DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains

Yanis Labrak[*1,4] **Adrien Bazoge**[*2,3] **Richard Dufour**[2] **Mickael Rouvier**[1]
Emmanuel Morin[2] **Béatrice Daille**[2] **Pierre-Antoine Gourraud**[3]
[1]LIA, Avignon Université
[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[3]Clinique des données, CHU de Nantes, Nantes Université [4]Zenidoc
{firstname.lastname}@univ-avignon.fr
{firstname.lastname}@univ-nantes.fr

Abstract

In recent years, pre-trained language models (PLMs) achieve the best performance on a wide range of natural language processing (NLP) tasks. While the first models were trained on general domain data, specialized ones have emerged to more effectively treat specific domains. In this paper, we propose an original study of PLMs in the medical domain on French language. We compare, for the first time, the performance of PLMs trained on both public data from the web and private data from healthcare establishments. We also evaluate different learning strategies on a set of biomedical tasks. In particular, we show that we can take advantage of already existing biomedical PLMs in a foreign language by further pre-train it on our targeted data. Finally, we release the first specialized PLMs for the biomedical field in French, called DrBERT, as well as the largest corpus of medical data under free license on which these models are trained.

1 Introduction

During the last years, pre-trained language models (PLMs) have been shown to significantly improve performance on many Natural Language Processing (NLP) tasks. Recent models, such as BERT (Devlin et al., 2018) or RoBERTa (Liu et al., 2019), are more and more taking advantage of the huge quantities of unlabeled data thanks to recent unsupervised approaches like masked language models based on the Transformers architecture (Vaswani et al., 2017). Most of these PLMs are frequently pre-trained on general domain corpora, such as news articles, books or encyclopedia. An additional step of fine-tuning can also be applied to use these PLMs on a targeted task (Devlin et al., 2018).

Although these generic models are used in various contexts, recent works have shown that optimal performance in specialized domains, such as finance (Yang et al., 2020), medical (Yang et al., 2022) or traveling (Zhu et al., 2021), can only be achieved using PLMs adapted to the targeted conditions.

The adaptation of language models to a domain generally follows two strategies. The first is the training *from scratch* of a new model using only textual data of the targeted specialty. The second approach, called *continual pre-training* (Howard and Ruder, 2018), pursues the training of already pre-trained models, allowing them to pass from a generic model to a specialized one. Even if studies have shown that the first strategy generally offers better performance (Lee et al., 2019), the second requires a much more constrained number of resources (Chalkidis et al., 2020; El Boukkouri et al., 2022) whether in terms of computing resources or of amount of data.

However, domain specific data are generally difficult to obtain, resulting in quite a few specialized PLMs available. This difficulty is even greater for languages other than English. For the medical domain, data produced during clinical care contain the finesse of medical reasoning and are the most prevalent in terms of quantity. However, they are rarely accessible due to patient privacy constraints.

In this paper, we describe and freely disseminate DrBERT, the first RoBERTa-based PLMs specialized in the biomedical field in French, and the corpus that had allowed their trainings. We also propose an original study concerning the evaluation of different language model pre-training strategies for the medical field, while comparing it with our model derived from clinical private data, called ChuBERT. In our experiments, PLMs with publicly available biomedical data can result in similar or better performance compared to highly specialized private data collected from hospital reports or to larger corpora having only generic data. Our contributions can be summarized as follows:

- A new benchmark aggregating a set of NLP

[*] Equal contribution.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 16207–16221
July 9-14, 2023 ©2023 Association for Computational Linguistics

tasks in the medical field in French has been set up, making it possible to evaluate language models at the syntactic and semantic level (multi-label classification, part-of-speech tagging, named entity recognition, etc.).

- A large textual data collection, called NACHOS, crawled from multiple biomedical online sources.
- The construction and evaluation of the first open-source PLMs in French for the biomedical domain based on RoBERTa architecture, called DrBERT, including the analysis of different pre-training strategies.
- A set of models using both public and private data trained on comparable data sizes. These models were then compared by evaluating their performance on a wide range of tasks, both public and private.
- The free distribution of the NACHOS corpus and of the public PLMs under the open-source MIT license[1].

2 Related work

BERT (Devlin et al., 2018) is a contextualized word representation model based on the concept of masked language model and pre-trained using bidirectional Transformers (Vaswani et al., 2017). Since its release, it obtains state-of-the-art (SOTA) performance on almost every NLP tasks, while requiring minimal task-specific architectural modifications.

However, the training cost of such a model is very high in terms of computation due to the complexity of each training objective and the quantity of data needed. Consequently, new methods emerge and propose more effective ways of performing pre-training. One of them is RoBERTa (Liu et al., 2019). In order to improve the initial BERT model, the authors made some simple design changes in its training procedure. They modify the masked-language strategy to perform dynamic masking, remove the next sentence prediction task, increase dramatically the batch sizes and use significantly more data during a longer training period. Nowadays, RoBERTa is the standard model for a lot of NLP tasks and languages, including French with CamemBERT model (Martin et al., 2020).

Recently, multiple language models have been developed for biomedical and clinical fields through unsupervised pre-training of Transformer-based architectures, mainly for English language. One of the first models was BioBERT (Lee et al., 2019), which is based on the initially pre-trained BERT model and further pre-trained using biomedical-specific data through continual pre-training. Other models like BlueBERT (Peng et al., 2019) and ClinicalBERT (Huang et al., 2019) also used this approach on various data sources. An alternative method, when enough in-domain data is available, is to directly pre-train models from scratch (SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2021), etc.). Note that SciBERT was trained on mixed-domain data from biomedical and computer science domains, while PubMedBERT on biomedical data only. Gu et al. (2021) disputed the benefits of mixed-domain data for pre-training, based on results obtained on tasks from BLURB benchmark.

In other languages than English, BERT-based models are much rarer and primarily rely on continual pre-training. Examples include German (Shrestha, 2021), Portuguese (Schneider et al., 2020), and Swedish (Vakili et al., 2022). Only the Spanish (Carrino et al., 2021) and Turkish (Türkmen et al., 2022) models were trained from scratch with biomedical and clinical data from various sources. For French, there is, to our knowledge, no publicly available model specifically built for the biomedical domain.

3 Pre-training datasets

In the biomedical domain, previous works (Gu et al., 2021) on PLMs highlighted the importance of matching the data sources used for its training to the targeted downstream tasks. Due to their sensitive nature (protection of user data, protected health information of patients, etc.), medical data are extremely difficult to obtain. Massive collection of web data related to this domain appears to be a solution that can overcome this lack. However, these web documents vary in terms of quality. No comparison has been made between PLMs based on specific domain data from the web and those on private documents from clinical data warehouses, whose quality can be controlled.

We extracted two different medical datasets for French. The first one gathers data crawled from a variety of free-of-use online sources, and the

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

second one private hospital stays reports from the Nantes University Hospital.

Table 1 gives a general overview of the two collected corpora. The public web-based data, detailed in Section 3.1, allowed the constitution of a corpus, called NACHOS$_{large}$, containing 7.4 GB of data. The private dataset, called NBDW$_{small}$ is described in Section 3.2 and contains 4 GB of data. In order to perform comparable experiments, we extracted a NACHOS sub-corpus (NACHOS$_{small}$) of the same size as the private data. Finally, Section 3.3 describes the pre-processing applied to both datasets.

Corpus	Size	#words	#sentences
NACHOS$_{large}$ (pub.)	7.4 GB	1.1 B	54.2 M
NACHOS$_{small}$ (pub.)	4 GB	646 M	25.3 M
NBDW$_{small}$ (private)	4 GB	655 M	43.1 M
NBDW$_{mixed}$ (both)	4+4 GB	1.3 B	68.4 M

Table 1: Overview of the public (NACHOS) and private (NBDW) collected datasets.

3.1 Public corpus - NACHOS

We introduce the opeN crAwled frenCh Healthcare cOrpuS (NACHOS), a French medical open-source dataset compiled by crawling a variety of textual sources around the medical topic. It consists of more than one billion words, drawn from 24 French speaking high-quality websites. The corpus includes a wide range of medical information: descriptions of diseases and conditions, information on treatments and medications, general health-related advice, official scientific meeting reports, anonymized clinical cases, scientific literature, thesis, French translation pairs, university health courses and a large range of data obtained from raw textual sources, web scrapping, and optical character recognition (OCR). Table 2 summarizes the different data sources of NACHOS.

We use heuristics to split the texts into sentences and aggressively filter out short or low-quality sentences like those obtained from OCR. Finally, we classified them into languages by using our own classifier trained on the multilingual Opus EMEA (Tiedemann and Nygaard, 2004) and MASSIVE (FitzGerald et al., 2022) corpora to keep only the sentences in French.

For the 4 GB version of NACHOS (NACHOS$_{small}$), we shuffled the whole corpus and selected randomly 25.3M sentences to maximize data sources homogeneity. The full NACHOS corpus is now freely available online[2].

Resource name	**# words**
HAL	638,508,261
Haute Autorité de Santé (HAS)	113,394,539
Drug leaflets	74,770,229
Medical Websites Scrapping	64,904,334
ANSES SAISINE	51,372,932
Public Drug Database (BDPM)	48,302,695
ISTEX	44,124,422
CRTT	26,210,756
WMT-16	10,282,494
EMEA-V3	6,601,617
Wikipedia Life Science French	4,671,944
ANSES RCP	2,953,045
Cerimes	1,717,552
LiSSa	235,838
DEFT-2020	231,396
CLEAR	225,898
CNEDiMTS	175,416
QUAERO French Medical Corpus	72,031
ANSM Clinical Study Registry	47,678
ECDC	44,482
QualiScope	12,718
WMT-18-Medline	7,673
Total	**1,088,867,950**

Table 2: Sources of the NACHOS corpus.

3.2 Private corpus - NBDW

The private corpus, called Nantes Biomedical Data Warehouse (NBDW), was obtained using the data warehouse from Nantes University Hospital. This data warehouse includes different dimensions of patients' related data: socio-demographic, drug prescriptions and other information associated with consultation or hospital stays (diagnosis, biology, imagery, etc.). The authorization to implement and exploit the NBDW dataset was granted in 2018 by the CNIL (*Commission National de l'Informatique et des Libertés*), the French independent supervisory authority in charge of application of national and European data privacy protection laws; authorization N°2129203.

For this work, a sample of 1.7 million de-identified hospital stays reports was randomly selected and extracted from the data warehouse. As described in Table 3, the reports are from various hospital departments, emergency medicine, gynecology and ambulatory care being the most frequent.

Each reports was split into tokens sequence with an average of 15.26 words per sequence. Then, all tokens sequences from all reports were shuffled to build the corpus. This corpus contains 655M words, from 43.1M sentences, for a total size of approximately 4 GB.

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

Medical Specialty	**# documents**	**# words**
Other	474,588	192,832,792
Emergency Medicine	235,579	90,807,406
Ambulatory Care	119,149	50,975,472
Consultation	95,135	38,335,804
Gynecology	132,983	38,204,495
Cardiology	29,633	22,654,583
Medical Oncology	45,603	22,587,869
Gastroenterology	46,600	21,340,794
Orthopaedic Surgery	82,084	18,983,791
Hematology	41,776	18,285,983
Critical Care Medicine	20,819	16,472,785
Otolaryngology	69,343	16,131,214
Dermatology	51,804	15,035,412
Rheumatology	31,527	14,647,543
Urology	51,535	14,272,231
Colon and Rectal Surgery	45,987	13,334,550
Internal Medicine	23,904	13,282,253
Psychiatry	26,628	12,496,503
Neurosurgery	34,481	10,360,533
Nephrology	19,171	9,548,533
Ophthalmology	19,700	4,464,515
Total	**1,698,029**	**655,055,061**

Table 3: Sources of the NBDW corpus.

3.3 Pre-processing step

The supplied text data has been split into subword units using SentencePiece (Kudo and Richardson, 2018), an extension of Byte-Pair encoding (BPE) (Sennrich et al., 2016) and WordPiece (Wu et al., 2016) that does not require pre-tokenization (at the word or token level), thereby avoiding the requirement for language-specific tokenizers. We employ a vocabulary size of 32k subword tokens. For each model pre-trained from scratch (see Section 4.2), tokenizers were built using all the sentences from the pre-training dataset.

4 Models pre-training

In this section, we describe the pre-training modalities of our studied models from two points of view: 1) the influence of the data used (size and nature), and 2) the pre-training strategies of the models. These two levels are respectively detailed in Sections 4.1 and 4.2. Section 4.3 finally presents the existing state-of-the-art pre-trained models that will be used for comparison purposes.

4.1 Influence of data

One issue is to identify the amount of data required to create a model that performs well and can compete with models trained on general domains. Recent studies, such as those by Zhang et al. (2020) and Martin et al. (2020), discuss the impact of the size of pre-training data on model performance. According to these studies, some tasks are performing better with fewer data while others, such as commonsense knowledge and reasoning tasks, keep improving performance when pre-training data are added.

In the medical field, no study has been conducted to compare the impact of varying the amount of domain-specific data during pre-training, or to assess the impact of the supposedly variable quality of the data depending on their source of collection.

We thus propose to evaluate the pre-training of several language models on either NACHOS$_{small}$ or NBDW$_{small}$ corpus, as described in Section 3. Additionally, we propose a model pre-trained on NACHOS$_{large}$ to investigate if having almost twice as much data improves model performance. Finally, a combination of both public NACHOS$_{small}$ and NBDW$_{small}$ sources for a total of 8 GB (NBDW$_{mixed}$) is explored, to demonstrate if combining private and public data is a viable approach in low-resource domains.

4.2 Pre-training strategies

In addition to the analysis on the size and the sources of data, we also seek to evaluate three training strategies of PLMs for the medical domain:

- Training a full model from scratch, including the subword tokenizer.
- Continuing the pre-training of the state-of-the-art language model for French, called CamemBERT, on our medical-specific data while keeping the initial tokenizer.
- Continuing the pre-training of a state-of-the-art domain specific language model for medical but here in English, called PubMedBERT, on our French data while keeping the initial tokenizer.

Regarding the last strategy, our objective is to compare the performance of an English medical model further pre-trained on our French medical data, against another one based on a generic French model. Indeed, the medical domains shares many terms across languages that make relevant the mixture of resources from two languages.

Table 4 summarizes all the configurations evaluated in this paper, integrating both the study of data size and pre-training strategies.

Model name	Strategy	Corpus
DrBERT	From scratch	NACHOS$_{large}$
DrBERT	From scratch	NACHOS$_{small}$
ChuBERT	From scratch	NBDW$_{small}$
ChuBERT	From scratch	NBDW$_{mixed}$
CamemBERT	continual pre-training	NACHOS$_{small}$
PubMedBERT	continual pre-training	NACHOS$_{small}$
CamemBERT	continual pre-training	NBDW$_{small}$

Table 4: List of studied pre-trained model configurations.

Model architecture All models pre-trained from-scratch use the CamemBERT $_{\text{base}}$ configuration, which is the same as RoBERTa $_{\text{base}}$ architecture (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters). We did not train the large version of our models due to resources limitations.

Language modeling We train the models on the Masked Language Modeling (MLM) task using HuggingFace library (Wolf et al., 2019). It consists of randomly replacing a subset of tokens from the sequence by a special token, and asking the model to predict them using cross-entropy loss. In BERT and RoBERTa models (including CamemBERT), 15% of the tokens are randomly selected. Of those selected tokens, 80% are replaced with the `<mask>` token, 10% remain unchanged and 10% are randomly replaced by a token from the vocabulary. We keep this masking probability of 15% for the training of our models.

Optimization & Pre-training We optimize the models for 80k steps with batch sizes of 4,096 sequences, each sequence filled with 512 tokens, allowing to process 2.1M tokens per step. The learning rate is warmed up linearly for 10k steps, going up from zero to the initial 5×10^{-5} learning rate. Models are trained on 128 Nvidia V100 32 GB GPUs for 20 hours on Jean Zay supercomputer. We use mixed precision training (FP16) (Micikevicius et al., 2017) to reduce the memory footprint, allowing us to enlarge the batch size to 32 sequences on each GPU.

4.3 Baseline models

We describe some existing pre-trained models used as baselines in our comparative study.

CamemBERT (Martin et al., 2020) is a RoBERTa based model pre-trained totally from scratch on the French subset of OSCAR corpus (138 GB). In our case, this model is our main baseline to compare our results on, since it is the state-of-the-art model for French. We also use the 4 GB model's variants of CamemBERT to compare the impact of the nature and quantity of the data.

PubMedBERT (Gu et al., 2021) is a BERT based biomedical-specific model pre-trained totally from scratch on the 3.1 billions words of PubMed corpus (21 GB).

ClinicalBERT (Huang et al., 2019) is a clinical-specific model based on BERT tokenizer and weights, which has been further pre-trained on the 0.5 billion words of MIMIC corpus (3.7 GB).

BioBERT v1.1 (Lee et al., 2019) is a biomedical-specific model based on BERT tokenizer and weights which has been further pre-trained using the 4.5 billion words of PubMed corpus.

5 Downstream evaluation tasks

To evaluate the different pre-training configurations of our models, a set of tasks in the medical domain is necessary. While this NLP domain-specific benchmark exists in English (BLURB (Gu et al., 2021)), none exist for French. In this section, we describe an original benchmark, summarized in Table 5, integrating various NLP medical tasks for French. Among them, some are from publicly-available datasets (Section 5.1), allowing the replication of our experiments. Other tasks come from private datasets (Section 5.2) and cannot be shared. However, they are useful to evaluate our models more accurately.

5.1 Publicly-available tasks

ESSAIS / CAS: French Corpus with Clinical Cases The ESSAIS (Dalloux et al., 2021) and CAS (Grabar et al., 2018) corpora respectively contain 13,848 and 7,580 clinical cases in French. Some clinical cases are associated with discussions. A subset of the whole set of cases is enriched with morpho-syntactic (part-of-speech (POS) tagging, lemmatization) and semantic (UMLS concepts, negation, uncertainty) annotations. In our case, we focus only on the POS tagging task.

FrenchMedMCQA The FrenchMedMCQA corpus (Labrak et al., 2022) is a publicly available Multiple-Choice Question Answering (MCQA) dataset in French for medical domain. It contains 3,105 questions coming from real exams of the French medical specialization diploma in pharmacy, integrating single and multiple answers.

QUAERO French Medical Corpus The QUAERO French Medical Corpus (Névéol et al.,

Thematic / Corpus name	Task	Metric	Train	Dev	Test
Public Corpus					
ESSAIS (Dalloux et al., 2021)	POS Tagging	Macro F1	9,693	2,077	2,078
CAS: French Corpus with Clinical Cases (Grabar et al., 2018)	POS Tagging	Macro F1	5,306	1,137	1,137
MUSCA-DET - Social Determinants of Health extraction (Task 1)	Nested NER	Macro F1	19,861	2,207	5,518
MUSCA-DET - Social Determinants of Health extraction (Task 2)	Multi-label Classification	Macro F1	19,861	2,207	5,518
QUAERO French Medical Corpus - EMEA (Névéol et al., 2014)	Nested NER	Weighted F1	11	12	15
QUAERO French Medical Corpus - MEDLINE (Névéol et al., 2014)	Nested NER	Weighted F1	833	832	833
FrenchMedMCQA (Labrak et al., 2022)	MCQA	EMR / Hamming Score	2,171	312	622
Private Corpus					
Medical report acute heart failure structuration	Named Entity Recognition	Macro F1	2,527	281	703
Acute heart failure (aHF) classification	Binary Classification	Macro F1	1,179	132	328
Technical Specialties Sorting	Classification Multi-class	Macro F1	4,413	1,470	1,473
Medical report structuration prescriptions	Named Entity Recognition	Macro F1	61	15	26

Table 5: Corpus, tasks and metrics synthesis for evaluating medical-specific models.

2014) introduces an extensive corpus of biomedical documents annotated at the entity and concept levels to provide NER and classification tasks. Three text genres are covered, comprising a total of 103,056 words obtained either from EMEA or MEDLINE. Ten entity categories corresponding to UMLS (Bodenreider, 2004) Semantic Groups were annotated, using automatic pre-annotations validated by trained human annotators. Overall, a total of 26,409 entity annotations were mapped to 5,797 unique UMLS concepts. To simplify the evaluation process, we sort the nested labels by alphabetical order and concatenate them together into a single one to transform the task to a usable format for token classification with BERT based architectures.

MUSCA-DET MUSCA-DET is a French corpus of sentences extracted from the "Lifestyle" section in clinical notes from Nantes University Hospital biomedical data warehouse. The corpus contains 27,000 pseudonymized sentences annotated with 26 entities related to Social Determinants of Health (living, marital status, housing, descendants, employment, alcohol, smoking, drug abuse, physical activity). The corpus includes two tasks: nested name entity recognition (NER) and multi-label classification.

5.2 Private tasks

Technical Specialties Sorting This classification task has to assign the specialty of a medical of a medical report based on its transcription. The dataset consists of 7,356 French medical reports that have been manually annotated and equally sampled across 6 specialties: Psychiatry, Urology, Endocrinology, Cardiology, Diabetology, and Infectiology.

Medical report structuration prescriptions (NER) The task seeks to identify named entities in a gold sample of 100 long medical reports obtained from French speech transcriptions. The named entities are annotated using the BIO format and fall into 12 classes: *O*, *AGE*, *CITY*, *DATE*, *EMAIL*, *HOSPITAL*, *PHONE*, *DOSAGE*, *DURATION*, *FORM*, *MEDICATION* and *POSOLOGY*.

Medical report acute heart failure structuration (NER) This corpus contains 350 hospital stay reports (divided into 3,511 sentences) from Nantes University Hospital. The reports are annotated with 46 entity types related to the following clinical information: cause of chronic heart failure, triggering factor for acute heart failure, diabetes, smoking status, heart rate, blood pressure, weight, height, medical treatment, hypertension and left ventricular ejection fraction. Overall, the corpus contains 6,116 clinical entities.

Acute heart failure (aHF) classification This task consists of the classification of hospital stays reports according to the presence or absence of a diagnostic of acute heart failure. This corpus consists of 1,639 hospital stays reports from Nantes university hospital, which are labeled as positive or negative to acute heart failure.

6 Results and Discussions

As previously described, we evaluate the performance of our pre-trained language models proposed for the biomedical domain on a set of public and private NLP downstream tasks related to the medical domain. We first propose to analyze the results according to the different pre-training strategies used (Section 6.1) then to focus on the impact of the pre-training data, whether in terms of size or nature (Section 6.2). Finally, we are interested in the generalization capacities of our domain-specific

	aHF NER			aHF classification			NER Medical Report			Specialities Classification		
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
CamemBERT OSCAR 138 GB	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 GB	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

Table 6: Performance on our private biomedical downstream tasks. Best model in bold and second is underlined.

models by applying and comparing them on general domain NLP tasks (Section 6.3).

Note that all the PLMs have been fine-tuned in the same way for all downstream tasks and all the reported results are obtained by averaging the scores from four runs. Performance on biomedical downstream tasks are reported in Tables 6 and 7 for respectively private and public tasks. For readability reasons, the first part of each table presents the existing baseline models results, the second part our specialized models trained from-scratch, and the last part our models using continual pre-training.

6.1 Impact of pre-training strategies

As observed both in Tables 6 and 7, models pre-trained completely from scratch (DrBERT NACHOS and ChuBERT NBDW) tend to produce the best results for both types of data sources and tasks (*i.e.* private and public). Indeed, considering the F1-score, they obtain the best results on all private tasks and on almost all public ones (5 tasks out of 7). The two public remaining tasks (MUSCA-DET T2 and QUAERO-MEDLINE) are then better handled using PubMedBERT NACHOS$_{small}$, a model that has already been pre-trained on domain-specific data (biomedical English data) then further pre-trained with our French medical data (NACHOS$_{small}$).

We also observed that continual pre-training from domain generic models (CamemBERT NACHOS$_{small}$ or CamemBERT NBDW$_{small}$) does not allow reaching the performance of the other specific models, neither of these two models reaching the first or second place (in terms of performance) on any task.

Finally, the baseline models trained on generic data (CamemBERT OSCAR) and those trained on biomedical data in English (PubMedBERT, ClinicalBERT and BioBERT) remain competitive in few biomedical public tasks (CAS POS, FrenchMCQA or MUSCA-DET T2), while none of them are placed in first or second place on private tasks. This seems to highlight the difficulty of private tasks when non-matching data are used.

6.2 Effect of data

Regarding the amount of data used for pre-training models ($small$ vs. $large$ or $mixed$), results show that, the larger the data are, the better the model performs, no matter the pre-training strategy or the source of data (private or public). However, the difference is very low for most tasks, with $small$ systems often being ranked second behind large models, even though they contain half as much data.

We notice a clear dominance of models that were pre-trained on web-based sources, specifically OSCAR and NACHOS, when applied to public tasks. Indeed, models relying on private NBDW data only achieve the best performance (in terms of F1-score) on the MUSCA-DET T1 task. This trend is not quite observed on private tasks, where NBDW-based models obtain more acceptable or even better performance when mixed with public biomedical data (ChuBERT NBDW$_{mixed}$), as seen in Table 6. We believe this discrepancy is mainly due to the different nature of processed data.

Finally, we observe that English-based models perform closely to the French-based CamemBERT model. This shows the usefulness of pre-training on domain specific data. For example, better results are obtained with continual pre-training of the PubMedBERT model with our specialized data in French (PubMedBERT NACHOS$_{small}$), corroborating our hypothesis about the effectiveness of cross-language knowledge transfer.

	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*	*Hamming*	*EMR*	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 GB	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
PubMedBERT NACHOS$_{small}$	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

Table 7: Performance on public biomedical downstream tasks. Best model in bold and second is underlined.

6.3 Performance on general-domain tasks

Table 8 gives the results obtained by all PLMs on general domain downstream tasks. These tasks come from Martin et al. (2020) who used them to evaluate the CamemBERT model. The first four are POS tagging tasks (GSD, SEQUOIA, SPOKEN and PARTUT), the last being a natural language inference task (XNLI).

All results of our models decrease in performance on all tasks. The most important drop is for the natural language inference task, with a performance of ChuBERT NBDW$_{small}$ almost 13% lower than CamemBERT 138 GB. We also observe that the specialized models in English are as efficient as our biomedical models in French. It seems quite clear from the previous observations that specialized models are difficult to generalize to other tasks, but that specialized information captured in one language could transfer to another language.

7 Conclusion

In this work, we proposed the first biomedical and clinical Transformer-based language models, based on RoBERTa architecture, for French language. An extensive evaluation study of these specific models has been performed on an aggregated collection of diverse private and public medical tasks. Our open-source DrBERT models improved the state of the art in all medical tasks against both French general model (CamemBERT) and English medical ones (BioBERT, PubMedBERT and ClinicalBERT). In addition, we showed that pre-training on constrained resources (4 GB) of web-crawled medical makes it possible to compete with, and even frequently surpass, models trained with specialized data from medical reports.

Results also highlighted that continual pre-training on an existing domain-specific English model, here PubMedBERT, is a more viable solution than on a French domain-generalist model while targeting French biomedical downstream tasks. It needs to further investigate the performance of this approach using more data, similar to what we have done with DrBERT NACHOS$_{large}$.

The pre-trained models as well as the pre-training scripts[3] have been publicly released online under a MIT open-source license. The main purpose of NACHOS dataset is to promote the development of robust NLP tools by the community,

[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

Table 8: Performance on public domain-general downstream tasks. Best model in bold and second is underlined.

so, we have decided to make the corpora available for academic research.

8 Ethical considerations

Concerning the risks and biases, all the freely available models pre-trained on NACHOS can supposedly be exposed to some of the concerns presented by the work of Bender et al. (2021) and Sheng et al. (2021) since some of the NACHOS sub-corpora quality might be lower than expected, specifically for non-governmental sources. When using a BERT-based biomedical language model, potential biases can be encountered including fairness, gendered language, limited representation and temporal correctness.

9 Limitations

It is important to mention some limitations of our work. Firstly, it would be wise to evaluate the impact of the tokenizer on the performance of the models to ensure that this is not the main reason for the observed performance gains.

Furthermore, we can not affirm in this study whether the medical domain transfer observed from English to French using continual pre-training on PubMedBERT can be generalized to other languages or other domains.

Finally, it is possible that training a ChuBERT model with more diverse private clinical data and in a larger quantity could have brought notable performance gains on private tasks.

A considerable amount of computational resources was used to conduct this study, since approximately 18,000 hours of GPU computation were used to create the 7 models presented here, as well as about 7,500 hours of GPU for debugging due to technical issues related to model configurations and poor performance, for a total of 25,500 hours. The total environmental cost, according to the Jean Zay supercomputer documentation[4] is equivalent to 6,604,500 Wh or 376.45 kg CO2eq based on the carbon intensity of the energy grid mention by BLOOM environmental cost study also made on Jean Zay (Luccioni et al., 2022). This makes the present study difficult to reproduce and to transpose to other languages when limited material resources are available.

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715) and from CCIPL (Centre de Calcul Intensif des Pays de la Loire). This work was financially supported by ANR AIBy4 (ANR-20-THIA-0011) and Zenidoc.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Emanuel Silva Oliveira, Claudia Maria Cabral Moro, Yohan Bonescki Gumiel, and Deborah Ribeiro Carvalho. 2021. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 27(2):181–201.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or Train from Scratch? Comparing Pre-training Strategies of BERT in the Medical Domain. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, pages 2626–2633, Marseille, France. European Language Resources Association.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh,

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

Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French Corpus with Clinical Cases. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 1–7, Brussels, Belgium.

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*, 3(1):1–23.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abou Dhabi, United Arab Emirates.

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*.

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. *CoRR*, abs/1907.11692.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suá rez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203—-7219. Association for Computational Linguistics.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2017. Mixed precision training. *CoRR*, abs/1710.03740.

Aurélie Névéol, Cyril Grouin, Jérémy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus : A ressource for medical entity recognition and normalization.

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)*, pages 58–65.

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 Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges.

Manjil Shrestha. 2021. Development of a language model for medical domain. masterthesis, Hochschule Rhein-Waal.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Hazal Türkmen, Oğuz Dikenelli, Cenk Eraslan, and Mehmet Callı. 2022. Bioberturk: Exploring turkish biomedical language model development strategies in low resource setting.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings* of the Thirteenth Language Resources and Evaluation Conference, pages 4245–4252, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

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 Johnson, 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 Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Xi Yang, Aokun Chen, Nima PourNejatian, 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, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

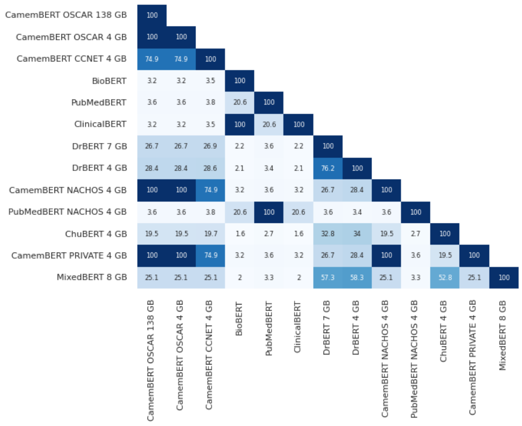
Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When do you need billions of words of pretraining data?

Hongyin Zhu, Hao Peng, Zhiheng Lyu, Lei Hou, Juanzi Li, and Jinghui Xiao. 2021. Travelbert: Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation.

A Appendix

A.1 Vocabularies Inter-coverage



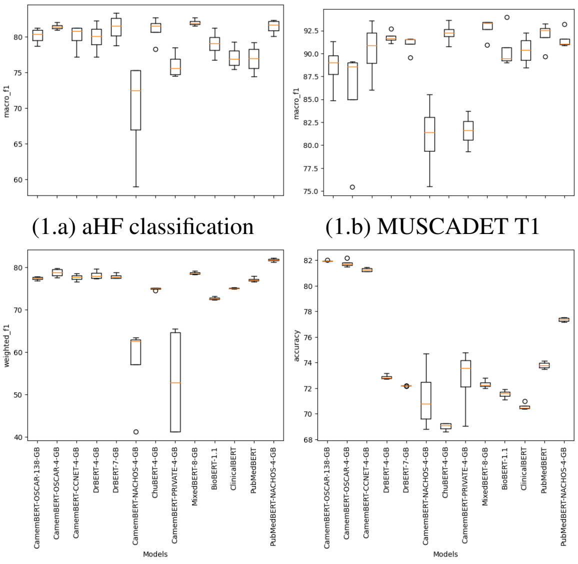
	CamemBERT OSCAR 138 GB	CamemBERT OSCAR 4 GB	CamemBERT CCNET 4 GB	BioBERT	PubMedBERT	ClinicalBERT	DrBERT 7 GB	DrBERT 4 GB	CamemBERT NACHOS 4 GB	PubMedBERT NACHOS 4 GB	ChuBERT 4 GB	CamemBERT PRIVATE 4 GB	MixedBERT 8 GB
CamemBERT OSCAR 138 GB	100												
CamemBERT OSCAR 4 GB	100	100											
CamemBERT CCNET 4 GB	74.9	74.9	100										
BioBERT	3.2	3.2	3.5	100									
PubMedBERT	3.6	3.6	3.8	20.6	100								
ClinicalBERT	3.2	3.2	3.5	100	20.6	100							
DrBERT 7 GB	26.7	26.7	26.9	2.2	3.6	2.2	100						
DrBERT 4 GB	28.4	28.4	28.6	2.1	3.4	2.1	76.2	100					
CamemBERT NACHOS 4 GB	100	100	74.9	3.2	3.6	3.2	26.7	28.4	100				
PubMedBERT NACHOS 4 GB	3.6	3.6	3.8	20.6	100	20.6	3.6	3.4	3.6	100			
ChuBERT 4 GB	19.5	19.5	19.7	1.6	2.7	1.6	32.8	34	19.5	2.7	100		
CamemBERT PRIVATE 4 GB	100	100	74.9	3.2	3.6	3.2	26.7	28.4	100	3.6	19.5	100	
MixedBERT 8 GB	25.1	25.1	25.1	2	3.3	2	57.3	58.3	25.1	3.3	52.8	25.1	100

Figure 1: Vocabularies Inter-coverage Matrix.

As we can see in Figure 3, despite having similar performances, some of the models do not share a lot of mutual vocabulary.

A.2 Models Stability

We observe during the evaluation phase that most of the models based on continual pre-training strategy from CamemBERT OSCAR 138 GB are suffering from bad consistency and stability during fine-tuning, which translates into fluctuation in performance between runs.



(1.a) aHF classification

(1.b) MUSCADET T1

(2.a) QUAERO MEDLINE

(2.b) XNLI

Figure 2: Box plot for each model.

We also notice during PubMedBERT NACHOS$_{small}$ pre-training that the model loss is globally stable during almost all the duration of the pre-training, until reaching the step 71,000, where the loss fall down until touching down zero at step 72,500.

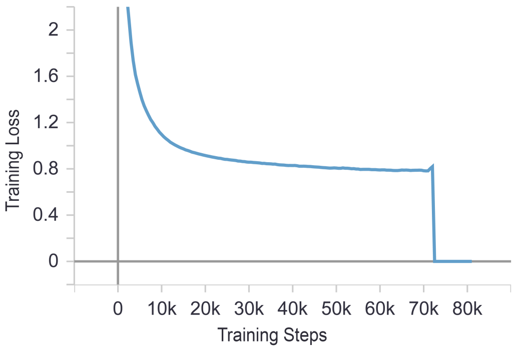


Figure 3: PubMedBERT NACHOS$_{small}$ loss.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9

☑ A2. Did you discuss any potential risks of your work?
In the Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

*The pre-trained models used as our baselines are presented in Section 2 and 4.3.
The datasets used are listed in the Table 5.
For the tools, we cite Huggingface library in Section 4.2 under the paragraph "Language modeling".
Our publicly available artifacts are explicitly enumerate in the contributions of our introduction (Section 1) and conclusion (Section 7). Will be made available online once the paper accepted: - Our training scripts (available at the anonymized repository link in the paper) under MIT license. - The web crawled pre-training corpus called NACHOS on Zenodo and HuggingFace (currently private) under CC0 1.0 license. - Our models trained on NACHOS on HuggingFace (currently private) under MIT license.*

☑ B1. Did you cite the creators of artifacts you used?
*We cite the used artifacts models during the Section 2 and 4.3.
The datasets used are listed in the Table 5 and cited in the section 5.1 and 6.3.
For the tools, we cite Huggingface library in Section 4.2 under the paragraph "Language modeling".*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The licenses are explicited at the end of the introduction (Section 1).

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*We used two corpora for our language models :
The first is trained on crawled data, hand-picked from high quality web sources, and doesn't require any anonymization step.
For our second model, it's trained on clinical data, and we specify in Section 3.2 that we used de-identified hospital stays reports and comply with GDPR and local authorities since we have official accreditations.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Coverage of domains are presented in Tables 2 and 3.
Concerning languages, we only consider French in our study, as the title suggest.
Demographic groups represented in the data are not explicited in the paper.*

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
The datasets used for evaluation and their splits are presented in the Table 5 for both used and created data.

C ☑ Did you run computational experiments?

Section 4.2 under paragraph "Optimization & Pre-training".

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, Section 4.2 under paragraphs "Model architecture" and "Optimization & Pre-training" for the GPUs used and number of parameters. End of the Section 9 for the total computational budget and infrastructure used, since it's one of the limitation of this article.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.2 under paragraph "Optimization & Pre-training".

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We are averaging 4 runs for all the experiments presented in the Tables 6, 7 and 9.
We also provide the Figure 2 in the Appendix to show the variability between the 4 runs of 4 of the tasks thanks to a box plot.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We have used hundreds of RegEx rules specific to each data sources to clean both corpus. It's not relevant for the public models, since the datasets NACHOS, used for pre-training them, will be available online soon after acceptance notification.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.