# CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data

**Rian Touchent, Laurent Romary, Eric de La Clergerie**
Inria, Sorbonne Université
2 rue Simone IFF 75012 Paris, 21 rue de l'école de médecine 75006 Paris
{rian.touchent,laurent.romary,eric.de_la_clergerie}@inria.fr

### Abstract

Clinical data in hospitals are increasingly accessible for research through clinical data warehouses. However these documents are unstructured and it is therefore necessary to extract information from medical reports to conduct clinical studies. Transfer learning with BERT-like models such as CamemBERT has allowed major advances for French, especially for named entity recognition. However, these models are trained for plain language and are less efficient on biomedical data. Addressing this gap, we introduce CamemBERT-bio, a dedicated French biomedical model derived from a new public French biomedical dataset. Through continual pre-training of the original CamemBERT, CamemBERT-bio achieves an improvement of 2.54 points of F1-score on average across various biomedical named entity recognition tasks, reinforcing the potential of continual pre-training as an equally proficient yet less computationally intensive alternative to training from scratch. Additionally, we highlight the importance of using a standard evaluation protocol that provides a clear view of the current state-of-the-art for French biomedical models.

**Keywords:** EHR, clinical NLP, CamemBERT, information extraction, biomedical, named entity recognition

## 1. Introduction

In recent years, there has been a development of clinical data warehouses (CDWs) in hospitals. These are clinical databases aimed at being more accessible for research purposes. These documents represent an opportunity for massive clinical studies using real data. They can take various forms within Electronic Health Records (EHR), such as reports, medical imaging, or prescriptions. However, most of the information is found in clinical reports. It is estimated that up to 80% of entities are missing from other modalities (Raghavan et al., 2014). Although these data are highly valuable, they are unstructured, which requires preprocessing before they can be used in a clinical study.

BERT-based models (Devlin et al., 2019) consistently demonstrate state-of-the-art results for a wide range of natural language processing tasks. The adaptation of BERT to the French language, particularly with the CamemBERT model (Martin et al., 2020), has replicated these performances in French natural language processing. CamemBERT is based on RoBERTa (Liu et al., 2019), which is a more efficient version of BERT. It is trained on a French corpus extracted from the web called OSCAR (Ortiz Suárez et al., 2019).

To extract information from medical reports, it is necessary to have high-performing language models trained on French clinical data, particularly for named entity recognition. It is possible to simply use CamemBERT; however, the results of this model on biomedical data are disappointing (Cardon et al., 2020), as it exhibits lower performance compared to heuristic models on certain evaluation datasets. These results are predictable because CamemBERT is trained on plain language, often sourced from web pages such as forums. However, biomedical data, especially clinical data, are significantly different. They contain technical terms that are very rare or absent in everyday language, and they have a radically distinct style, often telegraphic, rarely consisting of complete sentences, with varying abbreviations.

One of the major challenges with healthcare data warehouses is data confidentiality. These data are regulated and subject to strong regulations by the CNIL (French Data Protection Authority). As a result, adaptations of CamemBERT to the biomedical domain conducted within hospital infrastructures (Dura et al., 2022) cannot be publicly released. Their training datasets are subject to publication constraints. These constraints also apply to the resulting models. Therefore, it is not possible to exchange these models between different healthcare institutions. A publicly available model would not have these constraints and could be used in various institutions.

Using continual-pretraining on a new French biomedical corpus, we introduce a new model named CamemBERT-bio, which shows a 2.54 points improvement in F-score on several French biomedical named entity recognition tasks. Furthermore, we engage in a discussion on the evaluation of French clinical models, emphasizing the importance of adhering to established stan-

*LREC-COLING 2024*, pages 2692–2701
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

dard practices. By following this methodology, we successfully showcase the effectiveness of continual-pretraining in the context of a French model, which contrast with recent suggestions pertaining to the same domain and language (Labrak et al., 2023).

In this article, we present three main contributions :

- The creation of a new public French dataset specialized in the biomedical domain.
- The introduction of a publicly available adaptation of CamemBERT for the biomedical domain, which demonstrates improved performance on named entity recognition tasks.
- The demonstration that continual-pretraining on a French model is successful, necessitating a reevaluation of previous work due to the impact of evaluation methodology on result interpretation.

## 2. Related Works

Research on adapting language models to new domains is extensive. Gururangan et al. (2020) demonstrate that a second phase of pre-training on a target domain can improve performance on various tasks, even when the target domain corpus is small in size. In the biomedical domain, it has been observed that there can be up to a 3-point increase in F-measure compared to the same model without the second phase of pre-training.

This study by Gururangan et al. (2020) has inspired the creation of new models based on BERT, utilizing a second phase of pre-training on various specialized domains. Lee et al. (2019) introduced BioBERT, a BERT-based model specialized for biomedical text in English. BioBERT demonstrates improved performance on various biomedical NLP tasks, including a $0.62\%$ F-measure improvement on named entity recognition, a $2.80\%$ F-measure improvement on relation extraction, and a $12.24\%$ MMR improvement on question-answering. The second phase of pre-training is conducted on a corpus extracted from PubMed and PMC, consisting of approximately 18 billion words from biomedical scientific articles. While the corpus is substantial and solely composed of scientific-style text, performance gains are observed across all text styles. The presence of medical vocabulary in the corpus likely contributes to significant improvement compared to general language models.

Training new models from scratch is also a viable approach. This is explored in SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2022), two models specialized in biomedical scientific articles. PubMedBERT demonstrates that this method yields better performance than models trained with a second phase of specialization. However, the performance gains are relatively modest, and this approach is more computationally expensive. Starting from scratch requires longer training times and a larger corpus to achieve comparable performance.

For the French language, the reference models are CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). Several works have attempted to adapt CamemBERT to the biomedical domain.

Copara et al. (2020) explored a second phase of pre-training on 31,000 French biomedical scientific articles. However, they did not observe a significant improvement on a clinical named entity recognition task using the large version of CamemBERT. This could be explained by the combination of a relatively small corpus (31k documents compared to the 18 billion words in BioBERT) and the large version of the CamemBERT model.

Le Clercq de Lannoy et al. (2022) also adapted CamemBERT to the biomedical domain. They aggregated documents from various sources, including PubMed, Cochrane, ISTEX, and Wikipedia, forming a larger partially public corpus of approximately 136 million words. They observed a 2-point improvement in F-measure on a named entity recognition evaluation set composed of drug notices (EMEA), but no significant improvement on a set composed of scientific article titles (MEDLINE).

Dura et al. (2022) continued the pre-training of CamemBERT on 21 million clinical documents from the APHP (Assistance Publique - Hôpitaux de Paris) clinical data warehouse. They observed a significant 3% improvement on APMed, a private clinical named entity recognition dataset owned by APHP. They also achieved similar scores to CamemBERT on EMEA and MEDLINE. Their new model performs better on clinical data, yet it obtains scores similar to CamemBERT in other biomedical domains.

Labrak et al. (2023) introduced a public French biomedical model named DrBERT. Through their experiments, they explored both continual-pretraining and from-scratch training strategies. Their findings indicated a superior performance when training from scratch, suggesting that continual-pretraining with CamemBERT for French biomedical data may not be as effective. However, when applied to PubMedBERT, continual-pretraining yielded results nearly on par with the from-scratch approach.

Finally, Berhe et al. (2023) introduced AliBERT, which was trained on a French biomedical corpus primarily comprising articles from ScienceDirect and theses collected through Sudoc. The model leverages a new regularized Unigram-based tokenizer and underwent extensive training on 48

| Corpus | Details | Size |
|---|---|---|
| ISTEX | Scientific literature | 276 M |
| CLEAR | Drug leaflets | 73 M |
| E3C | Clinical cases and leaflets | 64 M |
| Total | | 413 M |

Table 1: Composition of the biomed-fr corpus (in millions of words)

GPUs for a total duration of 20 hours. Their pre-trained model outperforms notable French non-domain-specific models, such as CamemBERT and FlauBERT, in two biomedical downstream tasks. Unfortunately, the model is not currently available.

# 3. CamemBERT-bio

## 3.1. Corpus : biomed-fr

First, we built a French biomedical corpus composed exclusively of public documents to minimize the usage constraints mentioned earlier. The documents come from three different sources (see Table.1), the main one being ISTEX. This new corpus, named *biomed-fr*, consists of 413 million words, equivalent to 2.7 GB of data. Martin et al. (2020) have shown that with only 4 GB of data, it is possible to achieve performance almost comparable to the model trained with the 138 GB OSCAR dataset (Ortiz Suárez et al., 2019). For an adaptation of CamemBERT to the biomedical domain, this amount of data can be considered sufficient.

**ISTEX** The ISTEX database contains references to 27 million scientific publications. We extracted 108,183 French documents published in a biology or medical journal since 1990. Articles published before this date often contain numerous typographical errors, as they are often scanned articles that require optical character recognition algorithms, resulting in a certain number of errors. Such errors are found to a lesser extent in articles published after 1990. Some documents, although in French, contain passages in English. Therefore, there is an indeterminate amount of English in this corpus. However, it is unlikely that this will significantly impact pre-training. Typographical errors and the presence of other languages are aspects that can be addressed in future versions of biomed-fr.

**CLEAR** The CLEAR corpus (Grabar and Cardon, 2018) consists of encyclopedia articles, drug leaflets, and abstracts of scientific articles. Each document is available in two versions: one in technical language and the other in simplified language. We retrieved all of these documents in both versions. Regarding the drug leaflets, we removed redundant sentences at the beginning and end of each document, such as the website navigation bar from which the documents were extracted or information about the company selling the documents.

**E3C** This corpus (Magnini et al., 2020) is composed of three layers. The first two layers are annotated or semi-annotated and will be used for evaluation. The last layer is not annotated, and that is the one we retrieved. It consists of medical specialty admission competitions, drug leaflets, and medical thesis abstracts. There may be duplicates of some leaflets found in the CLEAR corpus.

**biomed-fr-small** By randomly selecting 10% of content from biomed-fr, we created a smaller corpus called *biomed-fr-small*. The corpus allows us to study the impact of corpus size.

## 3.2. Pre-training Strategies

For the adaptation of CamemBERT to the biomedical domain, we conducted a second phase of pre-training on both versions of the biomed-fr corpus, starting from the weights and configuration of the camembert-base model. We applied the Masked Language Modeling (MLM) task with whole-word masking, following the method of Martin et al. (2020). We used the Adam optimizer (Kingma and Ba, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, and a learning rate of $5e-5$. We performed 50,000 steps over 39 hours using two Tesla V100 GPUs. A batch size of 8 per GPU and gradient accumulation over 16 steps were used to achieve an effective batch size of 256.

## 3.3. Fine-tuning and Evaluation

Regarding model evaluation, we collected three named entity recognition evaluation datasets. These datasets cover various styles, allowing us to assess the model's versatility across different subdomains of biomedicine.

**QUAERO** The QUAERO corpus (Névéol et al., 2014) consists of two evaluation sets: EMEA, containing drug leaflets, and MEDLINE, containing scientific article titles. The entities are manually annotated following 10 semantic groups from the UMLS (Lindberg et al., 1993). As some of these entities are nested, we kept only the entities with the coarsest granularity. F-scores are calculated in the same way.

| | | | | CamemBERT-bio | |
|---|---|---|---|---|---|
| Style | Dataset | Score | CamemBERT | biomed-fr-small | biomed-fr |
| Clinical | CAS1 | F1 | 70.50 ± 1.75 | 72.94 ± 1.12 | **73.03 ± 1.29** |
| | | P | 70.12 ± 1.93 | **72.97 ± 0.84** | 71.71 ± 1.61 |
| | | R | 70.89 ± 1.78 | 72.92 ± 1.39 | **74.42 ± 1.49** |
| | CAS2 | F1 | 79.02 ± 0.92 | 80.00 ± 0.32 | **81.66 ± 0.59** |
| | | P | 77.3 ± 1.36 | 78.29 ± 0.91 | **80.96 ± 0.91** |
| | | R | 80.83 ± 0.96 | 81.80 ± 0.48 | **82.37 ± 0.69** |
| | E3C | F1 | 67.63 ± 1.45 | 67.96 ± 1.85 | **69.85 ± 1.58** |
| | | P | 78.19 ± 0.72 | 77.41 ± 1.01 | **79.11 ± 0.42** |
| | | R | 59.61 ± 2.25 | 60.57 ± 2.32 | **62.56 ± 2.50** |
| Leaflets | EMEA | F1 | 74.14 ± 1.95 | 75.93 ± 2.42 | **76.71 ± 1.50** |
| | | P | 74.62 ± 1.97 | 76.23 ± 2.27 | **76.92 ± 1.96** |
| | | R | 73.68 ± 2.22 | 75.63 ± 2.61 | **76.52 ± 1.62** |
| Scientific | MEDLINE | F1 | 65.73 ± 0.40 | 65.48 ± 0.31 | **68.47 ± 0.54** |
| | | P | 64.94 ± 0.82 | 64.43 ± 0.50 | **67.77 ± 0.88** |
| | | R | 66.56 ± 0.56 | 66.56 ± 0.16 | **69.21 ± 1.32** |

Table 2: F-scores on different biomedical named entity recognition tasks

**E3C** For evaluation, unlike the biomed-fr corpus, we use layers 1 and 2. These layers contain documents of different types, including clinical cases extracted from scientific articles. Layer 2 is semi-annotated, and it is used as the training set for fine-tuning, with 10% dedicated to the validation set. We evaluate on layer 1, which is fully manually annotated. There is only one class, and the objective is to find clinical entities in the text, regardless of their type.

**CAS** The CAS corpus (Grouin et al., 2019) also consists of clinical cases from scientific articles. We focus on task 3 of DEFT 2020 (Cardon et al., 2020), which is an information extraction task based on CAS. It includes two subtasks, and thus two sets of annotations. In the first subtask, two classes need to be identified: *pathology* and *signs or symptoms*. The second subtask concerns associated information, including *anatomy*, *dose*, *examination*, *mode*, *timing*, *substance*, *treatment*, and *value*. These two tasks will be referred to as CAS1 and CAS2, respectively.

**Fine-tuning** For fine-tuning, we used Optuna (Akiba et al., 2019) for hyperparameter selection. We set the learning rate to $5e-5$, the warmup ratio to 0.224, and the batch size to 16. We performed 2000 steps. Predictions were made using a simple linear layer on top of the model. None of the CamemBERT layers were frozen.

**Evaluation** Scores are measured using the seqeval tool (Nakayama, 2018) in strict mode with micro-average and the "**IOB2**" scheme. For each evaluation, the best fine-tuned model on the validation set is selected to measure the final score on the test set. We average the results over 10 evaluations with different seeds.

## 4. Results and Discussion

**CamemBERT vs CamemBERT-bio** We observe a significant performance gain on all evaluation datasets with our new model (see Table.2). On average, we achieve a 2.54-point improvement in F-score. This gain is observed across all styles, demonstrating the model's versatility for both clinical and scientific domains.

**biomed-fr-small vs biomed-fr** We observe a decrease in performance with the biomed-fr-small dataset, but there is still a significant gain on certain datasets compared to CamemBERT. This confirms that the size of the corpus positively influences the performance, even in a specialized domain like biomedicine.

**Comparison with the state of the art** We compared the performance of CamemBERT-bio with various previously mentioned approaches (see Table.3). CamemBERT-bio achieves the best results for almost all evaluation datasets. Dura et al. (2022) did not observe improvement on EMEA and MEDLINE compared to CamemBERT because their pretraining corpus (see Table.4) consists of documents from APHP, making it a less diverse corpus. However, they gain several points on their evaluation dataset, which is also based on APHP documents. Mulligen et al. (2016) presents the highest score on MEDLINE and the best recall on EMEA. Their

| | | CAS1 | CAS2 | EMEA | | | MEDLINE | | |
|---|---|---|---|---|---|---|---|---|---|
| Evaluator | Authors | F1 | F1 | F1 | P | R | F1 | P | R |
| seqeval | Dura et al. (2022)-fine-tuned | - | - | 72.90 | - | - | 59.70 | - | - |
| | Dura et al. (2022)-from-scratch | - | - | 69.30 | - | - | 60.10 | - | - |
| | ours | **73.03** | **81.66** | **76.71** | **76.92** | **76.52** | **68.47** | **67.77** | **69.21** |
| BRATeval | Le Clercq de Lannoy et al. (2022) | | | 67.4 | 73.4 | 62.2 | 55.3 | 62.2 | 49.7 |
| | Mulligen et al. (2016) | - | - | 74.9 | 71.6 | **78.5** | **69.8** | 68 | **71.6** |
| | Copara et al. (2020) | 61.53 | 73.7 | - | - | - | - | - | - |
| | ours | **84.97** | **83.25** | **77.80** | **79.77** | 75.93 | 56.16 | **75.33** | 44.82 |

Table 3: Comparison of CamemBERT-bio with different approaches on the 4 named entity recognition tasks. In the first part of the table, scores are measured with seqeval (Nakayama, 2018), and in the second part with BRATeval, which is the evaluation tool provided for the CLEF eHealth Evaluation lab 2016 campaign (Névéol et al., 2016).

| | Pre-training corpus | |
|---|---|---|
| Authors | Origin | Size[1] |
| Dura et al. (2022) | APHP | 21 MD |
| Le Clercq de Lannoy et al. (2022) | misc | 136 MW |
| Copara et al. (2020) | PubMed | 31 KD |
| ours | biomed-fr | 413 MW |

Table 4: Pre-training corpora of related works (cf. Table.3)

approach is based on a knowledge-based model, allowing them to achieve the best recall on both QUAERO evaluation datasets. Furthermore, their approach is the only one in this table capable of handling nested entities, giving them an advantage.

It is important to note that these different CamemBERT-based approaches have various experimental setups. The presence of CRF layers instead of a simple linear layer after CamemBERT, freezing CamemBERT layers, and variations in hyperparameters are examples of differences observed in addition to the pre-training corpus, which makes the comparison more challenging.

**Tokenization analysis** CamemBERT-bio is a biomedical-adapted model based on CamemBERT. Unlike a new model trained from scratch, it shares the same vocabulary. The vocabulary of CamemBERT was constructed using SentencePiece (Kudo and Richardson, 2018) on an OSCAR sample. Therefore, it is a general-purpose vocabulary designed for everyday language. We can hypothesize that the tokenization of CamemBERT may result in oversplitting of biomedical technical terms.

To investigate this possibility, we trained a specialized tokenizer on biomed-fr-small and calculated the intersection of the two vocabularies (Table.5).

We find a 45% intersection between the two vocabularies, which is quite close to the 42% intersection found by Beltagy et al. (2019) between the vocabulary of BERT and SciBERT. Therefore, there is a significant difference in the most frequent terms.

## 5. Evaluating Models: Methodology and Discussion

In a recent work, Labrak et al. (2023) introduced a new French biomedical language model named DrBERT. The authors contend that performing continual-pretraining on biomedical data from CamemBERT leads to reduced performance. They used a different methodology for the evaluation of named entity recognition, prompting us to examine the implications of each approach on the interpretation of the results.

Our evaluation approach is centered around micro F1, which is measured using seqeval in strict mode with the IOB2 scheme. Their approch is based on weighted F1 with independant token classification. Every token has a label and the model is evaluated for each of them, whereas with seqeval, the model is evaluated for each entity. Notably, the "O" token, representing non-entity, is the most frequent token and thus holds significant weight in the evaluation process. As all tokens are labeled, the number of predictions depends on the tokenizer, thereby influencing the final score. In order to explore the impact of token labeling variations on the evaluation, we conducted an additional experiment on EMEA and MEDLINE where we reproduced the DrBERT methodolody that we named *token-with-O*, and another one where we excluded the "O" token and only labeled the first token of each entity, that will be refered as *entity-without-O*. Furthermore, we included a seqeval strict score using the IOB2

[1] Units: MD (Million Documents), MW (Million Words), KD (Thousand Documents). As we rely on related articles, we can't provide a better estimation.

| Terms | general | specialized |
|---|---|---|
| échocardiographie | écho-cardi-ographie | échocardiographi-e |
| transthoracique | trans-thorac-ique | trans-thoracique |
| glimépiride | g-lim-épi-ride | gli-m-épi-ride |
| cardiopathie | cardio-pathie | cardiopathie |
| diastoliques | dia-s-tol-iques | diastolique-s |

Table 5: Comparison of tokenization between a general-purpose tokenizer and a specialized tokenizer for some biomedical technical terms.

scheme, although it is not directly comparable to our approach due to their implementation of nested entities by concatenating their names to form new entities, while we simply removed them.

We observe significant differences in scores between the two methodologies (Table.6). The methodology *token-with-O* consistently reports higher scores across all tasks compared to the alternative method, *entity-without-O*. However, the reported scores do not precisely align with those presented by Labrak et al. (2023), where DrBERT exhibits improvement over CamemBERT for EMEA, and match CamemBERT for MEDLINE. We posit that these discrepancies arise from disparities in hyperparameters. Nevertheless, the comparison with *entity-without-O* still provides insightful observations.

Notably, the best-performing model varies depending on the chosen methodology for one metric. CamemBERT achieves the highest macro F1 score on the EMEA dataset, whereas with *entity-without-O*, DrBERT-7GB emerges as the top model. This observation underscores the influential role of the "O" class in determining the best-performing model and prompts us to consider the aspect we aim to evaluate.

In terms of macro F1, DrBERT consistently outperforms CamemBERT-bio when evaluated using the *entity-without-O* methodology. On the other hand, CamemBERT-bio exhibits better performance across all other metrics. This indicates that DrBERT demonstrates a more balanced performance across different classes. These findings suggest that these models may possess complementary strengths, and it is important to focus on enhancing the performance of CamemBERT-bio in this aspect.

It is worth noting that the evaluations conducted by Labrak et al. (2023) encompass a wide range of tasks beyond EMEA and MEDLINE. This observation underscores the importance of establishing a unified benchmark to facilitate fair comparisons between models for the french biomedical domain. A noteworthy tool for this purpose is BRATeval from the CLEF eHealth Evaluation lab 2016 campaign (Névéol et al., 2016), as it evaluates based on exact match character offsets, thus avoiding any influence from dataset pre-processing methods.

Furthermore, Labrak et al. (2023) suggested that continual-pretraining from a French model on French biomedical data isn't effective, as it results in an F1-score loss of up to 20 points. However, considering our success with continual-pretraining and the points we raised about the methodology, we suggest a reassessment of these findings.

## 6. Environmental Impact

Considering the environmental implications is crucial when discussing language model training due to its potential compute intensity.

Considering estimated carbon emissions [1] during training, AliBERT emits 10 times the amount of CamemBERT-bio, while DrBERT releases 32 times more (Table.6). While a direct comparison with AliBERT wasn't possible, the minor performance variance with DrBERT, set against the pronounced disparities in computational and environmental costs, leads us to advocate for continual-pretraining as the preferred adaptation method for biomedical language models.

## 7. Limitations

It is important to discuss some limitations of our studies.

Firstly, our training corpus *biomed-fr* is limited in its diversity. This is because we chose to only use publicly available materials, which tend to lean towards scientific content. As a result, our analysis may not fully represent performances on real private clinical data.

Furthermore, we didn't explore any potential biases in our data and some part of our dataset would benefit from further cleaning. These aspects could impact the outputs of our model and should be investigated.

Additionally, our evaluation focused on one task which is Named Entity Recognition. This might mean our understanding of how well our model

[1] Estimations were conducted using the Machine-Learning Impact calculator presented by Lacoste et al. (2019)

| | | EMEA | | | | MEDLINE | | | |
|---|---|---|---|---|---|---|---|---|---|
| Methodology | Model | weighted-f1 | macro-f1 | micro-f1 | seqeval-f1 | weighted-f1 | macro-f1 | micro-f1 | seqeval-f1 |
| *token-with-O* | DrBERT-7GB | 87.45 | 34.95 | - | - | 75.52 | **15.07** | - | - |
| | CamemBERT-bio | **90.37** | 36.27 | - | - | **77.89** | 14.82 | - | - |
| | CamemBERT | 88.33 | **47.45** | - | - | 76.2 | 11.92 | - | - |
| *entity-without-O* | DrBERT-7GB | 66.72 | **24.72** | 68.34 | 59.39 | 60.70 | **10.80** | 63.40 | 50.45 |
| | CamemBERT-bio | **73.53** | 24.15 | **75.05** | **67.58** | **62.04** | 8.695 | **65.44** | **52.9** |
| | CamemBERT | 71.85 | 22.71 | 72.93 | 64.23 | 60.95 | 9.413 | 63.47 | 51.75 |

Table 6: Performance comparison of CamemBERT, CamemBERT-bio, and DrBERT on EMEA and MEDLINE using the evaluation methodology *token-with-O*, along with a modified variant *entity-without-O*. The reported scores are averaged over 10 runs.

| | Training time (hours) | Hardware type | Total GPU-hours | Estimation of carbon emitted (kg CO2 eq.) |
|---|---|---|---|---|
| DrBERT | 20h | 128xV100 | 2560 | 26.11 |
| AliBERT | 20h | 48xA100 | 960 | 8.16 |
| CamemBERT-bio | 39h | 2xV100 | 78 | 0.8 |

Table 7: Carbon emitted estimation based on hardware and training time. We used a rate of 34g CO2eq. per kWh, reflecting the average over the last 12 months in France starting from September 2022. This time frame and location coincide with when and where all experiments were conducted.

performs is restricted. Future studies should aim to assess our model across a wider range of tasks and datasets to get a clearer picture of its strengths and weaknesses.

This underscores the need for further research to address these constraints and enhance our understanding of the subject matter.

## 8. Conclusion and Perspectives

We have introduced a new French biomedical corpus called *biomed-fr* consisting of 413 million words, composed of drug leaflets and documents from scientific literature in medicine and biology. This new corpus has allowed us to adapt CamemBERT to the biomedical domain through a second phase of pre-training. We observe an improvement in performance on all our named entity recognition evaluation datasets, with an average gain of 2.54 F-score points. Our model establishes a new state-of-the-art on these French biomedical language processing tasks. We have some directions for future versions of *biomed-fr*. Firstly, we can further clean the data by removing passages within the documents that are not in French or by excluding documents with a high number of typographical errors. Secondly, we can increase the amount of data. This could involve leveraging archived documents on HAL related to life sciences, particularly those published by INSERM, or retrieving abstracts of French articles from PubMed.

The analysis of tokenization prompts us to consider expanding the vocabulary for CamemBERT-bio. The relatively modest performance gain of PubMedBERT compared to BioBERT and the similar performance of DrBERT compared to CamemBERT-bio despite their specialized vocabulary, the over-segmentation of technical terms and the low intersection rate between the generalist vocabulary and the specialized vocabulary demonstrate the value of the experiment. However, taking into account the comparable performance levels and significant difference in environmental impact, we advocate strongly for the continual-pretraining method in adapting language models to the biomedical domain.

Finally, in recent months, numerous generative models, often with billions of parameters, have demonstrated remarkable performance on biomedical tasks, sometimes surpassing specialized models like BioBERT (Agrawal et al., 2022; Singhal et al., 2023). This is a promising research direction for biomedical information extraction. However, we have reasons to believe that BERT-type models still have value (Lehman et al., 2023). Firstly, in a clinical context, models are often used within healthcare institution infrastructures, which entails resource constraints. It is easier to deploy small specialized models than large generalist models in such cases. Secondly, the use of these generative models often requires accessing remote servers, typically through APIs, which makes their utilization challenging considering the confidentiality constraints imposed on clinical documents.

## 9. Bibliographical References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Rémi Cardon, Natalia Grabar, Cyril Grouin, and Thierry Hamon. 2020. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases ). In *Actes de la 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 1–13, Nancy, France. ATALA et AFCP.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized French Language Models for Biomedical Named Entity Recognition. In *Traitement Automatique des Langues Naturelles (TALN, 27e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France. ATALA.

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. Technical Report arXiv:2207.12940, arXiv. ArXiv:2207.12940 [cs, stat] type: article.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [cs].

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Tiphaine Le Clercq de Lannoy, Romaric Besançon, Olivier Ferret, Julien Tourille, Frédérique Brin-Henry, and Bianca Vieru. 2022. Stratégies d'adaptation pour la reconnaissance d'entités médicales en français (Adaptation strategies for biomedical named entity recognition in French). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 215–225, Avignon, France. ATALA.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR.

Erik M. van Mulligen, Zubair Afzal, Saber A. Akhondi, Dang Vo, and Jan A. Kors. 2016. Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts.

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

Aurélie Névéol, K. Bretonnel Cohen, Cyril Grouin, Thierry Hamon, Thomas Lavergne, Liadh Kelly, Lorraine Goeuriot, Grégoire Rey, Aude Robert, Xavier Tannier, and Pierre Zweigenbaum. 2016. Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016. *CEUR workshop proceedings*, 1609:28–42.

Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. 2014. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Summits on Translational Science Proceedings*, 2014:218–223.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

## 10. Language Resource References

Beltagy, Iz and Lo, Kyle and Cohan, Arman. 2019. *SciBERT: A Pretrained Language Model for Scientific Text*. Association for Computational Linguistics.

Berhe, Aman and Draznieks, Guillaume and Martenot, Vincent and Masdeu, Valentin and Davy, Lucas and Zucker, Jean-Daniel. 2023. *AL-IBERT: A PRETRAINED LANGUAGE MODEL FOR FRENCH BIOMEDICAL TEXT*. Working paper or preprint.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics.

Grabar, Natalia and Cardon, Rémi. 2018. *CLEAR – Simple Corpus for Medical French*. Association for Computational Linguistics.

Grouin, Cyril and Grabar, Natalia and Claveau, Vincent and Hamon, Thierry. 2019. *Clinical Case Reports for NLP*. Association for Computational Linguistics.

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*, 3(1):1–23. ArXiv:2007.15779 [cs].

Kudo, Taku and Richardson, John. 2018. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing*. Association for Computational Linguistics.

Labrak, Yanis and Bazoge, Adrien and Dufour, Richard and Rouvier, Mickael and Morin, Emmanuel and Daille, Béatrice and Gourraud, Pierre-Antoine. 2023. *DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains*. Association for Computational Linguistics.

Le, Hang and Vial, Loïc and Frej, Jibril and Segonne, Vincent and Coavoux, Maximin and Lecouteux, Benjamin and Allauzen, Alexandre and Crabbé, Benoît and Besacier, Laurent and Schwab, Didier. 2020. *FlauBERT: Unsupervised Language Model Pre-training for French*. European Language Resources Association.

Lee, Jinhyuk and Yoon, Wonjin and Kim, Sungdong and Kim, Donghyeon and Kim, Sunkyu and So, Chan Ho and Kang, Jaewoo. 2019. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*.

Lindberg, D. A. and Humphreys, B. L. and McCray, A. T. 1993. *The Unified Medical Language System*.

Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv. ArXiv:1907.11692 [cs].

Bernardo Magnini and Begoña Altuna and Alberto Lavelli and Manuela Speranza and Roberto Zanoli. 2020. *The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases*.

Martin, Louis and Muller, Benjamin and Ortiz Suárez, Pedro Javier and Dupont, Yoann and Romary, Laurent and de la Clergerie, Éric and Seddah, Djamé and Sagot, Benoît. 2020. *CamemBERT: a Tasty French Language Model*. Association for Computational Linguistics.

Névéol, Aurélie and Grouin, Cyril and Leixa, Jeremy and Rosset, Sophie and Zweigenbaum, Pierre. 2014. *The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization*.

Pedro Javier Ortiz Suárez and Benoît Sagot and Laurent Romary. 2019. *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures*. Leibniz-Institut für Deutsche Sprache, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019.

## Appendices

|  |  |  |  | CamemBERT-bio |  |
|---|---|---|---|---|---|
| Style | Dataset | Score | CamemBERT | biomed-fr-small | biomed-fr |
| Clinical | CAS1 | F1 | 67.85 | 70.18 | **71.37** |
|  |  | P | 73.13 | 74.11 | **75.22** |
|  |  | R | 64.38 | 67.39 | **68.34** |
|  | CAS2 | F1 | 72.40 | **74.69** | 74.32 |
|  |  | P | 73.22 | **74.46** | 72.46 |
|  |  | R | 72.03 | 75.23 | **77.56** |
| Leaflets | EMEA | F1 | 50.74 | 53.1 | **55.69** |
|  |  | P | 51.99 | 55.88 | **56.01** |
|  |  | R | 50.67 | 51.78 | **56.09** |
| Scientific | MEDLINE | F1 | 45.09 | 47.73 | **48.18** |
|  |  | P | 46.57 | 48.04 | **49.16** |
|  |  | R | 47.18 | **52.62** | 50.38 |

Table 8: F-scores on different biomedical named entity recognition tasks with macro-average

## A. Macro-average

In our main results (Table.2), scores are measured using the seqeval tool (Nakayama, 2018) in strict mode with micro-average and the "**IOB2**" scheme. We also conducted the same evaluation using macro-average (Table.8).