ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 10, Number 2
december 2024
Special Issue

Natural Language for Artificial Intelligence in the Era of LLMs

aA Accademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255001430

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_10_2

IJCoL

Volume 10, Number 2
december 2024
Special Issue

Natural Language for Artificial Intelligence in the Era of LLMs

editors: *Elisa Bassignana, Dominique Brunato, Marco Polignano, Alan Ramponi*

CONTENTS

BERTinchamps: Cost-Effective In-House Training of Language Models in French

Amaury Fierens *
UCLouvain

Sébastien Jodogne **
UCLouvain

Many in-house applications are envisioned for Language Models (LMs) across various fields. In the medical domain, LMs could automate tasks such as summarizing the health condition of a patient and codifying electronic health records. They also hold potential in the legal field and in journalism. While training LMs directly inside an institution is desirable for leveraging local data and addressing data privacy concerns, this process demands a costly and complex computational infrastructure. This paper explores the recent Cramming approach as a cost-effective way to locally train medium-sized LMs, in one day and using one graphics processing unit (GPU). We show that the Cramming approach that was originally designed for the English language can be transposed to French, that the resulting models can be fine-tuned to domain-specific tasks in the French language, and that pre-training by including in-house data increases the performance of the models for journalism data. This research opens the path to the creation of medium-sized LMs that are tailored to the specific needs of institutions that handle sensitive textual data in another language than English.

1. Introduction

The field of Natural Language Processing (NLP) is currently attracting significant interest for in-house applications, notably in the context of healthcare (Li et al. 2024; Zhou et al. 2022), law (Wang et al. 2023), or journalism (Cheng 2024). Specifically in the clinical field, automating the codification of electronic health records (EHRs) could be highly valuable for monitoring the quality of treatments (Pronovost, Cole, and Hughes 2022), assisting with hospital payment reimbursements (Zhou et al. 2020), providing summaries of the health condition of patients (Watzlaf et al. 2007), and detecting diseases at an early stage by using clinical codes as biomarkers (Poongodi et al. 2021).

In particular, the recent major advances in the field of Language Models (LMs) are opening great opportunities for NLP (Anil and others 2023; Khurana et al. 2023; OpenAI 2023; Touvron et al. 2023). A growing number of physicians are interested in leveraging advanced tools such as the widely recognized ChatGPT chatbot to assist them in medical tasks (Biswas 2023; Li et al. 2024). At the time of writing, ChatGPT internally uses the closed-source, proprietary LMs GPT-3.5 and GPT-4 (OpenAI 2023), which introduces a strong dependency upon the proprietary infrastructure of the OpenAI platform. But, in clinical settings specifically, the protection of health information prevents the direct

* Institute for Information and Communication Technologies, Electronics and Applied Mathematics, INGI department – Place Sainte Barbe 2/L5.02.01, 1348 Louvain-la-Neuve, Belgium.
E-mail: amaury.fierens@uclouvain.be

** Institute for Information and Communication Technologies, Electronics and Applied Mathematics, INGI department – Place Sainte Barbe 2/L5.02.01, 1348 Louvain-la-Neuve, Belgium.
E-mail: sebastien.jodogne@uclouvain.be

use of proprietary LMs, primarily due to their cloud-based nature, while regulations such as the General Data Protection Regulation (GDPR) in Europe forbid patient data from leaving hospitals unless it has been at least pseudonymized. Similar difficulties are encountered when a hospital seeks to exploit LMs in the context of clinical research.

This calls for the development of LMs that can be entirely self-hosted inside the computational infrastructure of an institution, such as a hospital or a court. Self-hosting is highly desirable for protecting sensitive information during the process of inference on electronic health records (EHRs) in hospitals or on trial reports in courts. However, in addition to inference, it is also crucial to have the capability to train LMs directly within the institution. This allows for fine-tuning the LMs to the specific patient population of a hospital or a particular clinical department of interest. Similar considerations apply in legal contexts, where the characteristics of the population of one court can differ significantly from another. Self-hosting can be notably achieved by taking advantage of LMs whose architectures have been published as open-source code, and whose pre-trained weights are available as open data. Early Language Models available as open-source and open-data assets include BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019). More advanced models such as BLOOM (Scao and others 2023), Cerebras-GPT (Dey et al. 2023), or LLaMA (Touvron et al. 2023) are now available.

A difficulty with open, general-purpose LMs is that they have been primarily trained on English datasets, without a specialization on the clinical, legal, or journalistic language. This has motivated researchers to train LMs using corpora containing specialized documents such as medical or legal texts. This is possible for English, for which corpora of sufficient size have emerged over the years. For instance, BioBERT (Lee et al. 2019) was trained on a dataset made of PubMed abstracts, while ClinicalBERT (Alsentzer et al. 2019) was trained on MIMIC-III (Johnson et al. 2016). In the context of legal applications, Legal-BERT (Chalkidis et al. 2020) was trained on a large corpus containing all EU legislation, UK legislation, and all publicly accessible case reports from United States. Unfortunately, there is still a lack of large-scale medical and legal corpora for most languages besides English. In particular, only a handful of pre-trained LMs for the French clinical language are currently available. Those include DrBERT that leverages the RoBERTa architecture (Liu et al. 2019) and that is trained from scratch using the biomedical corpus NACHOS (Labrak et al. 2023), as well as the CamemBERT-bio model (Touchent, Romary, and De La Clergerie 2023) that is based on the CamemBERT architecture (Martin et al. 2020). The evaluation of such language-specific models on real-world clinical tasks is still a work in progress. In the context of the French legal language, two main pre-trained LMs are available: CamemBERT Judiciaire (Mahmoudi et al. 2022) and JuriBERT (Douka et al. 2021), which are both based on the CamemBERT architecture. Unfortunately, due to the lack of a publicly available corpus, neither of these models could be trained on actual French-speaking court reports.

An alternative to the use of LMs that are externally pre-trained for specialized applications would be to train the LMs internally, on the local computational infrastructure, directly on the data privately hosted by the institution. This approach would have the great advantage of training models that better reflect the local setup, while bringing privacy by design. It is commonly considered that training a sufficiently expressive LM from scratch is extremely demanding in terms of time and computational resources, because standard training processes require dozens of days of computation on powerful Graphics Processing Units (GPU) that come at a high budget. However, recent work has shown that it is possible to drastically reduce the training time of medium-sized, BERT-type LMs for the English language by slightly modifying the BERT architecture and

the way datasets are preprocessed (Geiping and Goldstein 2023). This simplification is referred to as "Cramming" and the resulting LMs are called "crammed BERT" models. To the best of our knowledge, the Cramming recipe has only been studied in the context of English and has not been applied to the medical field so far.

This paper investigates the applicability of the Cramming recipe to train medium-sized LMs on specialized tasks in French. It consists of a revised, extended version of an article presented at the 7th edition of the Workshop on Natural Language for Artificial Intelligence (Bassignana et al. 2023; Fierens and Jodogne 2023). The original results showed that our crammed BERT model, referred to as BERTinchamps, achieves a performance that is close to that of CamemBERT, both on general-purpose and healthcare-related tasks, while requiring only one single day of training. In addition to this original material, this paper analyzes the possibility of training a BERTinchamps model for journalism in-house using the RTBF corpus, a public dataset of 750,000 Belgian French news articles published between 2008 and 2021 (Escouflaire et al. 2023). These new results indicate that the original BERTinchamps model, trained with 30GB of text from the OSCAR dataset, achieves performance comparable to that of the state-of-the-art CamemBERT model on specialized tasks related to journalism. Another BERTinchamps model, trained on 29GB of text from OSCAR and 1.6GB of text from the RTBF corpus, achieves slightly better performance than the original BERTinchamps model on journalism-related tasks, without any decline in performance on general tasks. These contributions open the path to the training of medium-sized LMs for the French language, directly inside institutions that generate sensitive textual data, notably hospitals, at a reasonable cost, while preserving the privacy of data.

2. Related Work

The idea of optimizing LM architectures to reduce resource requirements during training has already been explored in the literature. In 2015, Knowledge Distillation was introduced in neural networks to train smaller models from a bigger one, with little loss in performance (Hinton, Vinyals, and Dean 2015). This approach was used to create distilled versions of well-known LMs, such as DistilBERT (Sanh et al. 2020) and DistilGPT-2 (Li et al. 2021). Another technique consists in the quantization of the model (Fiesler, Choudry, and Caulfield 1990). Since its introduction in 1990, quantization has been widely applied to the Transformer-based architectures that underpin BERT, for instance in BinaryConnect (Courbariaux, Bengio, and David 2015), in the paper that introduced quantization for BERT (Bondarenko, Nagel, and Blankevoort 2021), in Q8BERT (Zafrir et al. 2019), or in BinaryBERT (Bai et al. 2021). An even more recent paper has presented QLoRA, an efficient fine-tuning method for quantized models (Dettmers et al. 2024).

While Knowledge Distillation and quantization are extremely useful to reduce the size of an already existing LM, they are not designed to train LMs from scratch at a decent cost. The Cramming recipe is a recent contribution to serve this purpose (Geiping and Goldstein 2023). In this method, BERT-like models with 120 million of parameters are trained under extreme resource constraints, specifically limited to a single GPU and within a one-day timeframe. This is achieved by modifying the architecture of the model, by simplifying the data preprocessing pipeline, and by using highly optimized training techniques. Despite these computational limitations, the resulting models, known as "crammed BERT" models, demonstrate strong performance on common benchmarks, which proves that medium-sized Language Models can be trained effectively with limited hardware. This is made possible by employing techniques such as model simplifications, smaller batch sizes, mixed precision training, and efficient

learning rate schedules. The Cramming recipe challenges the assumption that high-performance models require massive computational resources, opening new perspectives for training accessible and efficient Language Models.

The Cramming recipe is motivated by the scaling laws that apply in the low-resource regime (Kaplan et al. 2020). These scaling laws suggest that it is not necessarily useful to reduce the number of parameters of BERT-like architectures. Instead of reducing the number of parameters, the Cramming recipe adjusts the training setup and optimizes the model architecture, notably by disabling the QKV biases in the multi-head attention block (Vaswani et al. 2017) and by adapting the embedding blocks. The Cramming recipe also explores architectural enhancements that speed up the computation of the gradients and proposes hyperparameters that are better suited for pre-training crammed BERT models. Finally, careful selection and processing of the training data is applied to extract well-suited tokens, enhancing the overall performance of the crammed models. These contributions have been shown to bring significant improvements, enabling the fast training of BERT-like models for English. However, the application of the Cramming recipe to other languages and the fine-tuning of crammed BERT models on domain-specific tasks is still largely unexplored. Our paper addresses this gap by applying the Cramming approach to the French language and by examining the performance of crammed BERT models on specialized tasks in French.

3. Methods

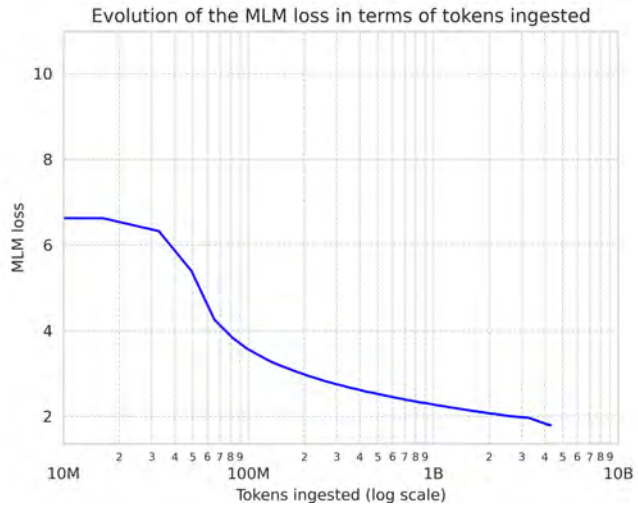
In this section, we describe our methodology and the parameters employed to mirror the Cramming recipe for the French language. The resulting crammed BERT model is referred to as BERTinchamps. The fine-tuning of BERTinchamps on selected healthcare-related tasks is then discussed, followed by pre-training experiments in the context of journalism using the RTBF corpus.

3.1 Initial Pre-Training

BERTinchamps was trained on the OSCAR dataset (Ortiz Suárez, Romary, and Sagot 2020). A subset of 30GB of the French part of the OSCAR dataset was randomly selected. This choice of 30GB of data corresponds to the amount of data that can be processed once within the limited time of 24 hours, as prescribed by the Cramming recipe. Tokens were extracted using the WordPiece algorithm (Schuster and Nakajima 2012) with a vocabulary size $n_{vocab} = 32768 = 2^{15}$, as proposed in the Cramming paper. This value is motivated by the claims of previous work indicating that choosing a total vocabulary of around 32k word pieces achieves both good accuracy and fast decoding speed (Wu et al. 2016). The cross-entropy loss was optimized, which is a standard practice in training BERT models. However, to accelerate the training, the context was reduced from a maximum sequence length of 512 tokens to 128 tokens, as recommended in the Cramming paper. In the same vein, the training objective was kept as masked language modeling, with the same masking proportions as in the original setup of BERT (Devlin et al. 2019).

AdamW was used as the optimizer (Loshchilov and Hutter 2019), which is a modified version of the Adam optimizer (Kingma and Ba 2015) where weight decay is performed after controlling the parameter-wise step size. The parameters of AdamW were set as follows: Weight decay equals 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-12}$. A gradient clipping of 0.5 was also included. The learning rate was set to 10^{-3} and, contrary to the original instructions of the Cramming recipe, a slanted triangular learning rate was

(a)



(b)

Figure 1
Evolution of the Masked Language Model (MLM) loss during the training, as a function of (a) the number of micro-batch steps, and (b) the number of tokens ingested by the model. Note that plot (b) is on a semi-log scale, with the $x-$axis expressed in base 10. These plots show that the loss decreases as either of those numbers increases, as expected.

used as the scheduler, with a base percentage of 25% and a falloff of 0.25 as parameters. This adaptation was experimentally found to be more efficient for the French language.

A micro-batch size of 128 and a batch size of 4096 were used, in accordance with the assumption of the Cramming recipe that only one single GPU is available. This policy is rescheduled by linearly increasing the average number of micro-batches over the training time. The model was trained for 24 hours, as required by the Cramming

recipe, on one NVIDIA A4500 GPU with 20GB VRAM. The resulting model is referred to as BERTinchamps. Figure 1 depicts the evolution of the Masked Language Model (MLM) loss during the training.

3.2 Fine-Tuning Experiments

In a first phase, three tasks in the FLUE benchmark were used to assess the overall performance of BERTinchamps. FLUE is an equivalent of the well-known GLUE benchmark for the French language (Le et al. 2020). In a second phase, to determine whether our crammed model was promising for specialized tasks related to the medical domain, BERTinchamps was fine-tuned on the QUAEROFrenchMed benchmark, which is based on two French datasets, namely EMEA and MEDLINE (Névéol et al. 2014). The use of QUAEROFrenchMed enables the comparison of BERTinchamps against the general-purpose CamemBERT model (Martin et al. 2020), as well as against two specialized models: DrBERT, a LM for biomedical tasks in French (Labrak et al. 2023), and Erasmus-run1, the best model originally built for the "*Task 1b of the CLEF eHealth Evaluation Lab 2015*" that introduced the QUAEROFrenchMed benchmark (Névéol et al. 2015).

3.2.1 Fine-Tuning on a General-Purpose Benchmark

The FLUE benchmark is widely used to evaluate general-purpose LMs for the French language. It has notably been used to assess the performance of CamemBERT (Martin et al. 2020) and FlauBERT (Le et al. 2020), two of the most powerful LMs for French at the time of writing. BERTinchamps was evaluated on three tasks that are included in the FLUE benchmark, namely the XNLI, CLS, and PAWS-X datasets. The remaining tasks of the FLUE benchmark were not selected because their public repositories are broken and no longer maintained. Each dataset included within the FLUE benchmark assesses a specific capability of the LM. The XNLI dataset, a French subset of MNLI, is related to the Natural Language Inference (NLI) task, which consists in identifying logical relationships between premise and hypothesis sentences. The CLS dataset is employed for text classification, using star-based labels to categorize reviews of books, DVDs, and music from the Amazon online store. In this work, these three categories are consolidated into a single category, which we assume will not significantly impact classifier performance. The PAWS-X dataset targets paraphrasing identification, where paired sentences are tagged as 1 if they are semantically equivalent and 0 otherwise.

Most of the FLUE tasks involve fine-tuning for sequence classification, for which the approach described in the Cramming paper was used. Precisely, the AdamW optimizer was employed, with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$, and a learning rate of $4 \cdot 10^{-5}$. The cosine-decay scheduler was experimentally found to provide better performance for the fine-tuning. The model was trained for 10 epochs on CLS and PAWS-X, and for 5 epochs on XNLI. The training batch size was set to 16, while the testing batch size was set to 128.

3.2.2 Fine-Tuning on a Healthcare-Related Benchmark

The QUAEROFrenchMed benchmark is a medical Named-Entity Recognition (NER) task in the French language, whose purpose is to associate a clinical entity to each token of a medical text. The QUAEROFrenchMed benchmark is made of two distinct datasets encoded using the BRAT format, namely EMEA and MEDLINE, that share the same clinical entities of interest. MEDLINE is composed of a lot of short sentences coming

from MEDLINE article titles, while EMEA is composed of a few long text documents coming from drug descriptions.

The QUAEROFrenchMed benchmark involves the classification of tokens, which contrasts with the FLUE benchmark, for which sequence classification fine-tuning was required. To this end, a linear layer was added at the end of the pre-trained BERTinchamps model, with an output size equal to the number of clinical entities in QUAEROFrenchMed. This linear layer was trained using cross-entropy loss. Moreover, the original datasets had to be adapted to meet the requirements associated with the classification task. For MEDLINE, the annotations were processed to extract all the spans for entities that had multiple occurrences. For EMEA, the dataset was first adjusted to make it like MEDLINE by segmenting the long documents into individual sentences. The annotations were then processed the same way as in MEDLINE. Each word of both datasets was tokenized and each sentence of EMEA was identified using the French blank model of the spaCy Python package along with its Sentencizer tool. For both EMEA and MEDLINE, all the resulting words, along with their labels, were stored as a JSON file for further processing.

The AdamW optimizer was again used to fine-tune BERTinchamps on the EMEA and MEDLINE tasks. The default implementation of the AdamW trainer in the PyTorch package was used (Paszke et al. 2019), with a learning rate of 10^{-4}. The slanted triangular learning rate scheduler was used, as it provided better performance in this case. The model was trained for 100 epochs, for both EMEA and MEDLINE. Both the training and testing batch sizes were set to 8.

3.3 Pre-Training on a Domain-Specific Task

Section 3.1 explained how the "classic," general-purpose version of BERTinchamps was pre-trained on 30GB of data extracted from the OSCAR corpus alone. To assess the possibility of pre-training a LM from scratch using in-house data, another version of BERTinchamps was pre-trained on 29GB of data from the OSCAR corpus combined with the entire RTBF corpus, which contains 1.6GB of data. The pre-training used the same settings as in Section 3.1. The resulting LM is referred to as the "RTBF version" of BERTinchamps, and is specialized on journalism-related tasks in the French language. To assess the performance of such a domain-specific LM, two custom benchmarks were defined on the RTBF corpus:

TOPIC benchmark. The RTBF corpus associates each newspaper article it contains with a label describing its general topic (information, sports, radio, culture, ...). The 5 labels that are the most frequent in the corpus were selected. For each selected label, 2000 newspaper articles associated with this label and 2000 not associated with this label were randomly selected. Each group of 2000 articles was further divided into a training set of 1000 articles and a test set of 1000 articles. This procedure allows defining the TOPIC benchmark as 5 binary classification tasks, each determining whether a given newspaper article is associated with the label of interest.

SIGNATURE benchmark. In addition to the label describing its general topic, each newspaper article of the RTBF corpus is associated with its signing author. Similar to the TOPIC benchmark, the 5 most frequent human authors were selected. For each selected author, 2000 newspaper articles signed by this person and 2000 not signed by this person were randomly selected. Each group of 2000 articles was further divided into a training set of 1000 articles and a test set of 1000 articles.

Table 1
Comparison of accuracy among three medium-sized LMs on the three tasks considered in the FLUE benchmark. The percentage difference compared to CamemBERT is displayed next to each score.

Model	CLS	PAWS-X	XNLI
$\text{CamemBERT}^{\dagger}_{\text{base}}$	**93.33***	**90.14**	**81.2**
$\text{FlauBERT}^{\dagger}_{\text{base}}$	93.21* (-0.12)	89.49 (-0.65)	80.6 (-0.6)
BERTinchamps	89.48 (-3.85)	86.47 (-3.67)	77.21 (-3.99)

†Results reported in FlauBERT paper (Le et al. 2020) .

*Results averaged from the 3 categories.

The SIGNATURE benchmark again corresponds to 5 binary classification tasks, each determining whether a given newspaper article was signed by the author of interest.

Because both the TOPIC and the SIGNATURE benchmarks consist of 5 binary classification tasks, CamemBERT, the "classic version" of BERTinchamps, and the "RTBF version" of BERTinchamps were fine-tuned on the training set of each of those binary classification tasks, using the same hyperparameters as in Section 3.2.2. This leads to a total of 10 fine-tuned models for each of the 3 pre-trained LMs. The performance of these fine-tuned models on the test sets will be assessed in the next sections.

4. Results

This section first presents the overall performance of the "classic version" of the BERTinchamps model according to the FLUE benchmark. Secondly, the specific capabilities of the model related to the medical language are tested on the QUAEROFrenchMed benchmark. Finally, the performance of the "RTBF version" of BERTinchamps is assessed on the FLUE benchmark and on the two custom tasks defined on the RTBF corpus.

4.1 Performance on General-Purpose Benchmark

As explained in Section 3.2.1, the "classic version" of BERTinchamps was compared to CamemBERT and FlauBERT, two state-of-the-art LMs for the French language, on the CLS, PAWS-X, and XNLI tasks of the FLUE benchmark. Because BERTinchamps is a crammed BERT model with 120 million parameters, versions of CamemBERT and FlauBERT with a comparable number of parameters were considered (i.e., $\text{CamemBERT}_{\text{base}}$ and $\text{FlauBERT}_{\text{base}}$ that respectively contain 110 and 138 million parameters). Table 1 reports the final accuracy of each model on each task. The results show a difference of less than 4% in performance between the crammed BERTinchamps model and the state-of-the-art models.

4.2 Performance on Healthcare-Related Tasks

The capabilities of the "classic version" of BERTinchamps on healthcare-related tasks in the French language were evaluated on EMEA and MEDLINE, the two datasets

Table 2
Top: Comparison of accuracy between the three LMs on the two parts of QUAEROFrenchMed benchmark (EMEA and MEDLINE). *Bottom:* Comparison of F_1-score between the "classic version" of BERTinchamps and the best model at the time QUAEROFrenchMed was introduced.

Metrics	Model	MEDLINE	EMEA
Accuracy	CamemBERT$_{\text{base}}$	**80.60**	**90.54**
	DrBERT$_{7\text{GB}}$	80.10	90.34
	BERTinchamps	80.18	90.06
F_1-score	Erasmus-run1[†]	66.5	75.6
	BERTinchamps	**79.95**	**89.58**

[†]Results reported in CLEF eHealth paper (Névéol et al. 2015).

Table 3
Label-level F_1-scores for DrBERT and the "classic version" of BERTinchamps on EMEA and MEDLINE, along with the support in terms of token appearances.

	MEDLINE				EMEA			
Label	DrBERT		BERTinchamps		DrBERT		BERTinchamps	
	F_1	support	F_1	support	F_1	support	F_1	support
0	88.5	10006	**88.9**	8881	**94.9**	10785	94.8	9648
ANAT	**53.2**	622	52.6	763	**59.5**	119	58.2	135
CHEM	70	786	**73.9**	959	85.2	1874	**89.6**	2501
DEVI	6.4	85	**30**	81	26.6	191	**57.4**	245
DISO	73.5	2421	**77.4**	2940	77.3	624	**77.7**	835
GEOG	35.2	109	**66.2**	73	78.3	28	**87.2**	22
LIVB	66.8	603	**73.9**	605	80.7	315	**81.6**	322
OBJC	6.4	57	**31.2**	53	2.1	70	**8.6**	59
PHEN	14.4	78	**20.6**	77	**17.01**	39	13.9	36
PHYS	35.3	264	**36.8**	267	48.5	118	**50.2**	150
PROC	**66**	1084	65.6	1227	**64.8**	345	57.4	380

of the QUAEROFrenchMed benchmark. To this end, BERTinchamps was compared to DrBERT, CamemBERT, and Erasmus-run1. At the time QUAEROFrenchMed was introduced, the latter model provided the best F_1-score on the benchmark (Névéol et al. 2015). The three LMs were fine-tuned on QUAEROFrenchMed for the classification of tokens, using the experimental setup described in Section 3.2.2. The results for the three LMs are reported in Table 2. As can be seen in this table, the accuracy of BERTinchamps is close to DrBERT and CamemBERT on the investigated tasks, with differences of less than 1%. This is an interesting finding, as BERTinchamps was trained using far less computational resources, and exclusively on the general-purpose OSCAR dataset, which only provides BERTinchamps with sparse information about the structure of the French medical language. The F_1-score comparison between BERTinchamps and Erasmus-run1 shows that the former outperforms any of the models originally designed for the benchmark by at least 13%.

Table 4
Comparison of accuracy between the CamemBERT model, the "classic version" of BERTinchamps, and the "RTBF version" of BERTinchamps on the three considered tasks of the FLUE benchmark. The two first lines correspond to an excerpt of Table 1. The difference in percentage versus CamemBERT is displayed next to each score.

Model	CLS	PAWS-X	XNLI
$\text{CamemBERT}^{\dagger}_{\text{base}}$	**93.33***	**90.14**	**81.2**
$\text{BERTinchamps}_{\text{classic}}$	89.48 (-3.85)	86.47 (-3.67)	77.21 (-3.99)
$\text{BERTinchamps}_{\text{rtbf}}$	89.95 (-3.38)	87.58 (-2.56)	78.03 (-3.17)

*Results averaged from the 3 categories.

Table 3 contains the results for BERTinchamps and DrBERT at the level of the individual labels. The F_1-score is reported together with the support for each label. The mismatch in support counts is due to the use of different tokenizers in the two models, as the labels are applied to the individual tokens that make up each word. This table shows that BERTinchamps outperforms DrBERT on 9 out of the 11 labels in the MEDLINE dataset and on 7 out of 11 labels in the EMEA dataset. This suggests that BERTinchamps slightly surpasses DrBERT on the considered tasks.

4.3 Impact of Pre-Training on a Journalism-Related Corpus

As discussed in the previous section, the fine-tuning of BERTinchamps on healthcare-related tasks yields surprisingly good performance that is in line with the state-of-the-art DrBERT model, despite BERTinchamps not being specifically trained on medical texts. Consequently, it would be highly valuable to pre-train a BERTinchamps model on a French clinical corpus, similarly to the way DrBERT was trained, to determine whether this improves the performance of the model in the context of healthcare. Unfortunately, there is currently a lack of publicly available datasets containing medical texts in French, which prevented a direct exploration of this question in this research work.

As a fallback investigation, this section evaluates a version of the BERTinchamps model pre-trained from scratch on a domain for which a public dataset is available. Specifically, the model is pre-trained on journalism data, by leveraging the RTBF corpus. The primary goal is to ensure that incorporating in-house data alongside the state-of-the-art OSCAR corpus does not degrade performance by introducing some sort of noise. More precisely, both the "classic version" of BERTinchamps introduced in Section 3.1 and the "RTBF version" of BERTinchamps introduced in Section 3.3 are compared with CamemBERT on the general-purpose FLUE benchmark. This comparison with CamemBERT is motivated by its status as the most efficient state-of-the-art model for FLUE at the time of writing. In addition, the performance of these three LMs is investigated on the TOPIC and SIGNATURE benchmarks that were defined on the RTBF corpus and that are specific to the domain of French journalism.

Table 4 compares the three considered LMs on the FLUE benchmark in terms of accuracy. Two interesting facts can be drawn from this table. Firstly, the difference in accuracy between the "classic" and "RTBF" versions of BERTinchamps stays in a range of 1%. This implies that the loss of information on the structure of the French language due to the removal of part of the OSCAR dataset has been compensated for

Table 5
Comparison of accuracy between CamemBERT and both versions of BERTinchamps on the 5 most frequent topics of the RTBF corpus.

Model	MONDE	RÉGIONS	FOOTBALL	SOCIÉTÉ	BELGIQUE
$\text{CamemBERT}_{\text{base}}$	**99.5**	92.2	99.6	86.9	**98.7**
$\text{BERTinchamps}_{\text{classic}}$	99.35	89.55	99.45	86.8	98.1
$\text{BERTinchamps}_{\text{rtbf}}$	**99.5**	**92.85**	**99.75**	**88.5**	98.65

Table 6
Comparison of accuracy between CamemBERT and both versions of BERTinchamps on the 5 most frequent human authors of the RTBF corpus.

Model	OPPENS	ROUQUET	BIOURGE	WEYNANTS	DELPARTURE
$\text{CamemBERT}_{\text{base}}$	**92.45**	**95.85**	**84.15**	**99.4**	79.55
$\text{BERTinchamps}_{\text{classic}}$	89.7	95.05	81.25	98.5	82.8
$\text{BERTinchamps}_{\text{rtbf}}$	90.75	95.6	81.7	99.0	**83.7**

by the content of the RTBF corpus, and makes it perform even better. Secondly, the "RTBF version" of BERTinchamps has an accuracy that stays in a range of 3% from CamemBERT, which is in line with the "classic version" of BERTinchamps. In summary, Table 4 indicates that the version of BERTinchamps that is pre-trained for the journalism domain maintains its performance on a general-purpose benchmark.

The performance of the three considered LMs is now investigated on the journalism-specific benchmarks TOPIC and SIGNATURE that were introduced in Section 3.3. Table 5 reports the results on the TOPIC benchmark. It shows that the difference in accuracy between CamemBERT and the two versions of BERTinchamps lies in a range of 3%. The $\text{BERTinchamps}_{\text{rtbf}}$ model outperforms or matches the performance of CamemBERT on every topic of the benchmark, while $\text{BERTinchamps}_{\text{classic}}$ is outperformed on every topic. For the SIGNATURE benchmark, Table 6 shows that the difference in accuracy between CamemBERT and the two BERTinchamps models is also in a range of 3%. Both BERTinchamps models are slightly less accurate than CamemBERT on 4 authors out of 5. $\text{BERTinchamps}_{\text{rtbf}}$ outperforms $\text{BERTinchamps}_{\text{classic}}$ on every author. An interesting outlier is the author DELPARTURE, for which both BERTinchamps models perform better than CamemBERT. Taken together, these findings indicate that replacing a part of the OSCAR corpus with the RTBF corpus increases its performance on tasks specific to journalism.

5. Discussion

The results of Section 4.1 on the FLUE benchmark provide evidence that the Cramming recipe can be transposed to the French language, even though it was originally designed for English. The crammed BERTinchamps model performs well compared to state-of-the-art, medium-sized LMs on general-purpose tasks, even though it has only been trained on a single NVIDIA A4500 GPU for 24 hours, which amounts to 0.61 exaFLOPs. In comparison, CamemBERT has been trained on 256 NVIDIA V100 GPUs for 24 hours, for a total of 110 exaFLOPs, while FlauBERT has been trained on 32 NVIDIA V100 GPUs for 410 hours, for a total of 230 exaFLOPs. This implies that the pre-training of a BERTinchamps model only requires 0.55% (resp. 0.26%) of the computational resources that are required for the pre-training of CamemBERT (resp. FlauBERT). This opens the path to the in-house pre-training of LMs directly on confidential data, using computational resources that are within the reach of many institutions.

Section 4.2 indicates that the BERTinchamps model is promising for medical tasks, competing with the specialized DrBERT LM on an experimental setup derived from the QUAEROFrenchMed benchmark. Also note that DrBERT was trained on the biomedical corpus NACHOS with 128 NVIDIA V100 GPUs for 20 hours, for a total of 44 exaFLOPs. Therefore, the training of BERTinchamps only requires 1.38% of the training of DrBERT. This raises the interesting research question of whether the performance of BERTinchamps can be further improved by pre-training it on the same corpus as DrBERT. This will be explored in future work, thanks to the recent announcement that the NACHOS dataset will be made available for academic research only.

Finally, the results of Section 4.3 show that pre-training a BERTinchamps model from scratch by replacing a part of the OSCAR dataset by the RTBF corpus does not degrade the performance with respect to a BERTinchamps model that is fully trained on the OSCAR dataset. A slight performance increase is even observed. In summary, both the "classic version" and the "RTBF version" of the BERTinchamps model share the same range of accuracy compared to CamemBERT on the general-purpose FLUE benchmark, as well as on the journalism-specific TOPIC and SIGNATURE benchmarks. This suggests that crammed BERTinchamps models could be trained on private datasets while maintaining good performance compared to state-of-the-art models that were pre-trained on public datasets, which will be explored in future work.

6. Conclusion

This paper contains three significant findings. Firstly, the Cramming recipe that was originally designed for English can be applied to the French language, with comparable effectiveness, resulting in the BERTinchamps model. This finding also suggests that Cramming is likely to be useful in other languages. Secondly, BERTinchamps can be fine-tuned on tasks related to the French medical language, achieving performance comparable to state-of-the-art models on the QUAEROFrenchMed benchmark. Further improvements in BERTinchamps models are likely achievable by leveraging a medical dataset of significant size during their pre-training. Taken together, these two first findings suggest that the self-hosted training of medium-sized LMs from scratch could possibly be within the reach of French-speaking institutions handling sensitive data, which includes hospitals. Indeed, by accelerating the training of LMs by two orders of magnitude, the Cramming recipe has the potential to strongly reduce the cost and complexity of the infrastructure to pre-train LMs. This could notably allow the creation of specialized, in-house medium-sized LMs that are tailored to the very specific needs

of the institutions where they are deployed. As a last finding, the BERTinchamps model can be trained by replacing a portion of the OSCAR training dataset by a domain-specific dataset, without compromising its performance on both general-purpose and domain-specific benchmarks, even improving it marginally.

Future work will involve the pre-training of crammed BERTinchamps models using biomedical data, leveraging the NACHOS dataset (Labrak et al. 2023) which will help determining whether the addition of in-house biomedical data can enhance performance in medical use cases. The NACHOS dataset, previously used to train DrBERT in a private setup, has indeed very recently become available upon request exclusively for academic research. At the time of writing, NACHOS stands as the sole large-scale corpus of medical texts in French that is available for research purposes. The next step would be to explore the pre-training of a BERTinchamps model that is specialized for the population of a hospital, by incorporating patient data from the local electronic health records alongside the OSCAR and NACHOS datasets. This pre-trained model could in turn be fine-tuned to help with real-world clinical tasks such as the automated codification of the clinical documents managed by the hospital. Another promising research path will consist in leveraging federated learning for the collaborative training of one crammed BERTinchamps model that is shared by a coalition of hospitals, while avoiding the communication of protected health information.

Reproducibility Statement

The pre-trained models $\text{BERTinchamps}_{\text{classic}}$ and $\text{BERTinchamps}_{\text{rtbf}}$, together with their pretraining corpora, are available at: `https://doi.org/10.14428/DVN/0OJU5N`. The source code of BERTinchamps is released at: `https://github.com/amauryfierens/BERTinchamps`.

References

Alsentzer, Emily, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Anil, Rohan et al. 2023. PaLM 2 technical report, May. arXiv:2305.10403.

Bai, Haoli, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2021. BinaryBERT: Pushing the limit of BERT quantization. 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 4334–4348, Online, August. Association for Computational Linguistics.

Bassignana, Elisa, Dominique Brunato, Marco Polignano, and Alan Ramponi. 2023. Preface to the seventh workshop on natural language for artificial intelligence (NL4AI). In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023)*, Rome, Italy, November. CEUR Workshop Proceedings.

Biswas, Som S. 2023. Role of ChatGPT in public health. *Annals of Biomedical Engineering*, 51(5):868–869, May.

Bondarenko, Yelysei, Markus Nagel, and Tijmen Blankevoort. 2021. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of EMNLP 2021 (The 2021 Conference on Empirical Methods in Natural Language Processing)*, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics:*

EMNLP 2020 (The 2020 Conference on Empirical Methods in Natural Language Processing), pages 2898–2904, Online, November. Association for Computational Linguistics.
Cheng, Sophia. 2024. When journalism meets ai: Risk or opportunity? *Digital Government: Research and Practice*, June. Just Accepted.
Courbariaux, Matthieu, Yoshua Bengio, and Jean-Pierre David. 2015. Binaryconnect: training deep neural networks with binary weights during propagations. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3123–3131, Cambridge, MA, USA, December. MIT Press.
Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, December. Curran Associates Inc.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June. Association for Computational Linguistics.
Dey, Nolan, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-GPT: Open compute-optimal language models trained on the cerebras Wafer-Scale cluster, April. arXiv:2304.03208.
Douka, Stella, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In Nikolaos Aletras, Ion Androutsopoulos, Leslie Barrett, Catalina Goanta, and Daniel Preotiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
Escouflaire, Louis, Jérémie Bogaert, Antonin Descampe, and Cédrick Fairon. 2023. The RTBF corpus: A dataset of 750,000 Belgian French news articles published between 2008 and 2021. In *Actes des 11èmes Journées Internationales de la Linguistique de Corpus*, pages 155–159, Grenoble, France, July. International Conference on Corpus Linguistics (JLC).
Fierens, Amaury and Sébastien Jodogne. 2023. BERTinchamps: Cost-effective training of large language models for medical tasks in French. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023)*, Rome, Italy, November. CEUR Workshop Proceedings.
Fiesler, Emile, Amar Choudry, and H. John Caulfield. 1990. Weight discretization paradigm for optical neural networks. In Hartmut Bartelt, editor, *Optical Interconnections and Networks*, volume 1281, The Hague, Netherlands, March. International Society for Optics and Photonics, SPIE.
Geiping, Jonas and Tom Goldstein. 2023. Cramming: training a language model on a single gpu in one day. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Hawaii, United States of America, July. JMLR.org.
Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems (NIPS 2014): Deep Learning and Representation Learning Workshop*, Montreal, Canada, March.
Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May. Number: 1 Publisher: Nature Publishing Group.
Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models, January. arXiv:2001.08361.
Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744, January.
Kingma, Diederik and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May.
Labrak, Yanis, 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 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada, July. Association for Computational Linguistics.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September.

Li, Jianning, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. Chatgpt in healthcare: A taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 245:108013, March.

Li, Tianda, Yassir El Mesbahi, Ivan Kobyzev, Ahmad Rashid, Atif Mahmud, Nithin Anchuri, Habib Hajimolahoseini, Yang Liu, and Mehdi Rezagholizadeh. 2021. A short study on compressing decoder-based language models. In *Thirty-Fifth Annual Conference on Neural Information Processing Systems (Neurips 2021), Efficient Natural Language and Speech Processing (ENLSP) Workshop*, Online, October.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach, July. arXiv:1907.11692.

Loshchilov, Ilya and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, United States of America, May.

Mahmoudi, Sid Ali, Charles Condevaux, Bruno Mathis, Guillaume Zambrano, and Stéphane Mussard. 2022. NER sur décisions judiciaires françaises : CamemBERT judiciaire ou méthode ensembliste ? In *Extraction et Gestion des connaissances EGC'2022*, Blois, France, January.

Martin, Louis, 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, July. Association for Computational Linguistics.

Névéol, Aurelie, Cyril Grouin, Xavier Tannier, Thierry Hamon, Liadh Kelly, Lorraine Goeuriot, and Pierre Zweigenbaum. 2015. CLEF eHealth evaluation lab 2015 task 1b: Clinical named entity recognition. In *Sixth Conference and Labs of the Evaluation Forum (CLEF 2015) Working notes.*, Toulouse, France, September.

Névéol, Aurélie, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proceedings of BioTxtM 2014 - 4thWorkshop on Building & Evaluating Resources for Health and Biomedical Text Processing at the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May. ELRA.

OpenAI. 2023. GPT-4 technical report, March.

Ortiz Suárez, Pedro Javier, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July. Association for Computational Linguistics.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary 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. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., Vancouver, Canada, December, pages 8024–8035.

Poongodi, T., D. Sumathi, P. Suresh, and Balamurugan Balusamy. 2021. Deep learning techniques for electronic health record (EHR) analysis. In Akash Kumar Bhoi, Pradeep Kumar Mallick, Chuan-Ming Liu, and Valentina E. Balas, editors, *Bio-inspired Neurocomputing*, Studies in Computational Intelligence. Springer, Singapore, pages 73–103.

Pronovost, Peter J., Melissa D. Cole, and Robert M. Hughes. 2022. Remote patient monitoring during COVID-19: An unexpected patient safety benefit. *Journal of American Medical Association (JAMA)*, 327(12):1125–1126, March.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, July.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, February. arXiv:1910.01108.

Scao, Teven Le et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model, March. arXiv:2211.05100.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto, Japan, March. IEEE.

Touchent, Rian, Laurent Romary, and Eric De La Clergerie. 2023. CamemBERT-bio: Un modèle de langue français savoureux et meilleur pour la santé. In Christophe Servan and Anne Vilnat, editors, *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1: travaux de recherche originaux – articles longs*, pages 323–334, Paris, France, June. ATALA.

Touvron, Hugo, 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, February. arXiv:2302.13971.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Los Angeles, United State of America, December. Curran Associates, Inc.

Wang, Yixu, Wenpin Qian, Hong Zhou, Jianfeng Chen, and Kai Tan. 2023. Exploring new frontiers of deep learning in legal practice: A case study of large language models. *International Journal of Computer Science and Information Technology*, 1(1):131–138, December. Number: 1.

Watzlaf, Valerie J.M., Jennifer Hornung Garvin, Sohrab Moeini, and Patricia Anania-Firouzan. 2007. The effectiveness of ICD-10-CM in capturing public health diseases. *Perspectives in Health Information Management / AHIMA, American Health Information Management Association*, 4:6, June.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, October. arXiv:1609.08144.

Zafrir, Ofir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8bit BERT. In *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, pages 36–39, Vancouver, Canada, December. IEEE.

Zhou, Binggui, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2022. Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*, 17:1–17, September.

Zhou, Lingling, Cheng Cheng, Dong Ou, and Hao Huang. 2020. Construction of a semi-automatic ICD-10 coding system. *BMC medical informatics and decision making*, 20:1–12, April.