AliBERT: A Pre-trained Language Model for French Biomedical Text

Aman Berhe*
Quinten
IRD, Sorbonne University, UMMISCO
91, bvd Hopital, F-75013, Paris, France
amanzaid.berhe@ird.fr

Guillaume Draznieks*
Quinten
8 rue Vernier, 75017 Paris
gdraznieks@student.ethz.ch

Vincent Martenot
Quinten
8 rue Vernier, 75017 Paris
v.martenot@quinten-france.com

Valentin Masdeu
Quinten
8 rue Vernier, 75017 Paris
v.masdeu@quinten-france.com

Lucas Davy
Quinten
8 rue Vernier, 75017 Paris
l.davy@quinten-france.com

Jean-Daniel Zucker
IRD, Sorbonne University, UMMISCO,
INSERM, Sorbonne University, NUTRIOMICS,
91, bvd Hopital, F-75013, Paris, France
jean-daniel.zucker@ird.fr

Abstract

Over the past few years, domain-specific pre-trained language models have been investigated and have shown remarkable achievements in different downstream tasks, especially in biomedical domain. These achievements stem on the well-known BERT architecture which uses an attention-based self-supervision for context learning of textual documents. However, these domain-specific biomedical pre-trained language models mainly use English corpora. Therefore, non-English, domain-specific pre-trained models remain quite rare, both of these requirements being hard to achieve. In this work, we proposed AliBERT, a biomedical pre-trained language model for French and investigated different learning strategies. AliBERT is trained using regularized Unigram based tokenizer trained for this purpose. AliBERT has achieved state-of-the-art F1 and accuracy scores in different down-stream biomedical tasks. Our pre-trained model manages to outperform some French non domain-specific models such as CamemBERT and FlauBERT on diverse down-stream tasks, with less pre-training and training time and with much smaller corpora.

1 Introduction

Recent contextual language models have achieved tremendous results in almost all domains using textual information. Transformers (Vaswani et al., 2017) based pre-trained language models (T-PLM) have contributed and continue to contribute to the success of natural language processing (NLP) in multiple domains of expertise. Furthermore, very large transformer based models which require hundreds of billions of parameters have shown extraordinary achievements and became more accessible.

The biomedical field is one of the most important domain and its associated textual corpora is one of the fast-growing sources of information in several languages. Hence, researchers have leveraged PLMs to represent biomedical knowledge from different sources, following their success in the general domain. There are plenty of Biomedical Pre-trained Language Models (B-PLMs) that have achieved interesting results and that help decision making in the biomedical field, such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2022), BioELECTRA (raj Kanakarajan et al., 2021), etc.

PLMs are trained using different training mechanisms. The most common are masked language modeling (MLM) (Devlin et al., 2019), replaced token detection (RTD) (Clark et al., 2020) or next sentence prediction (NSP) (Devlin et al., 2019). Training a biomedical language model using different strategies does benefit the different downstream tasks. Furthermore, B-PLMs apply various pre-training methods since they borrow some characteristics from already existing PLMs. Commonly used pre-training methods includes continual pre-training (CPT), mixed domain pre-training and domain-specific pre-training (DSPT). In this work, DSPT was used for training our proposed model from scratch using domain-specific French

*The first two authors have equal contribution.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 223–236
July 13, 2023 ©2023 Association for Computational Linguistics

corpora. Furthermore, B-PLMs use tokens as input. Tokenization is indeed the basic step of language model training since it is the tokens that are directly used as discrete input for model pre-training. There are different ways to tokenize a text input. The most common tokenization techniques are Byte Pair Encoding (BPE) (such as; SentencePiece, WordPiece, etc.) and Unigram sub-word based tokenization. Consideration and implementation of different tokenization techniques are equally important to achieve better performance of B-PLMs, especially when the model is language-specific. Language-specific PLMs can use common tokenization techniques like BPE, but they can also tailor the tokenization process and train a tokenizer that can fit a specific language and domain under consideration. In a similar way the biomedical text differs from general domain texts, so the use of custom tokenization allows for better representation of most biomedical vocabulary (words).

Biomedical language models in languages other than English, i.e. PLMs that are both domain and language-specific, are quite rare. In the field of non-English language-specific models, there are a few that focus on French language, such as CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). French is a very rich language and French-based PLMs (Martin et al., 2020; Le et al., 2020) have shown the importance of such language-specific model for different purposes. However, French biomedical textual information have not been implemented using transformers based PLM. Yet, there are a few French language word embedding in biomedical domains. (Dynomant et al., 2019) compared different word embedding techniques (word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014)) for French health-related documents. Given the disadvantages of embedding words for their representation it is necessary to build B-PLMs for better representation. In this work, we propose AliBERT (named after Jean-Louis-Marc AliBERT the French pioneer of dermatology), a BERT-based language-specific and domain-specific Biomedical language model. AliBERT uses a masked language model (MLM) pre-training mechanism which randomly masks some of the tokens from the input biomedical text and predicts the masked tokens based on the context of the input. Thereby learning the context of each word according to the biomedical text input. A Unigram based tokenizer with a novel regularization algorithm has been trained for AliBERT pre-training. In addition to the MLM, we have also trained an ELECTRA-based (Clark et al., 2020) model called AliBERT-ELECTRA. AliBERT-ELECTRA is trained using the replaced token detection mechanism using the same vocabularies and tokenization steps as AliBERT. In addition, the LAMB optimizer is studied to analyze its computational speed gain during model pre-training. Here are the main contributions of our work:

- A French biomedical language model, a language-specific and domain-specific PLM, which can be used to represent French biomedical text for different downstream tasks.
- A normalization of a Unigram sub-word tokenization of French biomedical textual input which improves our vocabulary and overall performance of the models trained.
- AliBERT outperforms other French PLMs in different downstream tasks. It is a foundation model that achieved state-of-the-art results on French biomedical text. Models are available on HuggingFace hub[1] and datasets are available to the public [2].

This paper is organized in the following manner: first the related work is discussed in section 2, different language-specific and domain-specific PLMs and their pre-training objectives and strategies are discussed. Second, section 3 presents our B-PLM AliBERT with details on architecture, tokenization and optimization. Then, section 4 discusses the fine-tuning and evaluation of our models in down-stream tasks. Next, section 5 explain the experiments and results on the down-stream tasks. Then, section 6 discusses the results found and the drawbacks we encountered in detail. Finally, section 7 concludes the findings of this paper and points out our future directions concerning the domain-specific and language-specific PLMs.

2 Related Work

In recent years, the number of language models based on Transformers (Vaswani et al., 2017) has grown rapidly and their performance has been remarkable in many areas. The pioneers of Transformers based PLMs (T-PLMs) are BERT (Devlin

[1] Quinten-datalab/AliBERT
[2] https://gitlab.par.quinten.io/qlab/alibert

et al., 2019) and GPT (Radford et al., 2018) which are a stack of encoders and decoders of transformers, respectively. Consequently, the T-PLMs can be mainly divided as transformer encoder based models such as ALBERT (Lan et al., 2020), RoBERTa (Zhuang et al., 2021), ELECTRA (Clark et al., 2020), and transformer decoder based model such as BART (Lewis et al., 2019), PEGASUS (Zhang et al., 2019), and T5 (Raffel et al., 2020). Devlin et al. (2019) played an important role for the increase of T-PLMs and fine-tuning many downsteam tasks. They also paved the way for other languages (other than English), such as (Martin et al., 2020; Le et al., 2020; Delobelle et al., 2020; Cañete et al., 2020), to develop language-specific (monolingual) language models.

There are very few French language models (Martin et al., 2020; Le et al., 2020; Copara et al., 2020; Douka et al., 2021; Cattan et al., 2021). CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) are trained on general knowledge French corpora. CamemBERT used OSCAR dataset which is composed of 130 Gigabytes (GB) of raw French text with 32.7 Billion tokens whereas FlauBERT utilized 71 GB of raw text with 12.7 Billion of token. BERTweetFR (Guo et al., 2021) is another French PLM trained on French tweets. BERTweetFR is a general domain which is initialized using CamemBERT utilizing the largest French tweet corpora which is composed of 16 GB of 226 Million tweets. They took tweets with an average length of 30 tokens. Kamal Eddine et al. (2021) developed a BART based French language model called BARThez which is a generative language model based on BART[3] (Lewis et al., 2019). BARThez used 66 GB (110 GB after tokenization) raw text for pre-training. Cattan et al. (2021) investigated the usability of transformer based models for French question answering task and provided a model known as FrALBERT which is based on the compact language models (parameter efficient BERT) called ALBERT (Lan et al., 2020). FrALBERT is a compact language model pre-trained on the French version of the Wikipedia encyclopedia as of 04/05/2021. Their dataset is composed of 4 GB of text and 17 million sentences. There are two French domain-specific PLMs. The first one is JuriBERT (Douka et al., 2021), it is a French legal language model (language and domain-specific) which is trained on 6.3 GB of raw legal text[4]. The second is CamemBioBERT (Copara et al., 2020), it is a fine-tuned CamemBERT (Martin et al., 2020) using biomedical text from a French language challenge known as DEFT ("Défi Fouille de Textes")[5]. Dura et al. (2022) introduced their ongoing work on a clinical French language model, known as EDS (Entrepôt des Données de Santé), that uses 21 million French clinical reports from electronic health records (EHR) from several hospitals in the Paris area. Dura et al. (2022) claimed that their preliminary results achieved better results than CamemBERT (Martin et al., 2020). They have trained EDS from scratch (EDS-from-scratch) and continuous training over CamemBERT (EDS-fine-tuned).

Regarding domain-specific-language models, Lee et al. (2020) built the first BERT based language model in English in the biomedical domain, known as BioBERT. BioBERT (Lee et al., 2020) is built on top of the BERT (Devlin et al., 2019) model using abstracts of biomedical articles. Following the publication of BioBERT, biomedical language models have gained considerable momentum. A survey (Kalyan et al., 2021) studied many publicly available language models in the biomedical domain and provided a survey of systematic literature review, known as AMMU. AMMU includes 121 articles of biomedical language models.

AMMU investigated the core B-PLMs concepts, such as pre-training methods, pre-training tasks, fine-tuning methods and embeddings. Furthermore, Kalyan et al. (2021) disclosed different types of corpora along with the language models that used the corpus. The main corpora included were electronic health record (EHR), radiology reports, social media texts and scientific literature. They have listed out the most common learning objectives such as Masked Language Modeling (MLM), Replaced Token Detection (RTD), Next Sentence Prediction (NSP), Sentence Order Prediction (SOP) and Span Boundary Objective (SBO). There are few non-English transformer-based biomedical PLMs (Schneider et al., 2020; Bressem et al., 2020; López-García et al., 2021; Vakili et al., 2022). Most of the models are pre-trained using the continual pre-training (CPT) approach which means they used already pre-trained language-specific general knowledge PTM. We invite readers to refer to the AMMU

[3] BART: De-noising Sequence-to-Sequence pre-training for Natural Language Generation, Translation, and Comprehension

[4] Number of token used in JuriBERT (Douka et al., 2021) not mentioned in the paper

[5] DEFT is a scientific evaluation campaign on Francophone text mining.

survey for details (Kalyan et al., 2021).

To the best of our knowledge, there is not yet a French biomedical transformer based PLM trained from scratch. However, as mentioned above we are aware of an ongoing work on a PLM for French clinical reports using proprietary EHR (Dura et al., 2022) From the literature we can clearly see that there is a gap in pre-trained language models for French biomedical text mining. Hence, our primary goal is to address this gap and enhance the tokenization of French biomedical texts. Instead of relying solely on general tokenization methods, we have standardized the tokenization process specifically for French biomedical texts.

3 Methods

This sections focuses on how the proposed pre-trained language model, AliBERT, was built. It describes the pre-training strategy and architecture, pre-training corpora, tokenization and optimization of our models.

3.1 Pre-training strategies

There are different kinds of pre-training strategies to train a transformers based models (Kalyan et al., 2021). Pre-training from scratch (PTS) is the strategy used for training AliBERT and its variants. They are trained from scratch using biomedical corpora to better represent the biomedical context of words. Training our models from scratch helps to represent vocabulary that only exists in biomedical text, which will be discussed in subsection 3.3.

The models developed are based on the transformers (Vaswani et al., 2017) architecture and RoBERTa (Zhuang et al., 2021) a variant of the BERT (Devlin et al., 2019) model is used as masked language model (MLM), transformers and BERT architecture will not be discussed here because they have been discussed extensively in many research works (Devlin et al., 2019; Martin et al., 2020). Therefore, AliBERT is trained in the course of self-supervised learning by masking 15% of the words from the input text (sequence of words). All necessary steps and configurations are discussed in the following subsections.

3.2 Pre-training data

The pre-training corpus was gathered from different sub-corpora of French biomedical textual documents. The sources used are a database of drug leaflets ("Base de données publique des médicaments"), a French equivalent of Physician's Desk Reference i.e. RCP[6], biomedical articles from ScienceDirect[7], Thesis manuscripts in French and articles from Cochrane database[8]. It can be inferred from the names of the corpora that they cover various topics in the biomedical domain and that they have different writing styles. Table 1 summarises the different corpora collected and used for pre-training AliBERT models.

Name	Quantity	Size
Drug database	23K	550Mb
RCP	35K	2200Mb
Articles	500K	4300Mb
Thesis	300K	300Mb
Cochrane	7.6K	27Mb

Table 1: Corpora used to pre-train AliBERT

The corpora were collected from different sources. Scientific articles are collected from ScienceDirect using an API provided on subscription and where French articles in biomedical domain were selected. The summaries of thesis manuscripts are collected from "Système universitaire de documentation (SuDoc)" which is a catalog of universities documentation system. Short texts and some complete sentences were collected from the public drug database which lists the characteristics of tens of thousands of drugs. Furthermore, a similar drug database known as "Résumé des Caractéristiques du Produit (RCP)" is also used to represent a description of medications that are intended to be utilized by biomedicine professionals. Pages of biomedical articles from Cochrane are also collected. Hence, our corpus for pre-training is composed altogether of around 7 gigabytes (GB) textual documents.

When compared with the corpora of already existing French T-PLMs, our corpus is big enough to represent a biomedical text. Table 2 compares the different corpora used for pre-training French language models.

[6] The "Résumé des Caractéristiques du Produit" (RCP) database aims at providing more accurate information than the instructions note for use of medicines.

[7] ScienceDirect is a website which provides access to a large bibliographic database of scientific and medical publications of the Dutch publisher Elsevier.

[8] Cochrane is a British international charitable organisation formed to organise medical research findings to facilitate evidence-based choices about health interventions involving health professionals, patients and policy makers.

Model	Domain	Size	Source
CamemBERT (Martin et al., 2020)	general	138 GB	OSCAR
FlauBERT (Le et al., 2020)	general	71 GB	WMT19, OPUS, Wikmedia
BERTweetFR (Guo et al., 2021)	general	16 GB	French tweets
JuriBERT (Douka et al., 2021)	legal	6.3 GB	LégalFrance & Court of Causation
FrAlbert (Cattan et al., 2021)	general	4.0 GB	Wikipedia
AliBERT	biomedical	7.0 GB	ScienceDirect, SuDoc, Drug databases and Cochrane

Table 2: Comparison of the AliBERT corpus with that of the existing French PLMs.

3.3 Tokenization

In the context of Pre-trained Language Models (PLMs), tokenization refers to the process of dividing the input text into subwords or words known as tokens that will serve as the input to the model. Most BERT based PLMs use sub-word tokenization scheme such as Byte Pair Encoding (BPE), WordPiece and SentencePiece. However, the tokenization process can be adapted or trained to meet a specific purpose and/or to represent a vocabulary in a specific domain. We chose to train our own tokenizer to ensure that its vocabulary encompasses the necessary biomedical terms.

A normalization step prior to tokenization, particularly adapted to French, was used to enhance our vocabulary. In this step we added a space after a selected list of punctuation mark. It normalises the representation of the text, and facilitates both the tokenization and learning by the neural network. Hence, this step leads to a significant reduction of duplicates, such as, ("MOT", "_MOT"),("_siècle","_siècles") which were introduced due to punctuation marks like "(", " :", "-", etc. in the text.

We have trained different tokenizers, such as Unigram, WordPiece with different parameters (vocabulary size, regularization). Unlike BPE, Unigram starts with a large vocabulary and removes tokens until it reaches the desired vocabulary size. During training, at every step, Unigram computes a loss over the corpus given the current vocabulary. Then, for each symbol it calculates how much the overall loss would increase if the symbol was removed, and looks for the symbols that would decrease it the most. Appendix A discusses the steps taken during tokenization with an example and compares Unigram tokenizers trained from scratch and the tokenizer from CamemBERT (Martin et al., 2020).

3.3.1 Training configurations

When training a large language model, it is necessary to take into account different configurations needed to build a well-performing model. Therefore, model architecture, training strategy, optimization and computation are key parameters to consider.

Model architecture and training: We have mainly developed two architectures of our French B-PLM namely AliBERT: a BERT (Lan et al., 2020) based and AliBERT-ELECTRA an ELECTRA (Clark et al., 2020) based, models. BERT and ELECTRA differ only in their learning strategy. The former uses masked language modeling (MLM) and the later uses replaced token detection (RTD). AliBERT$_{base}$ has the same architecture as BERT$_{base}$which has a length (L) of 12, height (H) of 512 and a self-attention head (A) of 12.

For MLM, a sequence of words is given as input and 15% of the words are hidden. The input goes through the tokenization stage and the words are tokenized. The tokens are padded or truncated to have a maximum length of 512 tokens. Hence, special tokens "[CLS]","[PAD]" are added if the sequence length is less than 512 tokens. Then the embeddings of the tokens are passed to the transformer layers to learn the context of the input and the relationship of the tokens. Finally, the output of the transformers is passed to a feed-forward neural network to compute the probability distribution of the token to predict the masked tokens/words. For more detail on this training method see the original work of BERT (Devlin et al., 2019).

Another strategy different from MLM is RTD, in RTD the objective is to predict which tokens have been replaced and which have not. A very simple pre-trained model is used as a generator to predict a masked word from the input text. Then, the predicted words are used to replace the masked inputs and the unmasked sentence is used as input text in the discriminator model. Eventually, the discriminator model is used to identify the original words of the original input text. For more details of the architecture, we invite our readers to refer to the original work of ELECTRA (Ozyurt, 2020).

Optimization: AliBERT was originally trained using the ADAM [9] optimizer for faster and better training as used in BERT. Meanwhile, a recent work by You et al. (2020) introduced an optimizer known as LAMB that reduces the training time of BERT from 3 days (4320 minutes) to 76 minutes. Therefore, AliBERT was also trained using LAMB optimizer for the purpose of comparing it with ADAM optimizer which is the default for our pre-trained models.

The models trained using LAMB optimizer trained much faster than their counter part (using ADAM). However the performance of the models trained with LAMB was not as good as the models trained with ADAM. The loss of the model quickly reduces when LAMB optimizer is used during training. Figure 3 in Appendix B shows the comparison of time taken to train using LAMB and ADAM atomizers on our models. Moreover, AliBERT trained with ADAM optimizer achieved better results in NER downstream task. Table 6 compares two AliBERT models trained with ADAM (AliBERT) and LAMB (AliBERT-LAMB) optimizers.

4 Fine-tuning and Model Evaluation

In order to evaluate the level of understanding of French biomedical tasks by AliBERT, we have fine-tuned AliBERT on standard pre-trained language model evaluation tasks such as biomedical named entity recognition (NER), biomedical text classification, etc. Below, we discussed how the tasks are trained.

4.1 Biomedical named entity recognition (NER)

For the NER task we have used HuggingFace[10] token classification pipeline using our AliBERT models. The first dataset used is "CAS dataset", from the work of Grouin et al. (2019), which is used in different challenges of French biomedical text challenge known as "DEFT (Défis Fouille de Texte)". It is composed of clinical French texts which focuses on specific specialties of medical domains such as cardiology, urology, oncology, obstetrics, pneumatic, etc. The annotation in this dataset include plenty of biomedical entities where some of them do not have adequate annotation. Hence, we have kept only the five most-annotated types, i.e, anatomy, pathology, symptom, substance and value. Appendix D describes the annotated dataset used in NER task for fine-tuning and evaluation purposes.

Meanwhile, QUAERO (Névéol et al., 2014) datasets is used for more experiment and fine-tuning. QUAERO datasets is composed of ten annotated entity categories corresponding to UMLS (Unified Medical Language System) semantic groups. The annotation was performed using automatic pre-annotations and validated by trained human annotators. We have selected five entities that are most related to biomedical concepts, from the QUAERO-MEDLINE datasets which consist of article titles from the MEDLINE[11] database. The five entities are selected according to their definition and their relatedness with biomedical domain. The entities and the dataset are discussed in appendix Appendix D.

4.2 Biomedical text classification

For the biomedical text classification, we have used a private dataset which is composed of 410,000 examples and 789 classes. Hence, it is an extreme classification problem. Classes that have more than 1000 examples have been selected. A sequence classification model from hugging face was used to fine-tune a downstream classification model.

5 Experiments and Results

AliBERT$_{base}$ was trained on 48 GPUs Nvidia A100 (12 nodes each with 4 GPUs) for 20 hours with 512 input tokens and a batch size of 960 (20 batch size for each GPU). We have used a vocabulary of 40K sub-word units which are built using Unigram tokenization algorithm.

Our models have been evaluated using the above-mentioned fine-tuning models and on the masked token prediction. The results found using our models have been compared to the CamemBERT (Martin et al., 2020) French PLM which is the state-of-the-art in French language. Unfortunately, we were not able to compare our models with biomedical

[9] Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments.

[10] HuggingFace: the AI community building the future. https://huggingface.co/

[11] http://www.ncbi.nlm.nih.gov/pubmed/

PLMs due to the lack of French PLM in biomedical domain.

The downstream tasks on which our models have been evaluated are Biomedical NER, classification and Masking Language Modeling (MLM). The downstream datasets are not included in the training dataset of our models. However, there might be an overlap with the QUAERO-MEDLINE article titles. The results obtained on these tasks are detailed below.

Biomedical Named Entity Recognition (NER) A token classification model was fine-tuned from the pre-trained models mainly in 5 biomedical entity types, these are **symptoms**, **anatomy**, **substance**, **value** and **pathology**. Our models have outperformed CamemBERT in most of the entities and in their macro average of precision (P), recall(R) and F1 score (F1).

The results found in Table 3 are trained upon a batch size of 80, learning rate (lr) of 2e-5 and weight decay of 0.01 and the dataset used for each of the entities is discussed on Appendix D. Table 3 illustrates that `AliBERT` and `AliBERT`-ELECTRA outperformed CamemBERT considering the precision of the models to detect the entities. CamemBERT achieved higher F1 score than our models' for the "Pathology" entity. This is due to the fact that the pathology entities in the dataset are very long text that includes many words that exist in the general French language words (CamemBERT vocabularies). For example, `"tumeur qui est d'allure maligne et qui envahissait la face postérieure et la corne vésicale droite"` is annotated as a single pathology entity. However, our models exhibited a noteworthy improvement in F1 score for the other entities when compared to CamemBERT. Furthermore, our model outperformed CamemBERT for disorder (including pathology) on the QUAERO dataset.

Table 4 shows the results of NER task on the QUAERO dataset and it compares the results with CamemBERT. Our model outperformed the two models on identifying different kinds of entities (Disorder, Anatomy, Device, Disorder and Procedure) in QUAERO dataset with around 15% macro average f1 score improvement. We selected the entities that are closely related to biomedical concepts. In, Table 4 CamemBERT was not able to identify any medical device whereas `AliBERT` and `AliBERT`-ELECTRA detected the devices with f1 score of 42%. Hence, we can say that the B-PTMs can identify to the specific terms used in the domain.

Masking language modeling and classification We have also compared the ability of the models to predict masked tokens and biomedical text classification. In the same way our proposed models have outperformed CamemBERT. For this experiment of unmasking evaluation a subset of 3000 text of clean texts (1000 articles of ScienceDirect, 1000 articles from Cochrane, 1000 thesis abstracts from SuDuc) was used. For the biomedical text classification, we selected classes with more than 1000 examples, resulting in 50 classes, from our private data. Table 5 illustrates the performance of different models for the prediction of the masked word and classification, in top 1, 3 and 5 Accuracy (Acc, 3-Acc and 5-Acc respectively).

`AliBERT` has outperformed CamemBERT on predicting a masked word prediction (see Figure B for examples). It can be seen in Table 5 `AliBERT` has an increase of 23% in accuracy when compared with CamemBERT. In the same way for text classification our models achieved better top 1 accuracy. Hence, it clearly shows that in-domain pre-trained language models are really important while dealing with a domain-specific texts and hence domain-specific downstream tasks.

6 Discussion

Our pre-trained language models trained on in-domain (biomedical) textual documents tend to outperform models that are trained on general domain textual documents which is also seen on the literature review of pre-trained language models for English language such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2022), etc. Training PLMs using the masked language model (MLM) objective shows somewhat better results, but the difference is not significant compared to the replaced token prediction (MLM) objective. Moreover, choosing the right optimizer like LAMB has an effect on the training speed of the pre-trained models but not on the performance of the models. During the training of our models different types of tokenizers, such as, Unigram, WordPiece, SentencePiece, BPE, etc. are trained and compared with each other. Unigram tokenizer along with our normalization (see section 3) step tend to outperform other tokenizers. Unigram was also trained into two ways, cased and uncased respectively. Lower casing the input text achieved better results than letting

	Models' performances on CAS dataset								
	CamemBERT			AliBERT			AliBERT-ELECTRA		
Entities	P	R	F1	P	R	F1	P	R	F1
Substance	**0.96**	0.87	0.91	**0.96**	**0.91**	**0.93**	0.95	0.91	0.93
Symptom	0.89	0.91	0.90	**0.96**	**0.98**	**0.97**	0.94	**0.98**	0.96
Anatomy	0.94	0.91	0.88	**0.97**	**0.97**	**0.98**	0.96	**0.97**	0.96
Value	0.88	0.46	0.60	**0.98**	**0.99**	**0.98**	0.93	0.93	0.93
Pathology	0.79	**0.70**	**0.74**	**0.81**	0.39	0.52	0.85	0.57	0.68
Macro Avg	0.89	0.79	0.81	**0.94**	0.85	0.88	0.92	**0.87**	**0.89**

Table 3: French Biomedical named entity recognition (NER) results. Performance in bold is the best achieved for the entity in question and the metrics in question

	Models' performances on QUAERO MEDLINE dataset								
	CamemBERT			AliBERT			AliBERT-ELECTRA		
Entity	P	R	F1	P	R	F1	P	R	F1
Anatomy	0.649	0.641	0.645	0.795	**0.811**	**0.803**	**0.799**	0.801	0.800
Chemical	0.844	0.847	0.846	0.878	**0.893**	**0.885**	**0.898**	0.818	0.856
Device	0.000	0.000	0.000	0.506	**0.356**	0.418	**0.549**	0.338	**0.419**
Disorder	0.772	0.818	0.794	0.857	**0.843**	**0.850**	**0.883**	0.809	0.845
Procedure	0.880	0.894	0.887	**0.969**	0.967	**0.968**	0.944	**0.976**	0.960
Macro Avg	0.655	0.656	0.655	0.807	**0.783**	**0.793**	**0.818**	0.755	0.782

Table 4: Biomedical named entity recognition (NER) results on the QUAERO MEDLINE dataset.Performance in bold is the best achieved for the entity in question and the measure in question

	MLM			Classification		
Model	Acc	3-Acc	5-Acc	Acc	3-Acc	5-Acc
CamemBERT	0.49	0.57	0.62	0.66	0.72	0.99
AliBERT	**0.72**	**0.83**	**0.87**	**0.68**	**0.73**	0.99
AliBERT-ELECTRA	0.71	**0.83**	**0.87**	**0.68**	**0.73**	0.99

Table 5: Results predicting the masked tokens (MLM) and biomedical classification

upper cases as it is. Biomedical text tend to have lots of words that are written in capital letters. But we have noted that they are not enough to be used for training our models as upper cases. Biomedical named entity recognition (B-NER) and biomedical text classification (private data, hence results not reported) were used to fine-tune our models to a specific task. Our models tend to generalize faster than the French counterpart general PLMs. For AliBERT or AliBERT-ELECTRA fewer examples of B-NER text inputs were required to start learning and generalize quickly and accurately. On the other hand, Camembert took more time to generalize and with less precision for biomedical entities. This is understandable as it was not trained using domain texts. In the same manner, this behaviour was reflected during biomedical text classification task. This can also be seen as a comparison to the vocabularies used by CamemBERT and our models. Our tokenizer's (Unigram) vocabulary and CamemBERT tokenizer's (SentencePiece) have a huge difference in content and size. The Unigram tokenizers used to train our models have a vocabulary size of 40008 while CamemBERT has a size of 32005. CamemBERT's vocabulary does not include most biomedical words. In fact, the two tokenizers have about 10,000 tokens in common in their vocabularies. Although the performance of our models is already very good, more and varied corpora could improve the models' capabilities. For example, medical notes, often found in electronic health records ("EHRs"), can help represent the knowledge and experience of practitioners.

In addition, to improve the models, continuous training on a general purpose pre-trained language, such as CamemBERT, could be implemented. Since our tokenizers were a bit different and our goal is to study a purely biomedical PLM, we have not investigated it yet.

7 Conclusion

This paper proposes a French biomedical pre-trained language model that was trained on several corpora of French biomedical textual materials. Two variants of the model are proposed using two different pre-training strategies. AliBERT is a pre-trained model based on BERT (Devlin et al., 2019) which used the pre-training strategy of masking language models (MLM). AliBERT-ELECTRA is based on ELECTRA (Clark et al., 2020) and used a replaced token prediction (RTP) learning strategy. Furthermore, a tokenization adaptation strategy was introduced as a building block for pre-training the two proposed models. A LAMB optimizer has also been tested to speed-up the learning of AliBERT. The proposed pre-training models have been tested on different downstream tasks and achieved state-of-the-art results on different tasks. Biomedical entity recognition (NER) and biomedical text classification downstream tasks are fine-tuned using different biomedical textual documents. Hence, AliBERT is expected to be used by different organization and practitioners that work with biomedical text for better understanding and to help make informed decisions regarding biomedical situations.

Limitations

Although our models performed well in all downstream tasks, the models also have some limitations. One of the limitations is the lack of varied biomedical corpus. Hence, we plan to work on integrating clinical documents e.g. EHR data, specifically physician notes, to make the model more robust to various kind of biomedical documents. The models can also be enlarged by using continual learning strategy from well-known French pre-trained language models. CamemBERT (Martin et al., 2020) can be used as a base model and the training can be continued using our biomedical corpus, like BioBERT (Lee et al., 2020) and others did. Moreover, our models used 512 sequence of tokens and more longer sequence lengths can be used as seen in the long language models like BigBird (Zaheer et al., 2020).

We are currently working on a new version of AliBERT with more data and a greater diversity of corpora that include text from EHR and medical notes in our corpora. Finally, we also plan to train AliBERT to generate biomedical texts for different purposes.

A reasonable amount of computational resources was used to conduct this study, since approximately 20,160 hours of GPU computation were used to create the three pre-trained models presented above. The total environmental cost according to Green Algorithm (Lannelongue et al., 2021)[12] is equivalent to 1.45 MWh or 71.11 kg CO2e. This computational cost and environmental impact should be taken into consideration when training such a model.

Ethics Statement

AliBERT, a BERT-based biomedical language model for the French language, has the potential to improve healthcare and research in French language. However, it is essential to address ethical considerations such as biases, privacy, misinformation, access and control, as well as accountability and transparency. We have implemented measures to mitigate biases, protect privacy, prevent malicious use and optimize efficiency as much as possible.

To responsibly develop and deploy AliBERT, collaboration between developers, researchers, policymakers, and healthcare professionals are crucial. By working together, stakeholders can ensure that AliBERT benefits a wide range of users and upholds ethical standards, ultimately maximizing its potential to improve healthcare and research in French biomedical domain.

Acknowledgements

The authors acknowledge a support from the French Institute for Sustainable Science (IRD) within the framework of the France Relance plan to support research and development (R&D) employment and collaborative research. This work was partly performed using HPC resources from a GENCI-IDRIS grant.

References

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski,

[12] http://calculator.green-algorithms.org/

Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. 2020. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics*, pages 5255–5261.

Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. On the usability of transformers-based models for a french question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 244–255.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC@ICLR*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations (ICLR)*, pages 1–18.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized french language models for biomedical named entity recognition. In *6e conférence conjointe journées d'études sur la parole (jep, 33e édition), traitement automatique des langues naturelles (taln, 27e édition), rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (récital, 22e édition). atelier défi fouille de textes*, pages 36–48.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: A dutch roberta-based language model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL): Student Research Workshop*, pages 203–209.

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 (ACL): Human Language Technologies*, pages 4171–4186.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101.

Basile Dura, Charline Jean, Xavier Tannier, Alice Calliger, Romain Bey, Antoine Neuraz, and Rémi Flicoteaux. 2022. Learning structures of the french clinical language: Development and validation of word embedding models using 21 million clinical reports from electronic health records. pages 1–10.

Emeric Dynomant, Romain Lelong, Badisse Dahamna, Clément Massonnaud, Gaétan Kerdelhué, Julien Grosjean, Stéphane Canu, and Stefan J Darmoni. 2019. Word embedding for the french natural language in health care: Comparative study. *JMIR Medical Informatics*, page e12310.

Cyril Grouin, Natalia Grabar, Vincent Claveau, and Thierry Hamon. 2019. Clinical case reports for NLP. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 273–282.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM transactions on computing for healthcare*, pages 1–23.

Yanzhu Guo, Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. Bertweetfr: Domain adaptation of pre-trained language models for french tweets. In *Proceedings of the seventh workshop on Noisy User-generated Text*, pages 445–450.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammu: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics*, page 103982.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. Barthez: A skilled pretrained french sequence-to-sequence model. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9369–9390.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *8th International Conference on Learning Representations (ICLR)*, pages 1–17.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, page 2100707.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the twelfth Language Resources and Evaluation Conference (LREC)*, pages 2479–2490.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, and Veselin Stoyanov andLuke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. Transformers for clinical coding in spanish. *IEEE access*, pages 72387–72397.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte 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)*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the Association for Computational Linguistics (ACL): Human Language Technologies*, pages 746–751.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization. pages 24–30.

Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Proceedings of the first workshop on Scholarly Document Processing*, pages 104–112.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. pages 1–12.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, pages 1–67.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on Biomedical Language Processing*, pages 143–154.

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.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information processing Systems (NIPS)*.

Yang You, Jing Li, Sashank J. Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *8th International Conference on Learning Representations (ICLR)*, pages 1–38.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, pages 17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 11328–11339.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, pages 1218–1227.

Appendix

A Tokenizers comparison and normalization

Figure 1 depicts the steps taken during tokenization with an example and compares Unigram tokenizers trained from scratch and the tokenizer from CamemBERT(Martin et al., 2020).

Different tokenizers are trained from scratch and are compared with one another according to their performance. Figure 2 shows the performance of Unigram, BPE and WordPiece tokenization algorithms. Unigram tokenization have higher proportion of words and learns faster than other tokenizers. It has also achieved the best results in training `AliBERT` and fine-tuning tasks. In Figure 2, tokenizers with a legend "_L_" describes that the text is lower cased and "_NoNo_" shows that the normalization step is ignored during training the tokenizer.

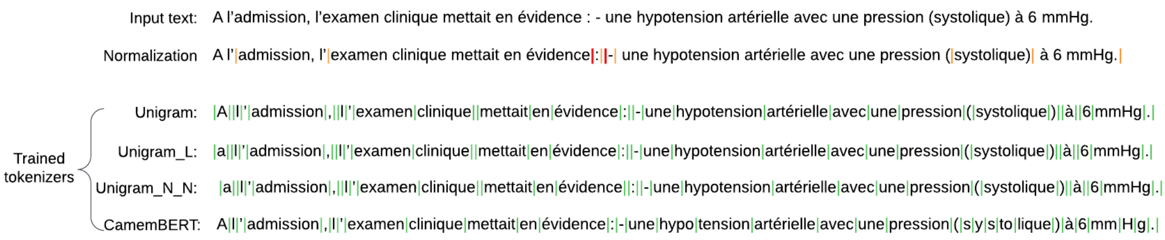

Input text: A l'admission, l'examen clinique mettait en évidence : - une hypotension artérielle avec une pression (systolique) à 6 mmHg.
Normalization A l'|admission, l'|examen clinique mettait en évidence|:||-| une hypotension artérielle avec une pression (|systolique)| à 6 mmHg.|
Trained tokenizers
Unigram: |A||l|'|admission|,||l|'|examen|clinique||mettait|en|évidence|:||-|une|hypotension|artérielle|avec|une|pression|(|systolique|)||à||6|mmHg|.|
Unigram_L: |a||l|'|admission|,||l|'|examen|clinique||mettait|en|évidence|:||-|une|hypotension|artérielle|avec|une|pression|(|systolique|)||à||6|mmHg|.|
Unigram_N_N: |a||l|'|admission|,||l|'|examen|clinique||mettait|en|évidence||:||-|une|hypotension|artérielle|avec|une|pression|(|systolique|)||à||6|mmHg|.|
CamemBERT: A|l|'|admission|,|l|'|examen|clinique|mettait|en|évidence|:|-|une|hypo|tension|artérielle|avec|une|pression|(|s|y|s|to|lique|)|à|6|mm|H|g|.|

Figure 1: Normalization and tokenization example. During normalization step the input text is normalized by adding a space after the punctuation (shown by the orange vertical lines) and removing a space before it (shown by the red vertical lines) and then used to train the tokenizer (Unigram). The Unigram tokenizers are trained from scratch while developing AliBERT, Unigram uses text input as it is (does not change the cases), Unigram_L lower cased the input text and Unigram_N_N is the not-normalized version of Unigram and CamemBERT is the tokenizer used by CamemBERT (Martin et al., 2020), a French PLM.

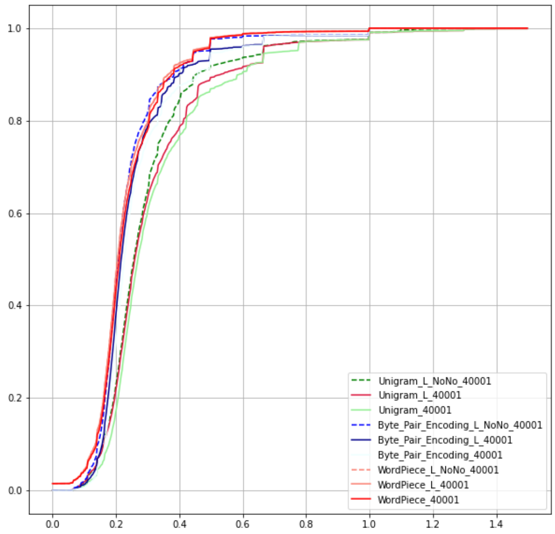


Figure 2: Proportion of individual words with less than x words

B Optimization:

The models trained using LAMB optimizer trained much faster than their counter part (using ADAM). However the performance of the models trained with LAMB was not as good as the models trained with ADAM. Figure 3 shows the comparison of time taken to train using LAMB and ADAM atomizers on our models. The loss of the model quickly reduces when LAMB optimizer is used during training.

Table 6 compares two same models with different optimizers. AliBERT uses ADAM optimizer and AliBERT-LAMB uses LAMB optimizer for pre-training. The two models are compared on NER task on the CAS dataset. AliBERT outperformed AliBERT-LAMP in terms of precision (p), recall (r) and f1 score (F1) for all the entity types except "Pathology".

C MLM examples

Figure 4 presents few biomedical text examples for the prediction of masked words. Predicted words colored in green are the correct predictions. Blue colors shows the prediction is correct in the top 2 predictions, purple color depicts that the correct prediction is the top 3 and the red colors show the correct word has not been predicted. As can be seen, Figure 4 AliBERT and AliBERT-ELECTRA outperformed the two French PLMs. This confirms that the need for training domain-specific language models, specifically B-PLMs.

D NER finetunning dataset

The two publicly available name entity recognition (NER) datasets used for fintunning and evaluating our models are CAS and QUAERO NER datasets which are described in Table 7 and Table 8 respectively. We have selected the biomedical entities

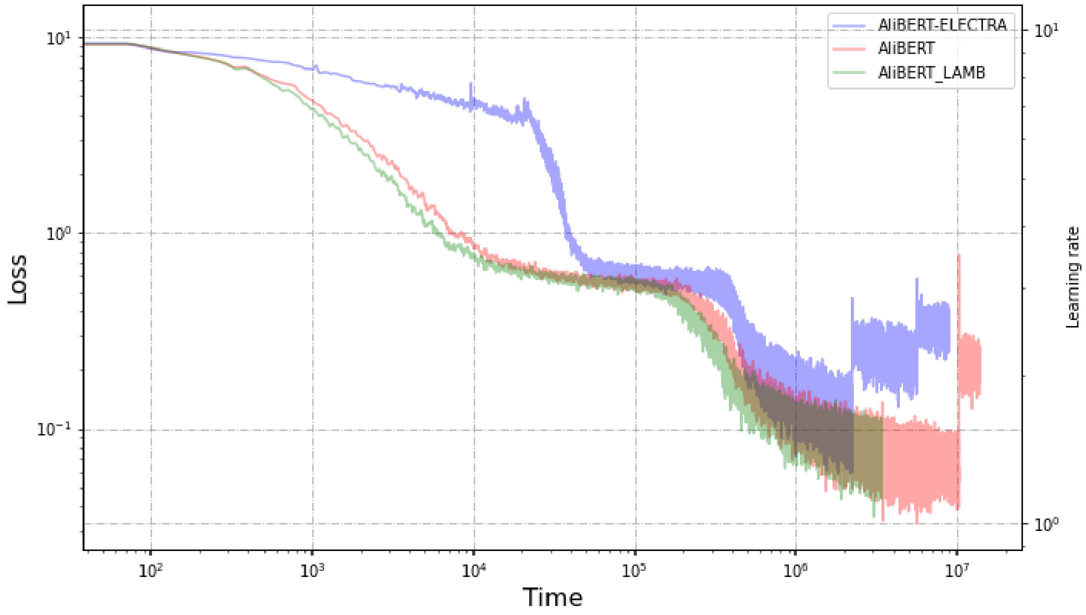
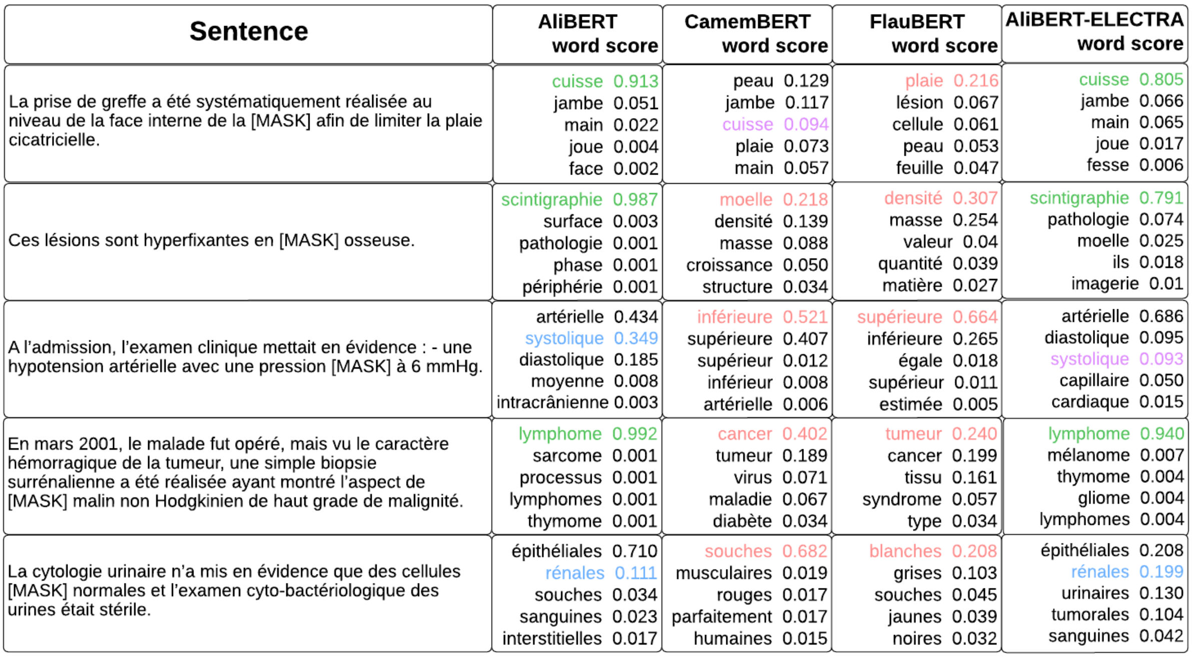


Figure 3: Training time comparison between models using the ADAM and LAMB optimizer. The latter allows for faster training but does not lead to better performance.

	Models' performances on CAS dataset					
	AliBERT			AliBERT-LAMB		
Entities	P	R	F1	P	R	F1
Substance	**0.96**	**0.91**	0.90	0.95	0.87	0.88
Symptom	**0.96**	**0.98**	**0.97**	0.95	0.97	0.96
Anatomy	**0.97**	**0.97**	**0.98**	0.97	0.95	0.96
Value	**0.98**	**0.99**	**0.98**	0.92	0.81	0.86
Pathology	0.81	0.39	0.52	**0.87**	**0.52**	**0.65**

Table 6: French Biomedical named entity recognition (NER) ADAM and LAMP optimizer comparison. Performance in bold is the best achieved for the entity in question and the metrics in question

from the whole datasets.



Sentence	AliBERT word score	CamemBERT word score	FlauBERT word score	AliBERT-ELECTRA word score
La prise de greffe a été systématiquement réalisée au niveau de la face interne de la [MASK] afin de limiter la plaie cicatricielle.	cuisse 0.913 jambe 0.051 main 0.022 joue 0.004 face 0.002	peau 0.129 jambe 0.117 cuisse 0.094 plaie 0.073 main 0.057	plaie 0.216 lésion 0.067 cellule 0.061 peau 0.053 feuille 0.047	cuisse 0.805 jambe 0.066 main 0.065 joue 0.017 fesse 0.006
Ces lésions sont hyperfixantes en [MASK] osseuse.	scintigraphie 0.987 surface 0.003 pathologie 0.001 phase 0.001 périphérie 0.001	moelle 0.218 densité 0.139 masse 0.088 croissance 0.050 structure 0.034	densité 0.307 masse 0.254 valeur 0.04 quantité 0.039 matière 0.027	scintigraphie 0.791 pathologie 0.074 moelle 0.025 ils 0.018 imagerie 0.01
A l'admission, l'examen clinique mettait en évidence : - une hypotension artérielle avec une pression [MASK] à 6 mmHg.	artérielle 0.434 systolique 0.349 diastolique 0.185 moyenne 0.008 intracrânienne 0.003	inférieure 0.521 supérieure 0.407 supérieur 0.012 inférieur 0.008 artérielle 0.006	supérieure 0.664 inférieure 0.265 égale 0.018 supérieur 0.011 estimée 0.005	artérielle 0.686 diastolique 0.095 systolique 0.093 capillaire 0.050 cardiaque 0.015
En mars 2001, le malade fut opéré, mais vu le caractère hémorragique de la tumeur, une simple biopsie surrénalienne a été réalisée ayant montré l'aspect de [MASK] malin non Hodgkinien de haut grade de malignité.	lymphome 0.992 sarcome 0.001 processus 0.001 lymphomes 0.001 thymome 0.001	cancer 0.402 tumeur 0.189 virus 0.071 maladie 0.067 diabète 0.034	tumeur 0.240 cancer 0.199 tissu 0.161 syndrome 0.057 type 0.034	lymphome 0.940 mélanome 0.007 thymome 0.004 gliome 0.004 lymphomes 0.004
La cytologie urinaire n'a mis en évidence que des cellules [MASK] normales et l'examen cyto-bactériologique des urines était stérile.	épithéliales 0.710 rénales 0.111 souches 0.034 sanguines 0.023 interstitielles 0.017	souches 0.682 musculaires 0.019 rouges 0.017 parfaitement 0.017 humaines 0.015	blanches 0.208 grises 0.103 souches 0.045 jaunes 0.039 noires 0.032	épithéliales 0.208 rénales 0.199 urinaires 0.130 tumorales 0.104 sanguines 0.042

Figure 4: MLM prediction examples and comparison between different Language Model for French Text. For each sentence where a word has been masked, the list of the first five most probable words according to the model are given. The colors show the position of the correct prediction, i.e. green is 1^{st}, blue is 2^{nd}, purple is 3^{rd} and red indicates the correct word is not within the list.

Annotation	Occurrences	Description
Substance	2,009	Refers to the pharmacological substances used by the patient (drugs, commercial names and generics)
Symptom	5,240	Entities that are used to make a diagnosis that reveals the pathology of the patient.
Anatomy	4,780	Refers to all anatomical parts (arms, cells, cytoplasm, etc.)
Value	1,743	Refers to values and units, grades, etc. corresponding to examination results, or descriptions of Symptoms
Pathology	764	Concerns diseases and all that is pathological (adenocarcinoma, carcinoma, fistula, etc.)

Table 7: Number of annotations in CAS (NER) dataset used for evaluation

Annotation	Occurrences	Description
Anatomy	1,464	A UMLS concept that refers to a particular part of the body
Chemical	1,028	Refers to chemicals and drugs inside and outside of the body, i.e. protein, enzyme, clinical drug, etc.
Device	126	Includes all devices that are used in the biomedical domain i.e, medical, drug delivery and medical devices
Disorder	2,825	Refers to any abnormality or disease of the body. E.g, disease, symptom, etc.
Procedure	1,631	Refers to procedures and activities practices in the biomedical domain.

Table 8: Number of annotations in QUAERO-MEDLINE NER dataset used for evaluation