Dalloux。2018年。CAS：法语临床病例语料库。第九届健康文本挖掘与信息分析国际研讨会（*LOUHI*）论文集，第 1–7 页，比利时布鲁塞尔。

Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。2021年。面向生物医学自然语言处理的领域特定语言模型预训练。*ACM* 医疗保健计算学报，3(1):1–23。

Jeremy Howard 和 Sebastian Ruder。2018年。用于文本分类的通用语言模型微调。载于第 *56* 届计算语言学协会年会（*ACL*）论文集，第 328–339 页，澳大利亚墨尔本。计算语言学协会。

Kexin Huang、Jaan Altosaar 和 Rajesh Ranganath。2019年。 Clinicalbert：对临床笔记进行建模并预测医院再入院。

Taku Kudo 和 John Richardson。2018年。Sentencepiece：一种用于神经文本处理的简单、语言无关的子词分词器和反分词器。*CoRR*，abs/1808.06226。

Yanis Labrak、Adrien Bazoge、Richard Dufour、Béatrice Daille、Pierre-Antoine Gourraud、Emmanuel Morin 和 Mickael Rouvier。2022年。FrenchMedM-CQA：一个面向医学领域的法语多项选择问答数据集。载于 第*13*届健康文本挖掘与信息分析国际研讨会（*LOUHI*）论文集，阿布扎比，阿拉伯联合酋长国。

Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。2019年。 BioBERT：一种用于生物医学文本挖掘的预训练生物医学语言表征模型。 生物信息学。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*CoRR*，abs/1907.11692。

Alexandra Sasha Luccioni、Sylvain Viguier 和 Anne-Laure Ligozat。2022年。估算 BLOOM（一种拥有 1760 亿参数的语言模型）的碳足迹。

Louis Martin、Benjamin Muller、Pedro Javier Ortiz Suá rez、Yoann Dupont、Laurent Romary、Éric de la Clergerie、Djamé Seddah 和 Benoît Sagot。2020年。CamemBERT：一种美味的法语语言模型。载于第*58*届计算语言学协会年会论文集（*ACL*），第7203—7219页。计算语言学协会。Paulius Micikevicius、Sharan Narang、Jonah Alben、Gregory F. Diamos、Erich Elsen、David García、BorisGinsburg、MichaelHouston、OleksiiKuchaiev、Ganesh Venkatesh 和 Hao Wu。2017年。混合精度训练。*CoRR*，abs/1710.03740。

Aurélie Névéol、Cyril Grouin、Jérémy Leixa、Sophie Rosset 和 Pierre Zweigenbaum。2014年。QUAERO 法语医学语料库：用于医学实体识别与规范化的资源。

Yifan Peng、Shankai Yan 和 Zhiyong Lu。2019年。生物医学自然语言处理中的迁移学习：在十个基准数据集上评估 BERT 和 ELMo。收录于《*2019*年生物医学自然语言处理研讨会论文集（*BioNLP 2019*）》，第 58–65 页。

Elisa Terumi Rubel Schneider、João Vitor Andrioli de Souza、Julien Knafou、Lucas Emanuel Silva e Oliveira、Jenny Copara、Yohan Bonescki Gumiel、Lucas Ferro Antunes de Oliveira、Emerson Cabr- era Paraiso、Douglas Teodoro 和 Cláudia Maria Cabral Moro Barra。2020年。BioBERTpt——用于临床命名实体识别的葡萄牙语神经语言模型。载于第三届临床自然语言处理研讨会论文集，第 65–72 页，线上。计算语言学协会。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年。使用子词单元进行低频词神经机器翻译。载于计算语言学协会第 *54* 届年会论文集（第 *1* 卷：长篇论文），第 1715–1725 页，德国柏林。计算语言学协会。

Emily Sheng、Kai-Wei Chang、Premkumar Natarajan 和 NanyunPeng。2021年。语言生成中的社会偏见：进展与挑战。

Manjil Shrestha。2021年。医学领域语言模型的开发。莱茵-瓦尔应用科学大学硕士学位论文。

Jörg Tiedemann 和 Lars Nygaard。2004。OPUS语料库——平行且免费：http://logos.uio.no/opus。载于第四届语言资源与评估国际会议（*LREC'04*）论文集，葡萄牙里斯本。欧洲语言资源协会（ELRA）。

Hazal Türkmen，O˘guz Dikenelli、Cenk Eraslan 和 Mehmet Callı。2022年。Bioberturk：探索低资源环境下土耳其生物医学语言模型的开发策略。

Thomas Vakili、Anastasios Lamproudis、Aron Henriks- son 和 Hercules Dalianis。2022年。 下游

使用自动去标识化临床数据预训练的 BERT 模型的任务性能。第十三届语言资源与评测会议论文集，第 4245–4252 页，法国马赛。欧洲语言资源协会。

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser 和 IlliaPolosukhin。2017年。注意力就是你所需要的一切。

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pier- ric Cistac, Tim Rault, Rémi Louf, Morgan Funtow- icz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest 和 Alexander M. Rush。2019年。HuggingFace 的 transformers：最先进水平的自然语言处理。

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin John- son, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rud- nick, Oriol Vinyals, Greg Corrado, Macduff Hughes 和 Jeffrey Dean。2016年。谷歌的神经机器翻译系统：弥合人与机器翻译之间的差距。

Xi Yang、Aokun Chen、NimaPourNejatian、Hoo Chang Shin、Kaleb E. Smith、Christopher Parisien、Colin Compas、Cheryl Martin、Anthony B. Costa、Mona G. Flores、Ying Zhang、Tanja Magoc、Christopher A. Harle、Gloria Lipori、Duane A. Mitchell、William R. Hogan、Elizabeth A. Shenkman、Jiang Bian 和 Yonghui Wu。2022年。用于电子健康记录的大型语言模型。*npj Digital Medicine*，5(1):194。

Yi Yang、Mark Christopher Siy UY 和 Allen Huang。2020年。Finbert：用于金融通信的预训练语言模型。

Yian Zhang、Alex Warstadt、Haau-Sing Li 和 Samuel R. Bowman。2020年。何时需要数十亿词的预训练数据?

Hongyin Zhu、Hao Peng、ZhihengLyu、LeiHou、Juanzi Li 和 Jinghui Xiao。2021年。TravelBERT：将领域特定的异构知识融入统一表示的语言模型预训练。

A 附录

A.1 词汇表互覆盖

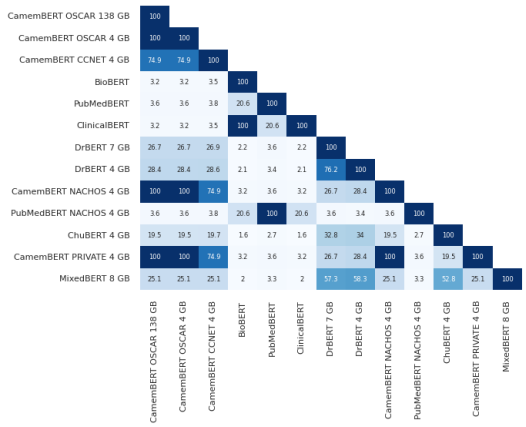
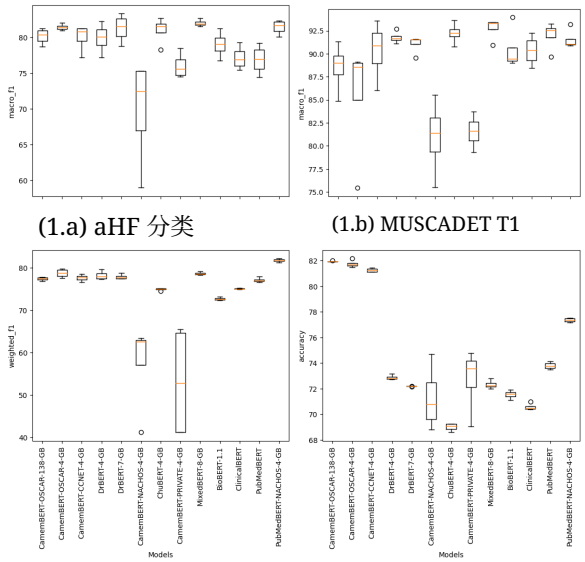


图 1：词汇表互覆盖矩阵。

如图3所示，尽管模型的性能相近，但部分模型之间并没有太多共同词汇。

A.2 模型稳定性

我们在评估阶段观察到，大多数基于 CamemBERT OSCAR 138 GB 持续预训练策略的模型，在微调过程中一致性和稳定性较差，导致不同运行之间的性能出现波动。



(1.a) aHF 分类　(1.b) MUSCADET T1

(2.a) QUAERO MEDLINE　(2.b) XNLI

图 2：每个模型的箱线图。

我们还注意到，在 PubMedBERT NACHOS$_{small}$ 预训练期间，模型

损失在几乎整个预训练持续时间内总体保持稳定，直到第 71,000 步时开始下降，并在第 72,500 步降至零。

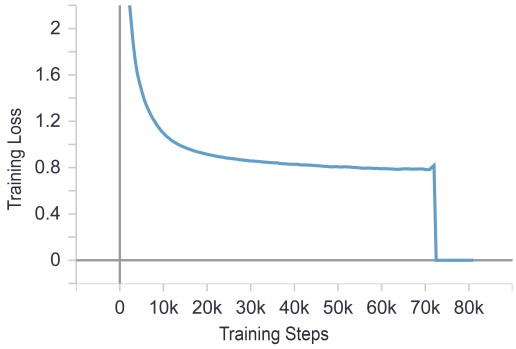


图 3：PubMedBERT NACHOS$_{small}$ 损失。