DrBenchmark: A Large Language Understanding Evaluation Benchmark for French Biomedical Domain

Yanis Labrak[*1,2], **Adrien Bazoge**[*3,4], **Oumaima El Khettari**[4]
Mickael Rouvier[1], **Pacôme Constant dit Beaufils**[5], **Natalia Grabar**[6], **Béatrice Daille**[4]
Solen Quiniou[4], **Emmanuel Morin**[4], **Pierre-Antoine Gourraud**[3] and **Richard Dufour**[1,4]
[1]LIA, Avignon Université [2]Zenidoc
[3]Nantes Université, CHU Nantes, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France
[4]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[5]Nantes Université, CHU Nantes, Service de Neuroradiologie diagnostique et interventionnelle, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[6]UMR 8163 – STL CNRS, Université de Lille
{first.last}@{univ-avignon.fr, univ-nantes.fr, chu-nantes.fr, univ-lille.fr}

Abstract

The biomedical domain has sparked a significant interest in the field of Natural Language Processing (NLP), which has seen substantial advancements with pre-trained language models (PLMs). However, comparing these models has proven challenging due to variations in evaluation protocols across different models. A fair solution is to aggregate diverse downstream tasks into a benchmark, allowing for the assessment of intrinsic PLMs qualities from various perspectives. Although still limited to few languages, this initiative has been undertaken in the biomedical field, notably English and Chinese. This limitation hampers the evaluation of the latest French biomedical models, as they are either assessed on a minimal number of tasks with non-standardized protocols or evaluated using general downstream tasks. To bridge this research gap and account for the unique sensitivities of French, we present the first-ever publicly available French biomedical language understanding benchmark called DrBenchmark. It encompasses 20 diversified tasks, including named-entity recognition, part-of-speech tagging, question-answering, semantic textual similarity, and classification. We evaluate 8 state-of-the-art pre-trained masked language models (MLMs) on general and biomedical-specific data, as well as English specific MLMs to assess their cross-lingual capabilities. Our experiments reveal that no single model excels across all tasks, while generalist models are sometimes still competitive.

Keywords: NLP evaluation, Benchmarking, Medical domain, French language, Transformers

1. Introduction

For the past few years, the field of Natural Language Processing (NLP) has witnessed major breakthroughs, particularly in the area of language modeling. Newer approaches such as the self-attention mechanism (Vaswani et al., 2017), Sparse Transformer (Child et al., 2019), and Replaced Token Detection (Clark et al., 2020) have emerged. These advancements have enabled the application of language models pre-trained on large corpora of textual data to a wide range of NLP tasks.

The evaluation of proposed models and approaches is an essential step in verifying their quality and performance. In the context of pre-trained language models (PLMs), this validation typically involves assessing their performance on targeted downstream tasks. This task-selection process is crucial, as the performance of models can vary depending on the chosen ones. Consequently, a model that performs well in one context may deliver disappointing results in another one. To address this issue and validate the models' generalizability, evaluation benchmarks have emerged, typically encompassing diverse sets of tasks. Hence, the availability of evaluation benchmarks plays a vital role in driving continuous progress, fostering the development of community members, and facilitating fair comparison between models.

While numerous open benchmarks exist for general tasks in NLP across multiple languages, the biomedical field remains an area with relatively few proposed benchmarks, mainly for English and Chinese, facilitating the availability of many biomedical models in these two languages. Even if the gap in other languages is beginning to narrow with new specialized models, the development of evaluation platforms has been comparatively slower.

Although the French language is generally considered as well-endowed, it is notably lacking in evaluation resources within the biomedical field. To address this issue, we introduce DrBenchmark, the first comprehensive open benchmark for the French biomedical domain, comprising 20 diverse tasks. These tasks encompass part-of-speech (POS) tagging, named-entity recognition (NER), classification, question-answering (QA), and semantic textual similarity (STS).

We also perform a quantitative study of 8 pre-trained state-of-the-art masked language models

LREC-COLING 2024, pages 5376–5390

(MLMs) with different configurations (languages and domains) on DrBenchmark. Our in-depth analysis integrates a comparison of these models' performance, fine-tuning with limited data and word tokenization. Our main contributions are:

- DrBenchmark, an original evaluation framework for French biomedical NLP domain aggregating a large set of 20 diversified, proven and challenging tasks.
- A quantitative study using our proposed biomedical benchmark on a wide range of 8 MLMs based on varied architectures, data sources and training strategies.
- The release under CC BY-SA 4.0 license on HuggingFace[1] of a new open biomedical dataset with clinical cases manually annotated into the 22 International Classification of Diseases 10th Revision (ICD-10) categories.
- A modular, reproducible and easily customizable automated protocol using identical training and evaluation scripts allowing a simple and fair comparison, with, as input, only the evaluated language models. DrBenchmark is freely available under MIT license on GitHub, HuggingFace and summarized as a leaderboard on the website[2].

The paper is organized as follows: Section 2 briefly introduces historical NLP benchmarks, including biomedical ones. Section 3 presents our proposed benchmark for French biomedical domain, with a focus on the downstream tasks. Section 4 presents the experimental protocol while Section 5 details the results and provides an analysis of the studied pre-trained models. Section 6 finally concludes the work and opens some perspectives.

2. Related work

In the recent years, several NLP open benchmarks have been created to facilitate direct comparison between proposed approaches. Among the earliest benchmarks, DecaNLP (McCann et al., 2018) and GLUE (Wang et al., 2018), focused on general English language understanding tasks rather than being specific to a particular domain. Thus, GLUE gathers nine tasks including text classification (linguistic acceptability, sentiment analysis, etc.), semantic analysis (paraphrase verification, sentence similarity, etc.), QA, coreference detection, and natural language inference (NLI). Following a similar concept, the French counterpart to GLUE, known as FLUE (Le et al., 2020), consists of 7 general-domain tasks in French, covering areas such as text classification, paraphrasing, NLI, parsing, and word sense disambiguation.

In the case of specialized domains, general benchmarks may not adequately evaluate the performance of in-domain models. Specifically, within the biomedical domain, only few benchmarks have been proposed, and they primarily focus on few languages. For instance, platforms like BLURB (Gu et al., 2021) and BLUE (Peng et al., 2019) predominantly offer benchmarks for English, while CBLUE (Zhang et al., 2022a) caters to the Chinese language. To provide more specific information, BLURB integrates 13 tasks, including NER, information and relation extraction, sentence similarity, text classification, and QA. BLUE encompasses 10 tasks, such as NER, sentence similarity, relation extraction, text classification, and inference. On the other hand, CBLUE covers 8 tasks, including NER, information extraction, text and intent classification, sentence similarity, and query relevance.

To our knowledge, aside the multilingual benchmark BigBIO (Fries et al., 2022) which includes only 4 corpora for French and is initially intended for generative text completion under zero-shot scenario, no large benchmark specialized in the French biomedical field exists. This makes the comparison of recent specialized models, such as (Labrak et al., 2023; Touchent et al., 2023; Copara et al., 2020; Berhe et al., 2023), extremely challenging.

3. DrBenchmark Overview

Our proposed benchmark comprises 20 French biomedical language understanding tasks, one of which is specifically created for this benchmark. The descriptions and statistics of these tasks are presented in Table 1. DrBenchmark encompasses the following overall aspects:

1. **A variety of tasks with different requirements and objectives:** Part-of-Speech (POS) tagging, Multi-class, Multi-label and Intent classification, Named-Entity Recognition (NER), Multiple-Choice Question-Answering (MCQA), and Semantic Textual Similarity (STS).

2. **A diverse range of data origins:** Scientific literature, clinical trials, clinical cases, speech transcriptions, and more as described in Table 2.

Please note that within DrBenchmark, we include classical tasks like NER and POS tagging, as well as more specific and challenging tasks like MCQA and multi-label classification. In Section 3.1, we provide an overview of the different French downstream tasks, while, in Section 3.2, we offer insights into the pipeline and its reproducibility.

[1] https://huggingface.co/DrBenchmark
[2] https://github.com/DrBenchmark/DrBenchmark

Dataset	Task	Metric	Train	Validation	Test	License
CAS	POS tagging	SeqEval F1	2,653	379	758	DUA
ESSAI	POS tagging	SeqEval F1	5,072	725	1,450	DUA
QUAERO	NER - EMEA	SeqEval F1	429	389	348	GFDL 1.3
	NER - MEDLINE	SeqEval F1	833	832	833	GFDL 1.3
E3C	NER - Clinical	SeqEval F1	969	140	293	CC BY-NC
	NER - Temporal	SeqEval F1	969	140	293	CC BY-NC
MorFITT	Multi-label Classification	Weighted F1	1514	1,022	1,088	CC BY-SA 4.0
FrenchMedMCQA	Question-Answering	Hamming / EMR	2,171	312	622	Apache 2.0
	Multi-class Classification	Weighted F1	2,171	312	622	Apache 2.0
Mantra-GSC	NER - EMEA	SeqEval F1	70	10	20	CC BY 4.0
	NER - Medline	SeqEval F1	70	10	20	CC BY 4.0
	NER - Patents	SeqEval F1	35	5	10	CC BY 4.0
CLISTER	Semantic Textual Similarity	EDRM / Spearman	499	101	400	DUA
DEFT-2020	Semantic Textual Similarity	EDRM / Spearman	498	102	410	DUA
	Multi-class Classification	Weighted F1	460	112	530	DUA
DEFT-2021	Multi-label Classification	Weighted F1	118	49	108	DUA
	NER	SeqEval F1	2,153	793	1,766	DUA
DiaMed	Multi-class Classification	Weighted F1	509	76	154	CC BY-SA 4.0
PxCorpus	NER	SeqEval F1	1,386	198	397	CC BY 4.0
	Multi-class Classification	Weighted F1	1,386	198	397	CC BY 4.0

Table 1: Descriptions and statistics of the 20 tasks included in DrBenchmark.

Dataset	Sources
CAS	Clinical cases
ESSAI	Clinical trial protocols
QUAERO	Drug leaflets & Biomedical titles
E3C	Clinical cases
MorFITT	Biomedical abstracts
FrenchMedMCQA	Pharmacy Exam
Mantra-GSC	Biomedical abstract / titles, drug labels, & patent
CLISTER	Clinical cases
DEFT-2020	Clinical cases, encyclopedia & drug labels
DEFT-2021	Clinical cases
DiaMed	Clinical cases
PxCorpus	Drug prescriptions transcripts

Table 2: Data sources covered by each datasets.

3.1. Downstream tasks

DEFT-2020 (Cardon et al., 2020) contains clinical cases, encyclopedia and drug labels introduced in the 2020 edition of an annual French Text Mining Challenge, called DEFT, and annotated for two tasks: (i) textual similarity and (ii) multi-class classification. The first task aims at identifying the degree of similarity within pairs of sentences, from 0 (the less similar) to 5 (the most similar). The second task consists in identifying, for a given sentence, the most similar sentence among three sentences provided.

DEFT-2021 (Grouin et al., 2021) is a subset of 275 clinical cases taken from the 2019 edition of DEFT. This dataset is manually annotated in two tasks: (i) multi-label classification and (ii) NER. The multi-label classification task focuses on identifying the patient's clinical profile based on the diseases, signs, or symptoms mentioned in the clinical cases. The dataset is annotated with 23 axes from Chapter C of the Medical Subject Headings (MeSH). The second task involves fine-grained information extraction for 13 types of entities (more detail in Appendix B.7).

E3C (Magnini et al., 2020) is a multilingual dataset of clinical cases annotated for the NER task. It consists of two types of annotations (more detail in Appendix B.4): (i) clinical entities (e.g., pathologies), (ii) temporal information and factuality (e.g., events). While the dataset covers 5 languages, only the French portion is retained for the benchmark. Since the dataset does not come with pre-defined subsets, we performed a 70 / 10 / 20 random split, as described in Table 3.

Subset	Train	Validation	Test
Clinical	87.38 % of layer 2	12.62 % of layer 2	100 % of layer 1
Temporal	70 % of layer 1	10 % of layer 1	20 % of layer 1

Table 3: Description of the sources for E3C.

The QUAERO French Medical Corpus (Névéol et al., 2014), simply referred to as QUAERO in this paper, contains annotated entities and concepts for NER tasks. The dataset covers two text genres (drug leaflets and biomedical titles), consisting of a total of 103,056 words sourced from EMEA or MEDLINE. 10 entity categories corresponding to the UMLS Semantic Groups (Lindberg et al., 1993) were annotated (more detail in Appendix B.3).. In total, 26,409 entity annotations were mapped to 5,797 unique UMLS concepts. Due to the presence of nested entities in annotations, we simplified the evaluation process by retaining only annotations at the higher granularity level from the BigBio (Fries et al., 2022) implementation, following the approach described in Touchent et al. (2023), which translates into an average loss of 6.06% of

the annotations on EMEA and 8.90% on MEDLINE. Additionally, considering that some documents from EMEA exceed the maximum input sequence length that most current language models can handle, we decided to split these documents into sentences.

MorFITT (Labrak et al., 2023) is a multi-label dataset annotated with medical specialties. It contains 3,624 biomedical abstracts from PMC Open Access. It has been annotated across 12 medical specialties (more detail in Appendix B.5), for a total of 5,116 annotations.

FrenchMedMCQA (Labrak et al., 2022) is a Multiple-Choice Question-Answering (MCQA) dataset for biomedical domain. It contains 3,105 questions coming from real exams of the French medical specialization diploma in pharmacy, integrating single and multiple answers. The first task consists of automatically identifying the set of correct answers among the 5 proposed for a given question. The second task consists of identifying the number of answers (between 1 and 5) supposedly correct for a given question.

Mantra-GSC (Kors et al., 2015) is a multilingual dataset annotated for biomedical NER. From the 5 languages covered, we included only the French subset in this benchmark. The dataset is obtained from 3 sources which have been partitioned to be evaluated separately by 2 annotation schemes (more detail in Appendix B.6): Medline (11 classes), and EMEA and Patents (10 classes). The sources cover different types of documents (biomedical abstracts/titles, drug labels and patents). To ensure evaluation consistency, we randomly split the dataset into 3 subsets: 70% for training, 10% for validation, and 20% for testing.

CLISTER (Hiebel et al., 2022) is a French clinical cases Semantic textual similarity (STS) dataset of 1,000 sentence pairs manually annotated by several annotators, who assigned similarity scores ranging from 0 to 5 to each pair. The scores were then averaged together to obtain a floating-point number representing the overall similarity. The objective of this dataset is to develop models that can automatically predict a similarity score that closely aligns with the reference score based solely on the two sentences provided.

CAS (Grabar et al., 2018) comprises 3,790 clinical cases that have been annotated for POS tagging with 31 classes using automatic annotations through Tagex [3], with an evaluation conducted by comparing the automatic outputs against manual annotations. This evaluation yielded 98% precision. Since the dataset does not come with predefined subsets, we made the decision to randomly split it into 3 subsets of 70%, 10% and 20% of the total data for training, validation and test respectively.

ESSAI (Dalloux et al., 2021) contains 7,247 clinical trial protocols annotated in 41 POS tags using TreeTagger (Schmid, 1994). As the dataset was not originally divided into 3 subsets, we applied the same procedure as on the CAS corpus.

PxCorpus (Kocabiyikoglu et al., 2022) is a spoken language understanding dataset in the domain of medical drug prescription transcripts. It includes 4 hours (1,981 recordings) of transcribed and annotated dialogues focused on drug prescriptions. The recordings were manually transcribed and semantically annotated. The first task involves classifying the textual utterances into one of the 4 intent classes (prescribe, replace, negate, none). The second task is a NER task where each word in a sequence is classified into one of 38 classes, such as drug, dose, or mode (more detail in Appendix B.9).

DiaMed is an original dataset created specifically for DrBenchmark. It comprises 739 new French clinical cases collected from an open source journal (The Pan African Medical Journal). The cases have been manually annotated by several annotators, one of which is a medical expert, into 22 chapters of the International Classification of Diseases, 10th Revision (ICD-10) (Wor, 2019). These chapters provide a general description of the type of injury or disease. To ease the annotation process, only label at the chapter level were used (more detail in Appendix B.8). The inter-annotator agreement between the 4 annotators has been computed for two annotation sessions (see Table 4), with 15 different clinical cases assessed per session.

	Session 1 - 0 to 15 docs		Session 2 - 15 to 30 docs	
Annotator ID	κ	$\mathcal{G}$	κ	$\mathcal{G}$
Annotator 1 & 2	0.538	0.566	0.697	0.705
Annotator 1 & 3	0.682	0.709	0.697	0.705
Annotator 1 & 4	0.397	0.429	0.548	0.558
Annotator 2 & 3	0.311	0.357	1.000	1.000
Annotator 2 & 4	0.472	0.497	0.672	0.707
Annotator 3 & 4	0.311	0.354	0.672	0.707
Average	0.452	0.485	0.714	0.730

Table 4: Inter-annotator agreement statistics. κ is referring to Kappa Cohen and $\mathcal{G}$ to Gwet's AC1.

3.2. Reproducibility and usage

To facilitate the adoption of DrBenchmark and ensure consistency in implementations, we have de-

[3] https://allgo.inria.fr/app/tagex

	Model	Tokenizer	Vocabulary	Pretraining	Corpus	Text Size
French Generalist	CamemBERTa	SentencePiece 32K	CCNET	from-scratch	CCNET	4 GB
	CamemBERT	SentencePiece 32K	OSCAR	from-scratch	OSCAR	138 GB
	FlauBERT	BPE 50K	Wiki + Web crawl	from-scratch	Wiki + Web crawl	71 GB
French Biomedical	DrBERT-FS	SentencePiece 32K	NACHOS	from-scratch	NACHOS	7.4 GB
	DrBERT-CP	WordPiece 30K	PubMed	continual pretraining	PubMed + NACHOS	21 + 4 GB
	CamemBERT-bio	SentencePiece 32K	OSCAR	continual pretraining	OSCAR + biomed-fr	138 + 2.7 GB
Cross-lingual Generalist	XLM-RoBERTa	WordPiece 30K	CC-100	from-scratch	CC-100	2.5 TB
English Biomedical	PubMedBERT	WordPiece 30K	PubMed	from-scratch	PubMed	21 GB

Table 5: Summary of the pre-training specifications for the different BERT-based models compared.

veloped a practical toolkit based on the HuggingFace Datasets library (Lhoest et al., 2021). This toolkit includes data loaders that adhere to normalized schemes and predefined data splits. It also provides pre-training and evaluation scripts for each of the tasks, utilizing the HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) libraries. For further guidance, we have integrated all the training details, including hyperparameters, in Appendix A. This information will help users to reproduce and customize the experiments conducted with DrBenchmark.

4. Experimental Protocol

In this section, we outline the experimental protocol used to compare the performance of existing language models within DrBenchmark. To guarantee fair comparison, we focus exclusively on pre-trained masked language models (MLMs) in this study. These MLMs are based on BERT-like architectures (Devlin et al., 2019).

We first provide a brief overview in Section 4.1 of the 8 pre-trained language models that were studied: French generalist models (CamemBERT, CamemBERTa and FlauBERT), cross-lingual generalist model (XLM-RoBERTa), French biomedical models (DrBERT and CamemBERT-bio), and English biomedical model (PubMedBERT). Subsequently, in Section 4.2, we describe the evaluation protocol employed to assess the performance of these models.

4.1. Pre-trained Masked Language Models

Table 5 summarizes the models and their parameters compared on DrBenchmark.

CamemBERT (Martin et al., 2020) is a RoBERTa based model for French, pre-trained from-scratch on the generalist French 138 GB subset of OSCAR corpus (Ortiz Suarez et al., 2019).

CamemBERTa (Antoun et al., 2023) is a DeBERTaV3 (He et al., 2023) based model pre-trained from-scratch on around 30% of the French subset of CCNET corpus (Wenzek et al., 2020) used for CamemBERT$_{CCNET}$, that had seen approximately 133 billion tokens during its pre-training.

FlauBERT (Le et al., 2020) is a BERT based model pre-trained from-scratch using a subsample of 71 GB of the French Common Crawl and Wikipedia corpora.

XLM-RoBERTa (Conneau et al., 2020) is a cross-lingual RoBERTa based model trained on 116 languages, including French, by using 2.5 TB of the CommonCrawl corpus.

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

DrBERT-FS and DrBERT-CP (Labrak et al., 2023) are French biomedical MLMs built using a from-scratch pre-training of RoBERTa (DrBERT-FS) and continual pre-training of PubMedBERT (DrBERT-CP) from the French public biomedical corpus NACHOS (Labrak et al., 2023) integrating 1.08 billion words (7.4 GB) and 646 million words (4 GB) respectively.

CamemBERT-bio (Touchent et al., 2023) is a French biomedical language model built using a continual pre-training of the CamemBERT$_{OSCAR-138GB}$ model. It was trained on the French public corpus biomed-fr (Touchent et al., 2023) with 413 million words (2.7 GB) and a wide range of data collected on the web.

4.2. Models evaluation

All the models are fine-tuned regarding a strict protocol using the same hyperparameters for each downstream task. The reported results are obtained by averaging the scores from four separate runs, thus ensuring robustness and reliability. We also report statistical significance computed using Student's t-test.

			French Generalist			French Biomedical			English Biomedical	Cross-lingual Generalist
Dataset	**Task**	**Baseline**	**CamemBERT**	**CamemBERTa**	**FlauBERT**	**DrBERT-FS**	**DrBERT-CP**	**CamemBERT-bio**	**PubMedBERT**	**XLM-RoBERTa**
CAS	POS	23.50	95.53**	96.56**	95.22**	**96.93**	96.46**	95.22**	94.82**	96.91
ESSAI	POS	26.31	97.38**	98.08**	97.05*	**98.41**	98.01**	97.39**	97.42**	98.34
QUAERO	NER EMEA	8.37	62.68**	64.86**	**74.86**	64.11**	67.05**	66.59**	53.19**	64.47**
	NER MEDLINE	4.92	55.25**	55.60**	48.98	55.82**	**60.10**	58.94	53.26**	51.12**
E3C	NER Clinical	4.47	54.70**	55.53	47.61	54.45	56.55	**56.96**	38.34	52.87**
	NER Temporal	21.74	**83.45**	83.22	61.64	81.48**	83.43	83.44	80.86**	82.6
MorFITT	Multi-Label CLS	3.24	64.21**	66.28**	70.25	68.70**	**70.99**	67.53**	68.58**	67.28**
FrenchMedMCQA	MCQA	21.83 / **11.57**	28.53 / 2.25**	29.77 / 2.57**	27.88 / 2.09**	31.07 / 3.22**	32.41 / 2.89**	**35.3** / 1.45	32.90 / 1.61**	34.74 / 2.09**
	CLS	8.37	66.21	64.44**	61.88	65.38	**66.22**	65.79	65.41*	64.69*
MantraGSC	NER FR EMEA	0.00	29.14**	40.84**	66.20	**66.23**	60.88	30.63**	40.14**	52.64*
	NER FR Medline	7.78	23.20**	22.55**	20.69	**42.38**	35.52	23.66**	27.53*	18.73*
	NER FR Patents	6.20	00.00**	44.16**	31.47**	**57.34**	39.68	00.00**	4.51**	8.58**
CLISTER	STS	0.44 / 0.00	0.55 / 0.33**	0.56 / 0.47**	0.50 / 0.29**	0.62 / 0.57**	0.60 / 0.49*	0.54 / 0.26**	**0.70** / **0.78**	0.49 / 0.23**
DEFT-2020	STS	0.49 / 0.00	0.59 / 0.58**	0.59 / 0.43**	0.58 / 0.51**	0.72 / 0.81*	0.73 / **0.86**	0.58 / 0.32**	**0.78** / **0.86**	0.60 / 0.26**
	CLS	14.00	96.31	**97.96**	42.37**	82.38	95.71*	94.78*	95.33*	67.66**
DEFT-2021	Multi-Label CLS	24.49	18.04**	18.04**	**39.21**	34.15**	30.04**	17.82**	25.53**	24.46**
	NER	0.00	62.76**	62.61**	33.51	60.44**	63.43*	**64.36**	60.27**	60.32**
DiaMED	CLS	15.36	30.40**	24.05**	34.08**	**60.45**	54.43**	39.57**	54.96**	26.69**
PxCorpus	NER	10.00	92.89**	95.05**	47.57	**95.88**	71.38	93.08**	94.66**	95.80
	CLS	84.78	94.41	93.95	93.45*	94.43	**94.52**	94.49	93.12	93.91

Table 6: Performance of the studied models over 4 runs. Best model in bold and second is underlined. Statistical significance is computed using Student's t-test: * stands for $p < 0.05$, ** stands for $p < 0.01$.

To ensure a fair and consistent comparison among systems for sequence-to-sequence tasks such as POS tagging and NER, we chose the SeqEval (Nakayama, 2018) metric in conjunction with the IOB2 format and the training of all the models to predict only the label on the first token of each word as mentioned by Touchent et al. (2023). It provides a tokenizer-agnostic evaluation and mitigates any correlation between models' performances and the tokenization process.

For STS tasks, the models' performance was assessed using two metrics: (1) the Spearman correlation, and (2) the mean relative solution distance accuracy (EDRM), as defined by the original authors of the DEFT-2020 dataset (Cardon et al., 2020).

5. Experiments and Results

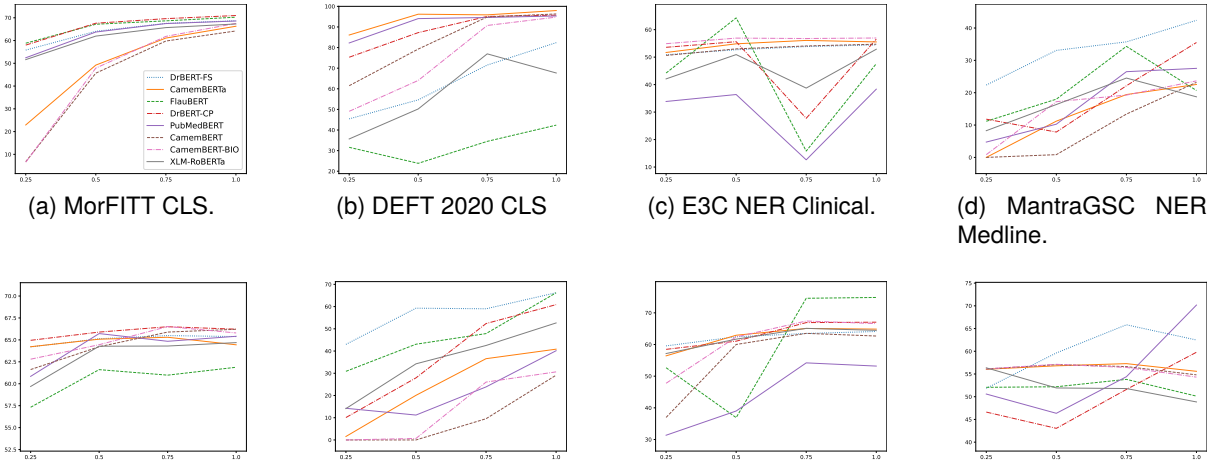
In Section 5.1, we compare the results obtained by each model within DrBenchmark, which permits to position a wide range of state-of-the-art models in the biomedical field across various NLP tasks. Then, we propose to gain a comprehensive understanding of the models' behavior by examining areas such as low-resource fine-tuning scenarios (Section 5.2) and the analysis of word tokenization of the studied models (Section 5.3).

5.1. Comparison of models performance

The results of the 8 models are reported in Table 6 and compared to a baseline obtained by considering the majority class for all predictions. Overall, although we might anticipate certain models to excel in all tasks, we discovered that no single model outperforms the rest in all application scenarios. Interestingly, most of the models examined manage to secure the top position in at least one of the French biomedical downstream tasks studied. The only exception pertains to the cross-lingual generalist model (`XLM-RoBERTa`), which manages to reach the second-best position on several tasks.

Despite this unexpected outcome, we observe that French biomedical language models (`DrBERT-FS`, `DrBERT-CP`, `CamemBERT-bio`), presumed to be the most aligned with the nature of the data of the benchmark, exhibit indeed superior performance across many tasks. More precisely, `DrBERT-FS` achieves the highest performance in 8 tasks, `DrBERT-CP` in 5 tasks, and `CamemBERT-bio` in 2 tasks. This indicates that domain and language-specialized models achieve the best performance in up to 75% of the DrBenchmark downstream tasks.

Biomedical vs. Generalist. The nature of the data appears to have an influence. Generalist models (`CamemBERT`, `CamemBERTa`, `FlauBERT` and `XLM-RoBERTa`) are more suitable for tasks that require extensive linguistic knowledge but may not perform as well as specialized models nor even reach their level of performance. We observe that all generalist models obtain better performance only on 4 out of the 20 tasks, but still remain competitive on most tasks. Furthermore, our experiments with `DrBERT-FS` indicate that biomedical models may require less pre-training data compared to generalist ones. However, it is important to note that this observation requires further confirmation. In some tasks, biomedical models that undergo continual pre-training from a generalist model, such as `CamemBERT-bio`, can prove to be the most effective, underscoring the value of pre-training on generalist



(a) MorFITT CLS. (b) DEFT 2020 CLS (c) E3C NER Clinical. (d) MantraGSC NER Medline.

(e) FrenchMedMCQA CLS (f) MantraGSC NER EMEA. (g) QUAERO NER EMEA. (h) CLISTER STS.

Figure 1: Performance with varying training subset sizes (25%, 50%, 75% and 100%). Results are reported on the full test set.

datasets.

From-scratch vs. Continual Pre-Training. `DrBERT-CP` and `CamemBERT-bio`, pre-trained from `PubMedBERT` and `CamemBERT` respectively, demonstrate improved performance compared to their initial models. Notably, `DrBERT-CP` outperforms `CamemBERT-bio` in 15 out of 20 tasks. These findings suggest that when it comes to continual pre-training, starting with a specialized model in the specific domain (here, `PubMedBERT`) may be a better choice than a generalist model (here, `CamemBERT`), even with different languages. Additionally, we observe that `DrBERT-FS` achieves the highest performance in 8 tasks, suggesting that starting *from-scratch* can be a competitive strategy compared to *continual pre-training*.

French vs. Other language. French models generally achieve better performance compared to English or multilingual ones. When considering the English `PubMedBERT` model, we observe that its performance in most tasks is comparable to that of the French models, with the exception of NER tasks where French models demonstrate superiority. Thus, we observe that the language appears to be less prominent when utilized in domain-specific tasks, such as those in the biomedical field.

RoBERTa vs. DeBERTaV3 architectures. Despite being trained on only 30% of the pre-training data used by $\texttt{CamemBERT}_{CCNET}$, `CamemBERTa` achieves identical or better performances in 68% of the tasks (12 out of 20), benefiting from the DeBERTaV3 architecture in domain-specific scenarios. However, all the models based on `CamemBERT` face difficulties in corpora with limited amount of data, such as MantraGSC Patents, where they fail to generate labels other than 'O'. On the other hand, in the same low-resource scenarios, `CamemBERTa` models exhibit greater robustness and achieve superior performance. The architecture on which the models are based therefore seems to play a role in the performance obtained.

5.2. Impact of fine-tuning with limited data

Unlike the process of training language models, the fine-tuning approach involves utilizing annotated data to adapt a pre-trained language model for solving specific downstream tasks. In the previous section, we observed that language models pre-trained on medical data generally achieved better performance on DrBenchmark compared to generalist models trained on much larger datasets. However, we now question the models' ability to be effectively applied to biomedical tasks when there is limited fine-tuning training data available. For this purpose, we conducted experiments by varying the amount of training data during the fine-tuning process by randomly choosing four percentages of the training data: 25%, 50%, 75% and 100%. To make the experiment as fair as possible, we did four runs for each percentage, model and dataset combination. The validation and test sets have not been changed for the sake of comparison.

We observe that on certain datasets, some models capture information more quickly than others, like in Figures 1b, 1f and 1a. Unsurprisingly, in almost all scenarios, having the complete training set yields better results than having only 25% of it. However, we note few exceptions in Figures 1a

		French Generalist			French Biomedical			English Biomedical	Cross-lingual Generalist
Dataset	**Task**	**CamemBERT**	**CamemBERTa**	**FlauBERT**	**DrBERT-FS**	**DrBERT-CP**	**CamemBERT-bio**	**PubMedBERT**	**XLM-RoBERTa**
CAS	POS	1.63	1.64	**1.34**	1.36	1.81	1.63	1.81	1.8
ESSAI	POS	1.55	1.56	**1.28**	1.29	1.78	1.55	1.78	1.75
QUAERO	NER EMEA	1.66	1.67	**1.37**	**1.37**	1.73	1.66	1.73	1.77
	NER Medline	2.01	2.01	**1.58**	1.64	1.97	2.01	1.97	2.18
E3C	NER FR Clinical	1.64	1.65	1.39	**1.32**	1.80	1.64	1.80	1.78
	NER FR Temporal	1.63	1.63	1.38	**1.31**	1.80	1.63	1.80	1.76
MorFITT	Multi-Label CLS	1.51	1.51	**1.33**	1.39	1.91	1.51	1.91	1.73
FrenchMedMCQA	MCQA	1.80	1.80	**1.55**	**1.55**	2.03	1.80	2.03	2.00
	CLS	1.80	1.80	**1.55**	**1.55**	2.03	1.80	2.03	2.00
MantraGSC	NER FR EMEA	1.50	1.46	**1.34**	1.37	1.99	1.50	1.99	1.71
	NER FR Medline	2.25	2.25	**1.88**	2.05	2.47	2.25	2.47	2.49
	NER FR Patents	1.58	1.58	**1.41**	1.51	2.06	1.58	2.06	1.86
CLISTER	STS	1.76	1.76	**1.55**	**1.55**	2.09	1.76	2.09	1.93
DEFT-2020	STS	1.43	1.43	**1.31**	1.45	1.92	1.43	1.92	1.64
	CLS	1.31	1.32	**1.20**	1.23	1.75	1.31	1.75	1.51
DEFT-2021	CLS	1.70	1.71	**1.48**	1.51	2.05	1.70	2.05	1.90
	NER	1.62	1.63	**1.35**	**1.35**	1.80	1.62	1.80	1.79
DiaMED	CLS	1.66	1.67	**1.45**	1.46	1.99	1.66	1.99	1.88
PxCorpus	NER	1.71	1.76	**1.63**	1.66	2.13	1.71	2.13	1.83
	CLS	1.71	1.76	**1.63**	1.66	2.13	1.71	2.13	1.83
Average		*1.67*	*1.67*	***1.43***	*1.47*	*1.90*	*1.67*	*1.90*	*1.85*

Table 7: Average sub-word units per word for each model and dataset. For each task, the lowest sub-word value is shown in bold, and the highest value is underlined. Models are grouped based on their tokenizer type. Cells in green indicate the best model in terms of performance for the task, while cells in red indicate the worst model.

and 1h with FlauBERT, where we observe the opposite trend. For intermediate percentages, 50% and 75%, we observe a decrease in performance with certain models, such as FlauBERT in Figures 1a and 1g, and DrBERT-CP in Figures 1d and 1h. In NER tasks (Figures 1a, 1d, 1f and 1g), DrBERT-FS achieves the best performance in scenarios with very little data, indicating good model robustness.

5.3. Analysis of word tokenization

Tokenizers play a crucial role in MLMs by utilizing size-limited vocabularies to split texts into sub-units, aiming to handle out-of-vocabulary (OOV) words. Due to variations in the training data, vocabularies differ across different models, as illustrated in Figure 2. As a result, tokenizers segment words in distinct ways, yet remarkably achieve similar performance levels as previously noted in Table 6.

So far, there has been a prevailing notion in the community that excessive segmentation of words in tokenization leads to a loss of morphological form and semantic meaning, introducing noise and adversely affecting performance (Church, 2020; Hofmann et al., 2021; Bostrom and Durrett, 2020). However, our experiments, as shown in Table 7, reveal that FlauBERT is the model with the least word segmentation (1.43 in average), while DrBERT-CP tends to have the highest average segmentation (1.90 in average). Surprisingly, when comparing the performance of these two models on the benchmark tasks, we observe that DrBERT-CP outperforms FlauBERT on 16 out of the 20 tasks, thus contradicting previous conclusions drawn by the

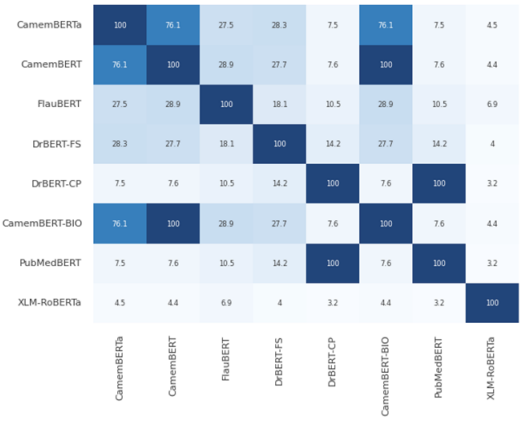


Figure 2: Vocabularies inter-coverage matrix.

community. Table 14 in Appendix C provides some examples of the tokenization done by each analyzed model, showcasing a list of commonly used biomedical terms. Yet, tokenization, as it is currently done in MLMs, seems to play a minor role in the performance of systems.

Table 8 summarizes the results obtained on average by the considered MLMs when aggregating the tasks into one of the five designated categories: POS, NER, MCQA, MCC (Multi-class classification), MLC (Multi-label classification), or STS tasks. Upon analyzing the average performance by task category, it becomes evident that the leading model, DrBERT-FS, does not excel in tasks such as MLC or STS. For example, the multilingual biomedical model PubMedBERT demonstrates a notable advantage, with nearly 18 EDRM points ahead of

CamemBERT-bio in the STS tasks.

	Tasks					
Models	**POS**	**NER**	**MCQA**	**MCC***	**MLC***	**STS**
CamemBERT	96.45	51.52	28.53 / 2.25	71.83	41.12	0.57 / 0.45
CamemBERTa	97.32	58.16	29.77 / 2.57	70.10	42.16	0.57 / 0.45
FlauBERT	96.13	51.85	27.88 / 2.09	57.94	**54.73**	0.54 / 0.40
DrBERT-FS	**97.67**	**64.23**	31.07 / **3.22**	75.66	51.42	0.67 / 0.69
DrBERT-CP	97.23	59.84	32.41 / 2.89	**77.72**	50.51	0.66 / 0.67
CamemBERT-bio	96.30	53.06	**35.30** / 1.45	73.65	42.67	0.56 / 0.29
PubMedBERT	96.12	46.93	32.90 / 1.61	77.20	47.05	**0.74** / **0.82**
XLM-RoBERTa	97.62	54.21	34.74 / 2.09	63.23	45.87	0.54 / 0.24

Table 8: Average results obtained by the different MLMs for each type of task. MLC stands for Multi-label classification and MCC for Multi-class classification.

6. Conclusion

In this paper, we introduced DrBenchmark, the first large language understanding benchmark tailored for the French biomedical domain. We conducted a qualitative evaluation of 8 state-of-the-art masked language models (MLMs) on this comprehensive benchmark, encompassing 20 diverse downstream tasks. Our findings illuminate the limitations of generalist models in tackling complex biomedical tasks, emphasizing the importance of employing domain-specific models to achieve peak performance. While the French biomedical models excel in most tasks, no single model emerges as universally superior. Remarkably, certain out-of-domain models or models trained in different languages exhibit superior performance in specific tasks and maintain competitiveness in others.

In conclusion, we have observed that several biomedical tasks in DrBenchmark exhibit relatively poor performance, even when utilizing specialized biomedical models. We postulate that the models examined in this study, here state-of-the-art MLMs, may not be the most effective choices for specific tasks such as question-answering or multi-label classification. In our future research, we intend to shift our focus towards generative approaches, such as LLaMA (Touvron et al., 2023), OPT (Zhang et al., 2022b), or GPT-NeoX-20B (Black et al., 2022), as well as their instruction-tuned counterparts (Iyer et al., 2023). These alternatives may offer more suitable solutions for addressing these types of tasks.

7. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715) and from CCIPL (Centre de Calcul Intensif des Pays de la Loire). This work was financially supported by ANR MALADES (ANR-23-IAS1-0005), ANR AIBy4 (ANR-20-THIA-0011) and Zenidoc.

Ethical considerations

All code for DrBenchmark is released under the MIT License. The licensing for all datasets remains unchanged from the original sources, and DrBenchmark has no intention of redistributing these datasets.

Limitations

The quantitative study we conducted on the PLMs requires further in-depth analysis to comprehend the impact of different parameters. Firstly, we investigated the influence of tokenizers based on the average number of sub-tokens they produce per word. It is important to note that various tokenization algorithms are employed, depending on the model under examination. Therefore, it is necessary to specifically assess the impact of these algorithms on model construction. Additionally, the size of the data has not been thoroughly investigated, particularly the significance of the pre-training data size, especially specialized data in the biomedical field. Analyzing the influence and importance of the amount of data used would be crucial for gaining deeper insights.

Although the benchmark is easily reproducible and customizable, it required a substantial amount of computational power to execute all runs. We utilized approximately 2,500 hours on V100 GPUs from the Jean-Zay supercomputer to complete the quantitative study. According to the Jean Zay supercomputer documentation [4], the total environmental cost is estimated to be equivalent to 647,500 Wh or 36.9 kgCO2eq/kWh, based on the carbon intensity of the energy grid mentioned in the BLOOM environmental cost study conducted on Jean Zay (Luccioni et al., 2022). While we acknowledge the significant cost of our study, we believe it will enable the research community to direct their future studies more efficiently by providing a comprehensive overview of the performance and behavior of existing models. This will help prevent redundant evaluations of the same models.

8. Bibliographical References

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023. Data-Efficient French Language Modeling with CamemBERTa. In *Findings of the 61th*

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

Annual Meeting of the Association for Computational Linguistics (ACL'23), Toronto, Canada.

Aman Berhe, Guillaume Draznieks, Vincent Martenot, Valentin Masdeu, Lucas Davy, and Jean-Daniel Zucker. 2023. AliBERT: A pre-trained language model for French biomedical text. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 223–236, Toronto, Canada. Association for Computational Linguistics.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers.

Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized French language models for biomedical named entity recognition. 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 36–48, Nancy, France. ATALA et AFCP.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

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*, pages 2479–2490, Marseille, France. European Language Resources Association.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Rian Touchent, Laurent Romary, and Éric Villemonte de La Clergerie. 2023. Camembert-bio: a tasty french language model better for your health.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022a. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models.

9. Language Resource References

2019. International statistical classification of diseases and related health problems 10th revision. World Health Organization.

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.

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

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. Bigbio: A framework for data-centric biomedical natural language processing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Cyril Grouin, Natalia Grabar, and Gabriel Illouz. 2021. Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021 (clinical cases classification and automatic evaluation of student answers : Presentation of the DEFT 2021 challenge). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 1–13, Lille, France. ATALA.

Nicolas Hiebel, Olivier Ferret, Karën Fort, and Aurélie Névéol. 2022. CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4306–4315, Marseille, France. European Language Resources Association.

Alican Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *13th Language Resources and Evaluation Conference (LREC 2022)*.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-Antoine Gourraud, Emmanuel

Morin, and Mickael Rouvier. 2022. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abou Dhabi, United Arab Emirates.

Yanis Labrak, Mickaël Rouvier, and Richard Dufour. 2023. MORFITT : A multi-label corpus of French scientific articles in the biomedical domain. In *30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) Atelier sur l'Analyse et la Recherche de Textes Scientifiques*, Paris, France. Florian Boudin.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The Unified Medical Language System. *Methods Inf Med*, 32(4):281–291.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

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. In *Proc of BioTextMining Work*, pages 24–30.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Appendices

A. Hyperparameters

For the experiments, we utilize the following hyperparameters that yield optimal performance from the models. To mitigate overfitting, we locally save the best model based on its validation metric.

Hyper-parameter	Value
Max sequence length	512
Epochs	20
Batch size	16
Learning Rate	2e-5
Weight Decay	0.01

Table 9: Hyper-parameters for the question-answering experiments.

Hyper-parameter	Value
Max sequence length	512
Epochs	10 / 25 / 35
Batch size	16
Learning Rate	2e-5
Weight Decay	0.01

Table 10: Hyper-parameters for the classification experiments. The number of epochs is by default 10 except for DEFT-2020 (25 epochs) and MorFITT (35 epochs).

Hyper-parameter	Value
Max sequence length	512
Epochs	10
Batch size	16
Learning Rate	1e-5
Weight Decay	0.01

Table 11: Hyper-parameters for the POS tagging experiments.

Hyper-parameter	Value
Max sequence length	512
Epochs	30
Batch size	16
Learning Rate	2e-5
Weight Decay	0.01

Table 12: Hyper-parameters for the regression experiments.

Hyper-parameter	Value
Max sequence length	512
Epochs	15
Batch size	16
Learning Rate	1e-4
Weight Decay	0.01

Table 13: Hyper-parameters for the NER experiments.

B. Dataset Classes

B.1. CAS

INT, *PRO:DEM*, *VER:impf*, *VER:ppre*, *PRP:det*, *KON*, *VER:pper*, *PRP*, *PRO:IND*, *VER:simp*, *VER:con*, *SENT*, *VER:futu*, *PRO:PER*, *VER:infi*, *ADJ*, *NAM*, *NUM*, *PUN:cit*, *PRO:REL*, *VER:subi*, *ABR*, *NOM*, *VER:pres*, *DET:ART*, *VER:cond*, *VER:subp*, *DET:POS*, *ADV*, *SYM* and *PUN*.

B.2. ESSAI

INT, *PRO:POS*, *PRP*, *SENT*, *PRO*, *ABR*, *VER:pres*, *KON*, *SYM*, *DET:POS*, *VER:*, *PRO:IND*, *NAM*, *ADV*, *PRO:DEM*, *NN*, *PRO:PER*, *VER:pper*, *VER:ppre*, *PUN*, *VER:simp*, *PREF*, *NUM*, *VER:futu*, *NOM*, *VER:impf*, *VER:subp*, *VER:infi*, *DET:ART*, *PUN:cit*, *ADJ*, *PRP:det*, *PRO:REL*, *VER:cond* and *VER:subi*.

B.3. QUAERO

O, *GEOG*, *PHEN*, *DISO*, *ANAT*, *OBJC*, *PHYS*, *PROC*, *DEVI*, *CHEM* and *LIVB*

B.4. E3C

Clinical: *O*, and *CLINENTITY*

Temporal: *O*, *EVENT*, *ACTOR*, *BODYPART*, *TIMEX3* and *RML*

B.5. MorFITT

microbiology, *etiology*, *virology*, *physiology*, *immunology*, *parasitology*, *genetics*, *chemistry*, *veterinary*, *surgery*, *pharmacology* and *psychology*

B.6. MantraGSC

Medline: *ANAT*, *PROC*, *CHEM*, *PHYS*, *GEOG*, *DEVI*, *LIVB*, *OBJC*, *DISO*, *PHEN* and *O*.

EMEA and Patents: *ANAT*, *PROC*, *CHEM*, *PHYS*, *DEVI*, *LIVB*, *OBJC*, *DISO*, *PHEN* and *O*.

B.7. DEFT-2021

Multi-label Classification: *immunitaire (immunology)*, *endocriniennes (endocrinology)*, *blessures (injury)*, *chimiques (chemicals)*, *etatsosy (signs and symptoms)*, *nutritionnelles (nutrition)*, *infections (infections)*, *virales (virology)*, *parasitaires (parasitology)*, *tumeur (oncology)*, *osteomusculaires (osteomuscular disorders)*, *stomatognathique (stomatology)*, *digestif (digestive system disorders)*, *respiratoire (respiratory system disorders)*, *ORL (otorhinolaryngologic diseases)*, *nerveux (nervous system disorders)*, *oeil (eye diseases)*, *homme (male genital diseases)*, *femme (female genital diseases)*, *cardiovasculaires (cardiology)*, *hemopathies (hemic and lymphatic diseases)*, *genetique (genertic disorders)* and *peau (dermatology)*.

Named-entity recognition: *O*, *ANATOMY*, *DATE*, *DOSAGE*, *DURATION*, *MEDICAL EXAM*, *FREQUENCY*, *MODE*, *MOMENT*, *PATHOLOGY*, *SOSY*, *SUBSTANCE*, *TREATMENT* and *VALUE*

B.8. DiaMed

- *A00-B99 Certain infectious and parasitic diseases*
- *C00-D49 Neoplasms*
- *D50-D89 Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism*
- *E00-E89 Endocrine, nutritional and metabolic diseases*
- *F01-F99 Mental, Behavioral and Neurodevelopmental disorders*
- *G00-G99 Diseases of the nervous system*

- *H00-H59 Diseases of the eye and adnexa*
- *H60-H95 Diseases of the ear and mastoid process*
- *I00-I99 Diseases of the circulatory system*
- *J00-J99 Diseases of the respiratory system*
- *K00-K95 Diseases of the digestive system*
- *L00-L99 Diseases of the skin and subcutaneous tissue*
- *M00-M99 Diseases of the musculoskeletal system and connective tissue*
- *N00-N99 Diseases of the genitourinary system*
- *O00-O9A Pregnancy, childbirth and the puerperium*
- *P00-P96 Certain conditions originating in the perinatal period*
- *Q00-Q99 Congenital malformations, deformations and chromosomal abnormalities*
- *R00-R99 Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified*
- *S00-T88 Injury, poisoning and certain other consequences of external causes*
- *U00-U85 Codes for special purposes*
- *V00-Y99 External causes of morbidity*
- *Z00-Z99 Factors influencing health status and contact with health services*

B.9. PxCorpus

Intent classification: *MEDICAL PRESCRIPTION*, *NEGATE*, *NONE* and *REPLACE*

Named-entity recognition: *O*, *A*, *CMA_EVENT*, *D_DOS_FORM*, *D_DOS_FORM_EXT*, *D_DOS_UP*, *D_DOS_VAL*, *DOS_COND*, *DOS_UF*, *DOS_VAL*, *DRUG*, *DUR_UT*, *DUR_VAL*, *FASTING*, *FREQ_DAYS*, *FREQ_INT_V1*, *FREQ_INT_V1_UT*, *FREQ_INT_V2*, *FREQ_INT_V2_UT*, *FREQ_STARTDAY*, *FREQ_UT*, *FREQ_VAL*, *INN*, *MAX_UNIT_UF*, *MAX_UNIT_UT*, *MAX_UNIT_VAL*, *MIN_GAP_UT*, *MIN_GAP_VAL*, *QSP_UT*, *QSP_VAL*, *RE_UT*, *RE_VAL*, *RHYTHM_HOUR*, *RHYTHM_PERDAY*, *RHYTHM_REC_UT*, *RHYTHM_REC_VAL*, *RHYTHM_TDTE* and *ROA*

C. Word tokenization

	French Generalist			French Biomedical	English Biomedical	Cross-lingual Generalist
Term	**CamemBERTa**	**CamemBERT CamemBERT-bio**	**FlauBERT**	**DrBERT-FS**	**PubMedBERT DrBERT-CP**	**XLM-RoBERTa**
asymptomatique	a-s-ym-pto-matique	a-s-y-mp-to-matique	as-ym-ptom-atique	✓	asympt-omat-ique	as-y-mp-tomat-ique
blépharorraphie	blé-phar-or-ra-phi-e	blé-phar-or-ra-phi-e	bl-é-phar-or-raph-ie	blé-ph-ar-or-ra-ph-ie	ble-pha-ror-ra-phi-e	b-lép-har-orra-phi-e
bradycardie	brad-y-cardi-e	brad-y-cardi-e	bra-dy-car-die	✓	brady-car-di-e	bra-dy-card-ie
bronchographie	bronch-ographie	bron-ch-ographie	bron-cho-graphie	bronch-ographie	bronch-ograph-ie	bron-ch-ographie
bronchopneumopathie	bronch-op-ne-um-opathie	bron-cho-p-ne-um-opathie	bron-chop-neu-mo-pathie	bronchop-neumopathie	bronch-op-neum-opath-ie	bron-chop-ne-umo-pathi-e
dysménorrhée	dys-mén-or-r-h-ée	dys-mén-or-r-h-ée	dys-mé-nor-rh-ée	dys-m-énorrhée	dysm-eno-rr-he-e	dys-mén-or-r-hé-e
glaucome	gla-uc-ome	gla-uc-ome	glau-come	✓	glauc-ome	gla-u-come
IRM	✓	✓	✓	✓	ir-m	I-RM
kystectomie	k-yst-ectomie	ky-st-ectomie	ky-st-ec-tomie	kys-tectomie	ky-st-ectom-ie	ky-st-ecto-mie
neuroleptique	neuro-le-p-tique	neuro-le-p-tique	neur-ol-ep-tique	neur-oleptique	neurol-ept-ique	neuro-lep-tique
nicotine	✓	✓	✓	✓	✓	nico-tine
poliomyélite	poli-om-y-élite	poli-om-y-élite	poli-omy-élite	poli-omyélite	poli-omyel-ite	poli-om-y-é-lite
rhinopharyngite	rh-ino-phar-y-ng-ite	rhin-oph-ary-ng-ite	rh-ino-phar-yn-gite	rhin-opharyng-ite	rhin-oph-aryng-ite	r-hin-op-har-y-ng-ite
toxicomanie	toxico-mani-e	toxico-mani-e	✓	✓	toxic-oman-ie	toxic-om-anie
vasoconstricteur	vas-oc-on-strict-eur	vas-oc-on-strict-eur	vas-o-cons-tri-cteur	vasoconstric-teur	vasoconstric-te-ur	vaso-con-strict-eur

Table 14: Visual comparison of models' tokenization on commonly used biomedical terms. A checkmark indicates that the word is present as a complete token, while hyphens separate subword units.