Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

An evaluation of current benchmarking strategies for French biomedical language models

Felix Herron[1,2]
(1) Laboratoire d'Informatique Grenoble, 700 Av. Centrale, 38401 Saint-Martin-d'Hères, France
(2) Laboratoire d'Analyse et de Modélisation de Systèmes d'Aide à la Décision, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
felix.herron@univ-grenoble-alpes.fr

ABSTRACT

We describe the current state of benchmarking for French language biomedical natural language processing (NLP). We note two important criteria in biomedical benchmarking: first, that a biomedical benchmark clearly simulate a specific use cases, in order to offer a useful evaluation of a biomedical model's real life applicability. Second: that a biomedical benchmark be created in collaboration with biomedical professionals. We note that many biomedical benchmarks, particularly in French, do not adhere to these criteria; however, we highlight other biomedical benchmarks which adhere better to those criteria. Furthermore, we evaluate some of the most common French biomedical benchmarks on an array of models and empirically support the necessity of domain-specific and language-specific pre-training for natural language understanding (NLU) tasks. We show that some popular French biomedical language models perform poorly and/or inconsistently on important biomedical tasks. Finally, we advocate for an increase in publicly available, clinically targeted French biomedical NLU benchmarks.

RÉSUMÉ

Évaluation de benchmarking actuel pour des modèles de langage biomédicaux français

Nous présentons dans cet article une réflexion à propos des tâches d'évaluation en traitement automatique des langues (TAL) biomédical et clinique pour la langue française. Nous soulignons l'insuffisance de référentiels reflétant des scénarios d'utilisation réels et concrets, limitant ainsi la pertinence de leurs résultats pour les professionnels de santé. De plus, il est réputé que certains sont élaborés sans la participation active de spécialistes du domaine. Notre examen d'une sélection de référentiels biomédicaux français classiques soutient le besoin d'un préentraînement spécifique au domaine biomédical en français destiné plus particulièrement aux tâches de compréhension du langage naturel (NLU). Nous montrons également que certains modèles préentrainés pour les domaines biomédicaux français affichent des performances médiocres voire incohérentes lorsqu'ils sont testés sur des tâches biomédicales courantes dans la littérature biomédicale française. En conclusion, nous plaidons pour une augmentation des référentiels librement disponibles et focalisés sur des situations cliniques réelles.

KEYWORDS : Benchmarking, biomedical language modeling, deep learning.

MOTS-CLÉS: Benchmarking, modélisation de langage biomédicale, apprentissage profond.

Actes de la 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, pages 1–16.

1 Introduction

Since the advent of the Transformer architecture and the subsequent rise of deep language models (LMs), the power of natural language processing (NLP) models has significantly improved (Vaswani *et al.*, 2017) (Devlin *et al.*, 2019). This improvement has led to the application of LMs in various domains, such as grading at universities (Fuchs, 2023), policing the internet for hate speech (Plaza-del Arco *et al.*, 2021), or helping doctors treat their patients (Agarwal *et al.*, 2018). However, with great power comes great responsibility; as these models become increasingly ubiquitous and their decisions increasingly relied upon, potential deployers must have a nuanced understanding of their abilities. One must know as precisely as possible how well an LM will perform on its assigned task in order to gauge the expected error in its calculations, and thus afford it adequate human supervision. A model ought not be deployed until it has been properly and thoroughly evaluated.

To perform this evaluation, the scientific community relies on benchmarks, which are series of tests designed to simulate real-life scenarios which an LM might encounter. In order for a new model to gain traction in the scientific community, it must perform well on certain benchmarks. For well established domains, these benchmarks have been studied for decades and have undergone multitudinous permutations and updates. At any given moment there are certain benchmarks that are understood by the community to be essential; a model not evaluated on these will not be taken seriously by the community, or reviewers at conferences or journals (Dehghani *et al.*, 2021)[1]. As the state-of-the-art (SOTA) improves, these benchmarks are continually updated or retired due to "degeneration", where human parity is reached (Dehghani *et al.*, 2021; Bowman & Dahl, 2021). However, domains in which there are not yet well-established benchmarks, such as French biomedical NLP, lack such self-regulation (Dehghani *et al.*, 2021). Therefore, the publishers of models in cutting edge domains must choose, without relying on significant precedent, on which benchmarks to evaluate their models. This freedom of choice can lead authors to primarily include benchmarks on which their models perform well compared with their competitors, a process referred to by Dehghani *et al.* (2021) as "rigging the lottery". This is counterproductive for a nascent domain, as it can motivate the reverse-engineering of evaluation systems to promote individual models, rather than the engineering of better models to solve known tasks.

In this paper, we show that French biomedical NLP benchmarking exhibits weaknesses consistent with an early stage domain as taxonimized by Dehghani *et al.* (2021). We consider challenges inherent in biomedical benchmark creation, and discuss ways in which benchmarks can be created more effectively. We then perform a review of benchmarks used in French biomedical NLP, and perform an independent evaluation of them using SOTA models. We show that more work towards benchmarking is necessary in order to better prepare French biomedical LMs for deployment.

2 Benchmarking biomedical LMs

2.1 Motivation

In machine learning, a series of tests on which a model can be evaluated. The purpose of a benchmark is to measure the quality of different models on identical input, both to rank the models amongst each

[1]For example, all of the English language masked language models (MLMs) published during the NLP boom promulgated by the release of the Transformer architecture, such as BERT, XLNet, RoBERTa, XLM-RoBERTa, were evaluated on GLUE and SQuAD (Wang *et al.*, 2018; Rajpurkar *et al.*, 2016)

other and to determine tractability of a problem using SOTA technology. Furthermore, a benchmark should mirror real life applications as closely as possible: the purpose for training and publishing biomedical LMs is for their eventual deployment to assist in some manner in the treatment of medical patients. Hence, when creating a biomedical benchmark, we should consider what real use cases exist for biomedical LMs. For example, Kanwal & Rizzo (2022) describe the task of summarizing dense clinical notes, Rabhi (2022) describes predicting patient outcomes based on previous visits in a multi-modal setting, and Carchiolo *et al.* (2019) the (semi)-automated prescription of medicines. Furthermore, Yang *et al.* (2023) identify three main phases of a patient's journey in which LMs could be applied.

1. Prior to formal medical care: screening without the input of human professionals, screening for potential medical conditions.
2. During medical care: diagnosing conditions based on written reports.
3. Post medical care: counseling patients, assisting in insurance billing.

In general, most perceived medical LM use cases involve automating a task that requires a nuanced understanding of medicine in general, and any individual patient likewise. Thus, we posit that most tasks envisioned for biomedical NLP fall under the umbrella of natural language understanding (NLU), which means a model's ability to parse texts semantically rather than merely syntactically[2]. Another important category of encoder LM benchmarks is named entity recognition (NER), which involves classifying individual words and phrases. In the biomedical domain, this could be useful for the extraction of keywords from long-form medical texts, and for text summarization based thereupon. However, according to the aforementioned clinical use cases for biomedical LMs, NER is in general of lesser significance than NLU tasks. Biomedical benchmarks in practice should reflect this proclivity towards NLU; however, in the following section we will discuss the challenges of creating biomedical NLU benchmarks.

2.2 Difficulties in biomedical NLP benchmarking

In order to create a biomedical benchmark, one must first have a biomedical corpus on which to build tasks. To the detriment of NLP scientists, access to and publication of medical data in general is heavily regulated in order to safeguard individuals' privacy ((European Parliament and Council, 2016; United States Department of Health and Human Services, 2013; Li & Qin, 2017)). In order to distribute data, patients' Protected Health Information (PHI) must be hidden from Electronic Health Records (EHR); however, PHI cannot simply be erased, as it is a critical piece of information in biomedical text analysis (Mamede *et al.*, 2016). Different anonymization standards and techniques exist for the automatic de-identifying of EHRs in order to facilitate data sharing, though there exists no industry gold standard (Sweeney, 2002; Machanavajjhala *et al.*, 2007; Li & Qin, 2017). For example, the most-frequently utilized English EHR corpus, MIMIC-III, uses a combination of regular expressions and dictionary lookups (Johnson *et al.*, 2016), though this system is continually updated and not guaranteed to completely de-identify all data. The difficulties of de-identifying data are exemplified in the `DrBERT` paper, which trains and evaluates a slew of models on private datasets,

[2]This is one substantial difference from traditional corpus linguistic use cases: in practical medical NLP, semantic understanding far outweighs syntactic precision. Classical general purpose LM use cases, such as grammar or spell checking, are superfluous for encoder biomedical LMs.

but these data remained siloed - i.e. private, accessible only to those with insider permissions (Labrak *et al.*, 2023; Lin *et al.*, 2022).

One technique to circumvent this problem is known as Federated Learning (FL), in which models are passed between secure data silos for on-site learning (Zhang *et al.*, 2021) as well as evaluation (Karargyris *et al.*, 2023). This way, no data must be transferred between institutions. Indeed, FL is gaining traction in many fields, including the biomedical one, as a means to avoid data leakage (Rieke *et al.*, 2020). Unfortunately, studies have shown that some models can be attacked to reveal training data, which defeats the purpose of privacy gains in private training in FL (Winograd, 2023). Furthermore, lack of data transparency further exacerbates the opacity inherent in highly parameterized LMs. FL is also expensive, as it requires a high degree of organizational cooperation, from thorough data inspection to functional model exchange platforms. While we advocate for this method in principle, its cost, both financially and administratively, renders its implementation challenging.

Another issue afflicting biomedical benchmarks is that they are often created by NLP scientists without significant input from biomedical professionals (Cardon *et al.*, 2020; Peng *et al.*, 2019; Carrino *et al.*, 2022). One solution to this problem is to collaborate directly with domain-specific experts. This is achieved in the Chinese and Russian biomedical benchmarks `CBLUE` and `RuMedBench` by working together with doctors (Zhang *et al.*, 2022; Blinov *et al.*, 2022). However, this collaboration can be challenging for any number of reasons, from pecuniary to bureaucratic to temporal. These challenges are particularly dire in the biomedical domain where, due to patient privacy concerns, there is an unusual abundance of administrative hurdles to clear in order to access, let alone share or publish data for potential benchmark usage. Thus, some benchmarks are created using sub-optimal corpora without the input of domain-specific experts, which can lead to self-professed ambiguity in quality of the resulting created benchmarks (Cardon *et al.*, 2020), further compounding the bias inherent in any human-based annotation (Schoch *et al.*, 2020). This results in benchmarks which are either insufficiently similar to real-life tasks, or potentially inaccurate. As noted in Cardon *et al.* (2020), where several common French biomedical benchmarks[3] were introduced: "...**the annotators' lack of medical training could diminish the annotation quality**" [4].

2.3 Evaluation of existing biomedical benchmarks

We compare the types of tasks in common biomedical NLP benchmarks (see Table 1). According to the two major criticisms interrogated in this paper (insufficient focus on NLU, non-medical annotators), some benchmarks are of higher quality than others. The Russian `RuMedBench`, for example, uses "clinician" annotators for each of their tasks, and focuses specifically on NLU tasks, introducing each with an explicit allusion to a clinical use case. For example, its `RuMedSymptomRec` symptom recommendation task helps users refine their (online) medical searches based on incomplete symptom lists. The Chinese `CBLUE` benchmark also highlights the medical credentials of its annotators ("doctors from class A tertiary hospitals"), and likewise is thorough in its motivation for each task. For example, in its `KUAKE-QIC` task, a biomedical LM must classify medically related search engine queries by category, such as diagnosis, treatment plan, or test result analysis. The English `BLURB` contains several tasks which were annotated by medical professionals. Like `CBLUE`, `BLURB` emphasizes the need for eclectically sourced corpora and a variety of different subtypes of tasks,

[3] `CAS-POS`, `CAS-SG`, and a semantic similarity task similar to `CLISTER` - see Section 2.4 for further details

[4] *Fr: l'absence de formation médicale des annotateurs peut également présenter un obstacle dans la qualité du travail d'annotation*

mainly of type NLU[5].

However, we find that not all biomedical benchmarks are as thorough as RuMedBench, CBLUE, and BLURB. For example, despite the greater importance of NLU tasks in biomedical NLP, both Bio-cli and CamemBERT-bio are evaluated on only NER tasks, as illustrated in Table 1. Furthermore, both jargon and DrBERT include part of speech (POS) tagging tasks as part of their principal analyses, despite little evidence for this being a useful clinical benchmark. In DrBERT, there is one particularly clinically relevant NLU task, aHF - the diagnosis of a heart condition based on a freeform text about a patient - but it is private, making it impractical for adoption by the community.

Despite compiling many tasks, of which some are NLU, neither Segonne *et al.* (2024) nor Labrak *et al.* (2023) discussed the clinical relevance that each task was trying to simulate, instead describing each task from a more technical NLP perspective. For example, they use the NLU task FrenchMedMCQA, which involves answering multiple choice questions from a real French pharmaceutical exam; the applicability of its results to a concrete use case are not immediately evident. However, its content was created by biomedical professionals, and thus the labels are as high quality as possible. To the contrary, CLISTER is a task based on judging the semantic similarity of pairs of sentences on a scale from 0 to 5. The clinical application of this is more immediately evident - for example, pairs of appointment summaries could be compared to determine whether a patient's health is changing. However, the four annotators of CLISTER were also the paper's four authors, none of whom has a background in medicine. Although they lay out a detailed annotation pipeline to ensure inter-annotator agreement, which emphasized "semantic similarity [of] medical concepts" (Hiebel *et al.*, 2022), given their lack of medical background, it appears they may be agreeing on potential shared medical misunderstanding. For example, consider this sample pair from the CLISTER corpus (similarity score 2.5):

Le reste de la vessie est strictement normal. *En.: The rest of the bladder was strictly normal*
Le reste du parenchyme rénal était normal. *En.: The rest of the renal parenchyma was normal*

To correctly annotate this pair, one must know what a renal parenchyma is (the author of this paper did not know what that was), as well as understand whether its normalcy is equivalent to bladder normalcy.

Table 1: Comparison of NLU focus for common biomedical benchmarks

Benchmark	DrBERT	CamemBERT-bio	jargon	BLURB	Bio-cli	CBLUE	RuMedBench
Citation	(Labrak *et al.*, 2023)	(Touchent *et al.*, 2023)	(Segonne *et al.*, 2024)	(Gu *et al.*, 2020)	(Carrino *et al.*, 2022)	(Zhang *et al.*, 2022)	(Blinov *et al.*, 2022)
Language	French	French	French	English	Spanish	Chinese	Russian
Grammar tasks	2	0	3	0	0	0	0
NER tasks	5[6]	5	4	6	3	2	1
NLU tasks	4[7]	0	3	7[8]	0	6	4

2.4 Benchmarks in our study

We will use a representative sample of six popular French biomedical benchmarks, as described in Table 2, on which we will evaluate several French biomedical LMs. Of the three publicly available

[5]Regarding English biomedical benchmarking: the ClinicalBERT paper uses clinic readmission from MIMIC-III longform clinical notes as a benchmark (Huang *et al.*, 2019) (Johnson *et al.*, 2016). This is a highly targeted use case! However, this benchmark has inexplicably not been reused in subsequent English biomedical literature.

[6]Of which two are private

[7]Of which two are private and one inaccessible

[8]Of which three are relation extraction tasks

French NLU benchmarks available, we chose two (CLISTER and FrenchMedMCQA), while leaving out the semantic similarity task from Cardon *et al.* (2020), given that CLISTER is basically its updated equivalent (Hiebel *et al.*, 2022). We are not aware of any other publicly available French biomedical NLU tasks at the time of writing.

Table 2: Statistics for each dataset included in this paper

	CAS-POS	ESSAI-POS	CAS-SG	QUAERO-MEDLINE	CLISTER	FrenchMedMCQA
Task	POS	POS	NER	NER	Semantic Similarity	Question Answering
Size (sentences)	3.8k	2.4k	4.5k	7.2k	1k	3.1k
DrBERT / CamemBERT-bio / jargon	✓✓✓	✓✗✓	✓✓✓	✓✗✓	✗✗✓	✓✗✓
Is task NLU?	✗	✗	✗	✗	✓	✓
Clinician annotated?	✗	✗	✗	✗	✗	✓

3 French biomedical LMs

In 2024, analysis of texts is accomplished using Masked Language Models (MLMs) based on the Transformer architecture (Devlin *et al.*, 2019). These models use fixed-length self-attention to process blocks of text and emit an encoded embedding for each sub-word of the input. MLMs are convenient because their pre-training is completely unsupervised, meaning it requires no labeled data. Given an input document composed of many tokens (syntactically selected sub-words), an MLM produces embeddings for each token, as well as a summarizing embedding which seeks to represent the document as a single unit. They can therefore be used for two main types of analysis: token-level analysis (using each token embedding) or document-level analysis (using the summarizing embedding). These embeddings can either be used out of the box or further refined by using end-to-end fine-tuning to create task-specific representations (Devlin *et al.*, 2019).

In this paper, we will examine three classes of French biomedical LMs. Each has in the order of 100M trainable parameters.

1. French bio-medical models (left portion of Table 3) - i.e. those pre-trained from scratch (or from general-purpose checkpoint) on French bio-medical corpora. We will test DrBERT-4 (Labrak *et al.*, 2023), CamemBERT-bio (Touchent *et al.*, 2023), Jargon-biomed and Jargon-gen-biomed (Segonne *et al.*, 2024). (The last was trained on a mixture of biomedical data and general data.)

2. General purpose French language models (middle portion of Table 3) - we seek to replicate the aforementioned necessity of domain-specific models for various biomedical tasks. We will be using CamemBERT (Martin *et al.*, 2020) trained on the CCNet corpus (Wenzek *et al.*, 2020), as well as FlauBERT-1 (Le *et al.*, 2020).

3. English biomedical models (right portion of Table 3) - i.e. those trained from scratch from English biomedical corpora. Because of the syntactical similarities of English and French, one strategy for creating French language biomedical LMs is simply to co-opt English language biomedical LMs, as was tested in Labrak *et al.* (2023); Touchent *et al.* (2023); Segonne *et al.* (2024). We will be testing ClinicalBERT (Huang *et al.*, 2019) and PubMedBERT (Gu *et al.*, 2020).

Table 3: Statistics for each model included in this paper

statistic	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
Language	French	French	French	French	French	French	English	English
Domain	Bio-med	Bio-med	Bio-med	Bio-med	General	General	Bio-med	Bio-med
Train steps	80k	50k	50k	100k	240k	50k	200k	63k
Data size	4GB	2.7GB	5.4GB	24GB	4GB	71GB	5GB	21GB

4 Experimental setup

The goal of this empirical section is two-fold:

1. We seek to assess the effect of model type on benchmark type. We want to evaluate in which contexts domain-specific MLMs are useful, and consistently reliable. We measure utility by mean performance on a certain task, and reliability by low variance on replications of different splits of the data for each task as well as and different classifier initializations. We will achieve this by evaluating the six benchmarks in Table 2 on different classes of model and comparing their performances.

2. We are interested in how much information is stored in each model during only pre-training, to ascertain whether the models are useful out-of-the-box. Many applications involve employing pre-trained token embeddings from MLMs without fine-tuning them to a specific downstream task, so it is important to test whether these token-embeddings are useful in a specific downstream setting. To test this, we train the models in two settings: first, in conventional, "unfrozen", end-to-end training, in which all model parameters may be updated during fine-tuning; and second, in "frozen" fine-tuning, in which the model's pre-trained weights remain fixed during fine-tuning, and only the classification layer(s)[9] are updated.

We train each model on each of the six benchmarks. We cross-validate the learning rate for each model and dataset using a random 80/10/10 train/valid/test split, each for up to 2000 steps, stopping early given validation set convergence. We repeat this training for frozen and unfrozen model weights. We replicate each experiment twenty times to gauge each model's consistency.

4.1 Summary and discussion of results

1. Each experiment, frozen or otherwise, saw a French biomedical LM perform best, as illustrated in Figures 1 and 2, and tabulated explicitly in Appendix Tables 4 and 5. For all experiments apart from non-frozen POS tagging, that best performing model was `CamemBERT-bio`. However, the other French biomedical models all fall short on some of the NER and NLU tasks: `DrBERT` is nearly as strong as `CamemBERT-bio` on `CLISTER` for non-frozen fine-tuning, but much worse on `FrenchMedMCQA` and frozen `CLISTER`. The two `jargon` models are significantly inferior on almost all tasks, despite performing best on the non-frozen POS tagging tasks. So as to the question whether French biomedical LM training is worthwhile, the answer appears to be yes for the specific case of `CamemBERT-bio`, but should be studied further to determine why the other models are unable to replicate its performance.

[9] We use a single linear classification layer for all tasks except `FrenchMedMCQA`, where we use two, on the advice of the authors (Labrak *et al.*, 2022).

Furthermore, the model CamemBERT-CCNet performs reliably worse than CamemBERT-bio, though never by too huge of a margin, while the English biomedical models are inferior to CamemBERT-CCNet at almost every task. This motivates the usage of general purpose same-language LMs over English biomedical models LMs for languages without dedicated biomedical LMs. However, we caution that given the fact that these benchmarks were created without the input of medical professionals, this trend could be misleading. It is worth noting that the only benchmark in our study which was created by medical professionals - FrenchMedMCQA - resulted in an English language model, PubMedBERT, outperforming CamemBERT-CCNet.

Lastly, we note the lack of consistency on the NLU datasets. The standard deviation of test scores (red lines in Figures 1 and 2) for FrenchMedMCQA (and to a slightly lesser extent for CLISTER) are great, illustrating the unreliability of using a French biomedical LM on related tasks. Such a wide performance range renders clinical models much less useful.

2. We show that some models are usable without end-to-end fine-tuning, while others should not be. For example, as shown in Figure 3, CamemBERT-bio has a consistently small improvement (even negative for CLISTER) when model weights are unfrozen, while both the English biomedical LMs and Jargon-biomed tend to improve significantly with unfrozen weights. For English LMs, this is not surprising, given the model is adapting to a new language. For Jargon-biomed, this suggests pre-training that is somehow inferior compared to that of CamemBERT-bio.

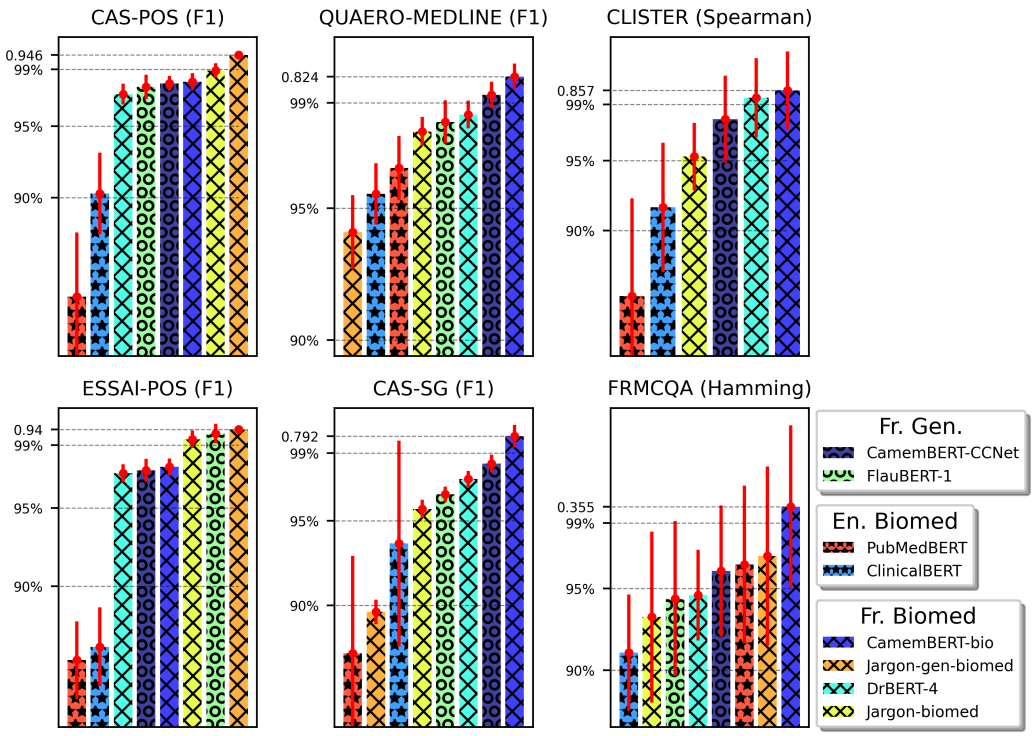

Figure 1: We compare the test-set scores of each model on each benchmark with **unfrozen** model weights. CamemBERT-bio is the best performer on all but the POS tasks. For scaling purposes, we left off models which performed significantly worse than the top model for each task.

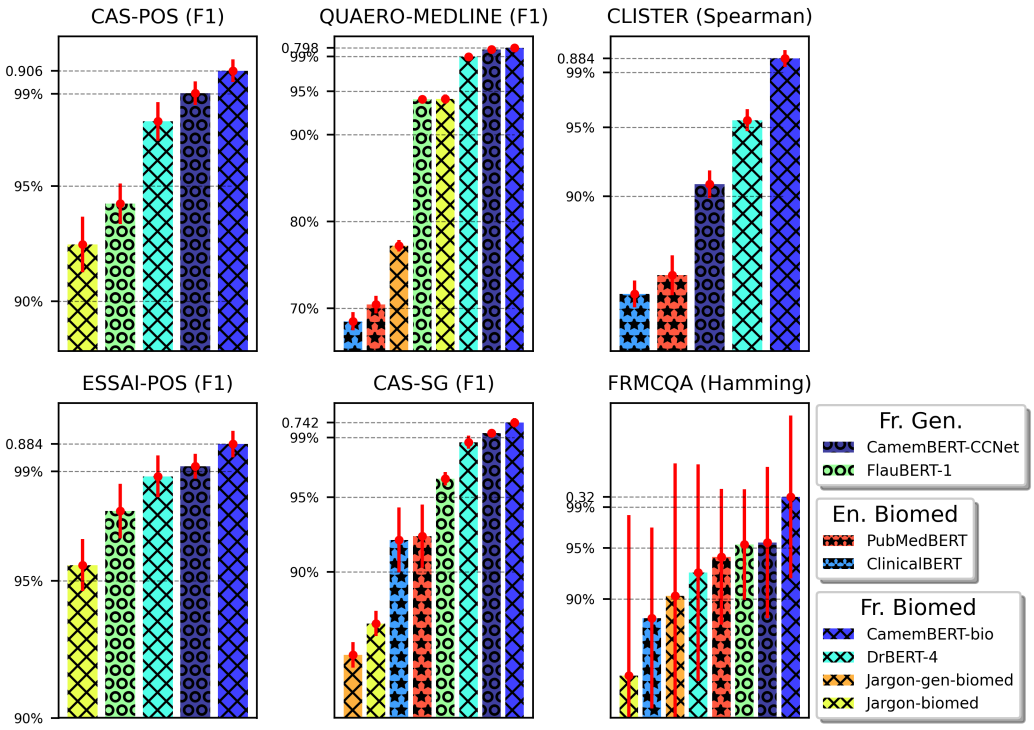

Figure 2: We compare the test-set scores of each model on each benchmark with **frozen** model weights. `CamemBERT-bio` is the best performing model on all tasks.

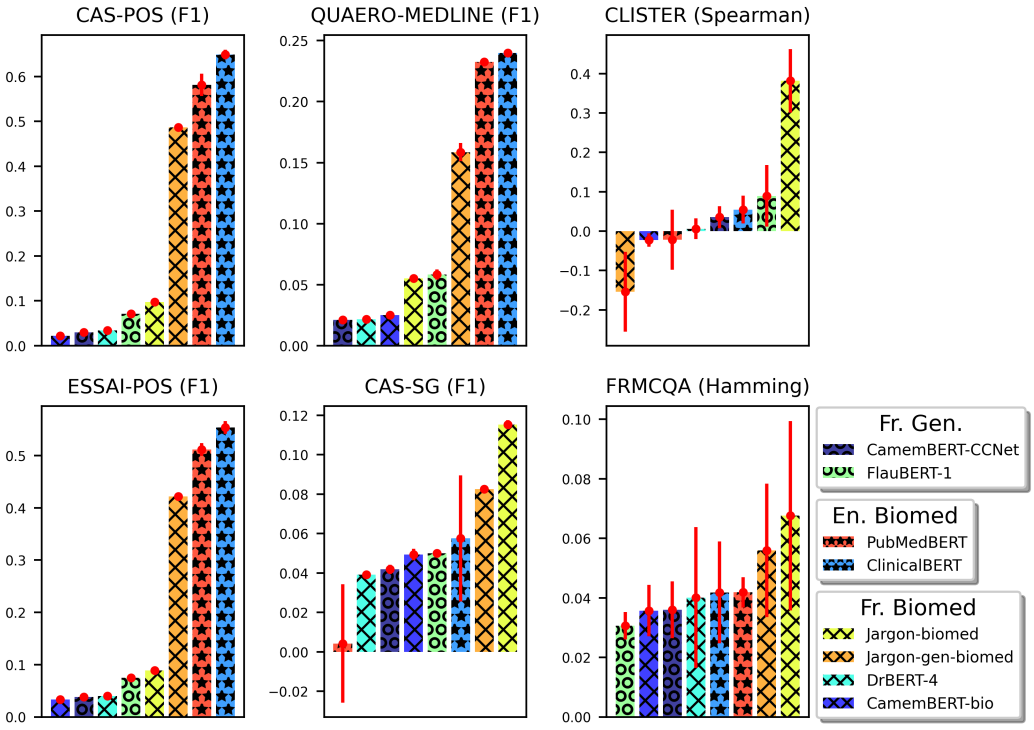

Figure 3: We calculate the difference between all pairs of tests (**frozen** and **unfrozen**) for each model and benchmark. A model that improves with unfrozen weights (as most do) has a positive score.

5 Conclusion and future work

We set out to study the state of benchmarking for French language biomedical LMs. We show that most clinical NLP tasks are best viewed through an NLU lens, and discuss the importance of benchmarks targeting specific use cases. Despite this, we show that the quantity and quality of biomedical NLU tasks is lacking in many languages, a trend particularly noticeable in French. With this in mind, we recommend that immediate future study of French biomedical NLP go towards improving benchmarking before it goes toward improving models. While benchmark creation may be less exciting than model development, it is essential to properly understand where our current models stand with respect to potential clinical application. We propose two criteria for benchmark design, inspired in large part by the excellent Russian and Chinese biomedical benchmarks described respectively in Blinov *et al.* (2022) and Zhang *et al.* (2022). Biomedical benchmarks should be:

1. constructed with a specific target use case in mind and in concert with biomedical professionals. These envisioned use cases should be briefly delineated in papers that apply them.

2. accompanied with a performance threshold above which a model could considered to be ready for some real life use. This will help users interpret the models' performances in an absolute sense, which is not currently the case for NLU benchmarks like `CLISTER` or `FrenchMedMCQA`.

Once this threshold for benchmark quality has been met, we can begin to pose more refined questions regarding a biomedical benchmark's quality, as has been done for domains with better established benchmarks (Bowman & Dahl, 2021; Dehghani *et al.*, 2021). For example, the AFLITE algorithm can be used to de-bias datasets for repetitiveness and prohibit models from picking up on spurious correlations (Sakaguchi *et al.*, 2021). However, given the nascent state of French biomedical NLP benchmarking, such sophisticated methods are not yet relevant.

Through experimentation, we observe that while all tasks benefit from domain-specific pre-training, the effect is most pronounced for NLU tasks[10]. While we identified one model which outperforms the others (`CamemBERT-bio`), even this model suffers from high variance under experimental replication. Therefore, we recommend further study of `CamemBERT-bio` and why it significantly outperforms its competitors. Are its training data higher quality, its architecture more effective, its pre-training strategy better? A brief analysis does not reveal any significant difference in construction and pretraining between any of the three French biomedical MLMs studied in this paper (Touchent *et al.*, 2023; Labrak *et al.*, 2023; Segonne *et al.*, 2024)[11].

Finally, we recommend a study into the rate at which LMs (French biomedical LMs included) are used without end-to-end finetuning. Barring a near-zero rate, we recommend regular frozen evaluation to complement end-to-end finetuning in subsequent publications.

[10]We note that our empirical conclusions were drawn based on results from two NLU benchmarks, a pittance when compared to the vast potential use cases for biomedical LMs. This conclusion should be re-evaluated once French biomedical benchmarking has advanced.

[11]The most notable exception is that `jargon` uses the Linformer architecture (Wang *et al.*, 2020), though studies have shown this architecture to perform like Transformer, and thus is unlikely to be the source of observed inferior performance.

References

AGARWAL A., BAECHLE C., BEHARA R. & ZHU X. (2018). A natural language processing framework for assessing hospital readmissions for patients with copd. *IEEE Journal of Biomedical and Health Informatics*, **22**(2), 588–596. DOI : 10.1109/JBHI.2017.2684121.

BLINOV P., RESHETNIKOVA A., NESTEROV A., ZUBKOVA G. & KOKH V. (2022). *RuMedBench: A Russian Medical Language Understanding Benchmark*, In *Lecture Notes in Computer Science*, p. 383–392. Springer International Publishing. DOI : 10.1007/978-3-031-09342-5_38.

BOWMAN S. R. & DAHL G. (2021). What will it take to fix benchmarking in natural language understanding? In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4843–4855, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.385.

CARCHIOLO V., LONGHEU A., REITANO G. & ZAGARELLA L. (2019). Medical prescription classification: a nlp-based approach. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, p. 605–609. DOI : 10.15439/2019F197.

CARDON R., GRABAR N., GROUIN C. & HAMON T. (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*, p. 1–13, Nancy, France: ATALA et AFCP.

CARRINO C. P., LLOP J., PÀMIES M., GUTIÉRREZ-FANDIÑO A., ARMENGOL-ESTAPÉ J., SILVEIRA-OCAMPO J., VALENCIA A., GONZALEZ-AGIRRE A. & VILLEGAS M. (2022). Pre-trained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, p. 193–199, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.bionlp-1.19.

DEHGHANI M., TAY Y., GRITSENKO A. A., ZHAO Z., HOULSBY N., DIAZ F., METZLER D. & VINYALS O. (2021). The benchmark lottery. *ArXiv*, **abs/2107.07002**.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota: Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EUROPEAN PARLIAMENT AND COUNCIL (2016). Regulation (eu) 2016/679. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

FUCHS K. (2023). Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? In *Frontiers in Education*, volume 8, p. 1166682: Frontiers.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2020). Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, **abs/2007.15779**.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2022). CLISTER : A corpus for semantic textual similarity in French clinical narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4306–4315, Marseille, France: European Language Resources Association.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, **abs/1904.05342**.

JOHNSON A. E. W., POLLARD T. J., SHEN L., LEHMAN L.-W. H., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., ANTHONY CELI L. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, **3**(1), 160035.

KANWAL N. & RIZZO G. (2022). Attention-based clinical note summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, p. 813–820, New York, NY, USA: Association for Computing Machinery. DOI : 10.1145/3477314.3507256.

KARARGYRIS A., UMETON R., SHELLER M. J., ARISTIZABAL A., GEORGE J., WUEST A., PATI S., KASSEM H., ZENK M., BAID U., NARAYANA MOORTHY P., CHOWDHURY A., GUO J., NALAWADE S., ROSENTHAL J., KANTER D., XENOCHRISTOU M., BEUTEL D. J., CHUNG V., BERGQUIST T., EDDY J., ABID A., TUNSTALL L., SANSEVIERO O., DIMITRIADIS D., QIAN Y., XU X., LIU Y., GOH R. S. M., BALA S., BITTORF V., PUCHALA S. R., RICCIUTI B., SAMINENI S., SENGUPTA E., CHAUDHARI A., COLEMAN C., DESINGHU B., DIAMOS G., DUTTA D., FEDDEMA D., FURSIN G., HUANG X., KASHYAP S., LANE N., MALLICK I., MASCAGNI P., MEHTA V., MORAES C. F., NATARAJAN V., NIKOLOV N., PADOY N., PEKHIMENKO G., REDDI V. J., REINA G. A., RIBALTA P., SINGH A., THIAGARAJAN J. J., ALBRECHT J., WOLF T., MILLER G., FU H., SHAH P., XU D., YADAV P., TALBY D., AWAD M. M., HOWARD J. P., ROSENTHAL M., MARCHIONNI L., LODA M., JOHNSON J. M., BAKAS S., MATTSON P., FETS CONSORTIUM, BRATS-2020 CONSORTIUM & AI4SAFECHOLE CONSORTIUM (2023). Federated benchmarking of medical artificial intelligence with MedPerf. *Nature Machine Intelligence*, **5**(7), 799–810. DOI : 10.1038/s42256-023-00652-2.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (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)*, p. 41–46, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. DOI : 10.18653/v1/2022.louhi-1.5.

LABRAK Y., BAZOGE A., DUFOUR R., ROUVIER M., MORIN E., DAILLE B. & GOURRAUD P.-A. (2023). DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 16207–16221, Toronto, Canada: Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.896.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français (FlauBERT : Unsupervised language model pre-training for French).

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). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278, Nancy, France: ATALA et AFCP.

LI X.-B. & QIN J. (2017). Anonymizing and sharing medical text records. *Information Systems Research*, **28**, 332–352. DOI : 10.1287/isre.2016.0676.

LIN B. Y., HE C., ZE Z., WANG H., HUA Y., DUPUY C., GUPTA R., SOLTANOLKOTABI M., REN X. & AVESTIMEHR S. (2022). FedNLP: Benchmarking federated learning methods for natural language processing tasks. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Findings of the Association for Computational Linguistics: NAACL 2022*, p. 157–175, Seattle, United States: Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-naacl.13.

MACHANAVAJJHALA A., KIFER D., GEHRKE J. & VENKITASUBRAMANIAM M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **1**(1), 3–es.

MAMEDE N., BAPTISTA J. & DIAS F. (2016). Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, p. 1287–1294. DOI : 10.1109/CEC.2016.7743936.

MARTIN L., MULLER B., SUÁ REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In D. DEMNER-FUSHMAN, K. B. COHEN, S. ANANIADOU & J. TSUJII, Éds., *Proceedings of the 18th BioNLP Workshop and Shared Task*, p. 58–65, Florence, Italy: Association for Computational Linguistics. DOI : 10.18653/v1/W19-5006.

PLAZA-DEL ARCO F. M., MOLINA-GONZÁLEZ M. D., URENA-LÓPEZ L. A. & MARTÍN-VALDIVIA M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, **166**, 114120. DOI : 10.1016/j.eswa.2020.114120.

RABHI S. (2022). *Optimized deep learning-based multimodal method for irregular medical timestamped data*. Theses, Institut Polytechnique de Paris. HAL : tel-03600526.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In J. SU, K. DUH & X. CARRERAS, Éds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392, Austin, Texas: Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

RIEKE N., HANCOX J., LI W., MILLETARÌ F., ROTH H. R., ALBARQOUNI S., BAKAS S., GALTIER M. N., LANDMAN B. A., MAIER-HEIN K., OURSELIN S., SHELLER M., SUMMERS R. M., TRASK A., XU D., BAUST M. & CARDOSO M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, **3**(1), 119. DOI : 10.1038/s41746-020-00323-1.

SAKAGUCHI K., BRAS R. L., BHAGAVATULA C. & CHOI Y. (2021). Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, **64**(9), 99–106. DOI : 10.1145/3474381.

SCHOCH S., YANG D. & JI Y. (2020). "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In S. AGARWAL, O. DUŠEK, S. GEHRMANN, D. GKATZIA, I. KONSTAS, E. VAN MILTENBURG & S. SANTHANAM, Éds., *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, p. 10–16, Online (Dublin, Ireland): Association for Computational Linguistics.

SEGONNE V., MANNION A., CANUL L. C. A., AUDIBERT A., LIU X., MACAIRE C., PUPIER A., ZHOU Y., AGUIAR M., HERRON F., NORRÉ M., AMINI M. R., BOUILLON P., ESHKOL-TARAVELLA I., ESPERANÇA-RODIER E., FRANÇOIS T., GOEURIOT L., GOULIAN J., LAFOURCADE M., LECOUTEUX B., PORTET F., RINGEVAL F., VANDEGHINSTE V., COAVOUX M., DINARELLI M. & SCHWAB D. (2024). Jargon: A suite of language models and evaluation tasks for french specialized domains. In *Proceedings of the LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation [Forthcoming]*.

SWEENEY L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, **10**(05), 557–570.

TOUCHENT R., ROMARY L. & DE LA CLERGERIE E. (2023). CamemBERT-bio : Un modèle de langue français savoureux et meilleur pour la santé. In C. SERVAN & A. VILNAT, Éds., *18e Conférence en Recherche d'Information et Applications*
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, p. 323–334, Paris, France: ATALA. HAL : hal-04130187.

UNITED STATES DEPARTMENT OF HEALTH AND HUMAN SERVICES (2013). 45 cfr parts 160 and 164.
https://www.govinfo.gov/content/pkg/FR-2013-01-25/pdf/2013-01073.pdf.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA: Curran Associates Inc.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. LINZEN, G. CHRUPAŁA & A. ALISHAHI, Éds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, p. 353–355, Brussels, Belgium: Association for Computational Linguistics. DOI : 10.18653/v1/W18-5446.

WANG S., LI B. Z., KHABSA M., FANG H. & MA H. (2020). Linformer: Self-attention with linear complexity.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4003–4012, Marseille, France: European Language Resources Association.

WINOGRAD A. (2023). Loose-lipped large language models spill your secrets: The privacy implications of large language models. *Harvard Journal of Law & Technology*, **36**(2).

YANG R., TAN T. F., LU W., THIRUNAVUKARASU A. J., TING D. S. W. & LIU N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, **2**(4), 255–263.

ZHANG C., XIE Y., BAI H., YU B., LI W. & GAO Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, **216**, 106775. DOI : https://doi.org/10.1016/j.knosys.2021.106775.

ZHANG N., CHEN M., BI Z., LIANG X., LI L., SHANG X., YIN K., TAN C., XU J., HUANG F., SI L., NI Y., XIE G., SUI Z., CHANG B., ZONG H., YUAN Z., LI L., YAN J., ZAN H., ZHANG K., TANG B. & CHEN Q. (2022). CBLUE: A Chinese biomedical language understanding evaluation benchmark. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 7888–7915, Dublin, Ireland: Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.544.

6 Appendix

Table 4: Test set results for frozen models

Dataset	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
CAS-POS	0.886 ± 0.008	0.906 ± 0.004	0.838 ± 0.011	0.459 ± 0.01	0.897 ± 0.005	0.854 ± 0.008	0.205 ± 0.019	0.204 ± 0.018
QUAERO-MEDLINE	0.79 ± 0.003	0.798 ± 0.003	0.751 ± 0.004	0.616 ± 0.005	0.797 ± 0.002	0.751 ± 0.003	0.547 ± 0.008	0.562 ± 0.008
CLISTER	0.845 ± 0.007	0.884 ± 0.005	0.435 ± 0.082	0.51 ± 0.078	0.804 ± 0.009	0.473 ± 0.018	0.733 ± 0.009	0.745 ± 0.013
ESSAI-POS	0.874 ± 0.007	0.884 ± 0.004	0.845 ± 0.008	0.517 ± 0.009	0.877 ± 0.004	0.862 ± 0.009	0.256 ± 0.03	0.29 ± 0.034
CAS-SG	0.733 ± 0.003	0.742 ± 0.002	0.643 ± 0.006	0.627 ± 0.006	0.737 ± 0.002	0.715 ± 0.003	0.684 ± 0.016	0.686 ± 0.016
FrenchMedMCQA	0.296 ± 0.034	0.32 ± 0.025	0.264 ± 0.05	0.289 ± 0.041	0.305 ± 0.024	0.305 ± 0.017	0.282 ± 0.028	0.301 ± 0.021

French bio-medical models are purple, French general-purpose models cyan, and English bio-medical models grey. POS and NER tasks are evaluated using F1 score; CLISTER is evaluated using the Spearman ranked correlation coefficient; FrenchMedMCQA is evaluated using either the Hamming distance between the (potentially) multiple correct answers and the answers chosen by the model.

Table 5: Test set results for non-frozen models

Dataset	DrBERT-4	CamemBERT-bio	Jargon-biomed	Jargon-gen-biomed	CamemBERT-CCNet	FlauBERT-1	ClinicalBERT	PubMedBERT
CAS-POS	0.92 ± 0.007	0.928 ± 0.006	0.936 ± 0.005	0.946 ± 0.002	0.927 ± 0.005	0.925 ± 0.008	0.854 ± 0.027	0.786 ± 0.042
QUAERO-MEDLINE	0.812 ± 0.004	0.824 ± 0.004	0.806 ± 0.005	0.775 ± 0.012	0.818 ± 0.004	0.809 ± 0.007	0.787 ± 0.01	0.795 ± 0.01
CLISTER	0.853 ± 0.024	0.857 ± 0.024	0.817 ± 0.021	0.367 ± 0.07	0.839 ± 0.027	0.563 ± 0.098	0.786 ± 0.04	0.731 ± 0.06
ESSAI-POS	0.913 ± 0.005	0.917 ± 0.005	0.934 ± 0.005	0.94 ± 0.003	0.915 ± 0.007	0.937 ± 0.006	0.809 ± 0.024	0.801 ± 0.023
CAS-SG	0.772 ± 0.004	0.792 ± 0.005	0.758 ± 0.004	0.71 ± 0.006	0.779 ± 0.004	0.765 ± 0.004	0.742 ± 0.048	0.69 ± 0.046
FrenchMedMCQA	0.336 ± 0.01	0.355 ± 0.018	0.331 ± 0.019	0.345 ± 0.019	0.341 ± 0.014	0.335 ± 0.017	0.324 ± 0.013	0.343 ± 0.017

Almost all models experienced performance improvement on all tasks when their weights were unfrozen.