

CamemBERT 2.0: A Smarter French Language Model Aged to Perfection

Wissam Antoun Francis Kulumba Rian Touchent
Éric de la Clergerie Benoît Sagot Djamé Seddah
Inria, Paris, France
{firstname,lastname}@inria.fr

Abstract

French language models, such as CamemBERT, have been widely adopted across industries for natural language processing (NLP) tasks, with models like CamemBERT seeing over 4 million downloads per month. However, these models face challenges due to temporal concept drift, where outdated training data leads to a decline in performance, especially when encountering new topics and terminology. This issue emphasizes the need for updated models that reflect current linguistic trends. In this paper, we introduce two new versions of the CamemBERT base model—CamemBERTav2 and CamemBERTv2—designed to address these challenges. CamemBERTav2 is based on the DeBERTav3 architecture and makes use of the Replaced Token Detection (RTD) objective for better contextual understanding, while CamemBERTv2 is built on RoBERTa, which uses the Masked Language Modeling (MLM) objective. Both models are trained on a significantly larger and more recent dataset with longer context length and an updated tokenizer that enhances tokenization performance for French. We evaluate the performance of these models on both general-domain NLP tasks and domain-specific applications, such as medical field tasks, demonstrating their versatility and effectiveness across a range of use cases. Our results show that these updated models vastly outperform their predecessors, making them valuable tools for modern NLP systems. All our new models, as well as intermediate checkpoints, are made openly available on Huggingface https://huggingface.co/almanach?search_models=camembert+v2.

1 Introduction

In recent years, French language models such as CamemBERT (Martin et al., 2020) have become integral to businesses leveraging natural language processing (NLP) to boost productivity

and efficiency. Since its release, CamemBERT has gained widespread adoption, receiving over 4 million downloads each month, and continues to be actively used by the NLP community. A notable example is ENEDIS, which used CamemBERT to automate the dispatch of 100,000 customer requests per day across 1,500 operators, significantly reducing manual workload and achieving a return on investment of approximately €3M per year (Gemignani et al., 2023; Akani et al., 2023).

However, a significant challenge that models like CamemBERT, developed in 2019, face is temporal concept drift (Loureiro et al., 2022; Agarwal and Nenkova, 2022; Jin et al., 2022). This phenomenon occurs when the data a model was originally trained on becomes outdated, leading to a decline in performance as new topics, events, and terminology emerge. For instance, when CamemBERT was trained, discussions around COVID-19, public health restrictions, and associated changes in language usage were not present in the training data. As a result, models like CamemBERT, which have not been updated, struggle to understand or generate accurate responses to these newer concepts. Temporal drift impacts their ability to remain relevant in evolving real-world applications, highlighting the need for continuous updates to keep the models aligned with current linguistic and contextual trends.

Given these challenges, it is essential to develop and deploy updated encoder models that can better serve modern NLP applications. In response, we aim to provide state-of-the-art models and fine-tuned versions for a range of common NLP tasks, including Named Entity Recognition (NER), Question Answering (QA), Natural Language Inference (NLI), and Part-of-Speech (POS) tagging. These updated models will play a major role in creating fast, efficient, and reliable AI systems.

In this paper, we present two new versions of the CamemBERT base model: CamemBERTav2

and CamemBERTv2. CamemBERTav2 is built on the DeBERTaV3 (He et al., 2021a) architecture as an update to the CamemBERTa model (Antoun et al., 2023), using the Replaced Token Detection (RTD) training objective for enhanced context and positional representation, while CamemBERTv2 is based on the RoBERTa (Liu, 2019) architecture, trained using the Masked Language Modeling (MLM) objective. Both models benefit from training on a much larger and more recent dataset, coupled with an updated tokenizer designed to better capture the nuances of the French language, and to support modern token requirements by adding new lines, tabulation and emojis to the vocabulary. This ensures improved tokenization performance across various NLP tasks.

To evaluate the performance of these new models, we conducted extensive tests on both general-domain NLP tasks and domain-specific applications, such as those in the medical field. This dual approach demonstrates the versatility of our models, highlighting their capacity to excel in diverse use cases, including highly specialized areas where domain-specific knowledge is necessary.

The contributions of this paper are as follows:

- We present CamemBERTav2 and CamemBERTv2 and trained on a larger, up-to-date dataset with an enhanced tokenizer to better capture the complexities of the French language.
- We evaluate the models on both general-domain NLP tasks and domain-specific applications, particularly in the medical field, to demonstrate their robustness and versatility.
- We are releasing all model artifacts, including intermediate checkpoints and fine-tuned models, enabling the community to further experiment, fine-tune, and deploy these models across various applications.¹

These contributions aim to advance French language modeling and provide the community with state-of-the-art tools for diverse NLP tasks.

2 Related Works

Pre-trained French Language Models French language models have predominantly been

trained using either Masked Language Modeling (MLM) or Causal Language Modeling (CLM) techniques. Among the leading French models, CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) have gained considerable traction, both relying on the MLM pre-training method. Other noteworthy models include FrALBERT (Cattan et al., 2021), a French adaptation of ALBERT (Lan et al., 2020), and LePetit (Micheli et al., 2020), a scaled-down version of CamemBERT. In addition, D’AlemBERT (Gabay et al., 2022), derived from RoBERTa (Liu et al., 2020), is tailored for Early Modern French. BARThez (Kamal Eddine et al., 2021) is a sequence-to-sequence model built on BART’s architecture (Lewis et al., 2020), while PAGnol (Launay et al., 2022) and Cedille (Müller and Laurent, 2022) represent models trained with CLM.

Recently, CamemBERTa, a French DeBERTa model based on the DeBERTaV3 architecture, has been introduced. CamemBERTa demonstrates superior performance on various French NLP tasks compared to traditional BERT-based models, despite using significantly fewer training tokens (Antoun et al., 2023). Additionally, CamemBERT-bio, a specialized model fine-tuned for French biomedical data, has shown substantial improvements in named entity recognition tasks within the biomedical domain, addressing the limitations of general-purpose models like CamemBERT when applied to specialized texts (Touchent and de la Clergerie, 2024). DrBERT further extends this focus by pre-training on both public and private medical data, offering specialized performance for the French biomedical field (Labrak et al., 2023). Finally, CharacterBERT-French, a character-based variant of BERT, offers robustness to noise by utilizing character embeddings instead of subword-based vocabularies. This model has shown promise, particularly in handling noisy experimental data (Riabi et al., 2021).

3 CamemBERT 2.0

In this paper, we introduce two updated versions of the CamemBERT base model, named CamemBERTav2 and CamemBERTv2, developed to improve upon the original CamemBERTa (Antoun et al., 2023) and CamemBERT (Martin et al., 2020) models

¹https://huggingface.co/almanach?search_models=camembert+v2

respectively. CamemBERTav2 is built on the DeBERTaV3 (He et al., 2021a,b) architecture and the Replaced Token Detection (RTD) (Clark et al., 2020) training objective, leveraging its improved attention mechanism for better context and positional representation. CamemBERTv2, on the other hand, is based on RoBERTa (Liu, 2019), using the Masked Language Modeling (MLM) objective for training, and is meant to be a drop-in replacement in task where computing having the pseudo-language modeling probability is needed. The models were trained on a much larger and up-to-date dataset, accompanied by an updated tokenizer that better captures the linguistic complexities of the French language, ensuring improved tokenization performance for various downstream tasks.

3.1 Pre-Training Dataset

Our new pre-training dataset is sourced mainly from the French subset of the CulturaX corpus (Nguyen et al., 2023). CulturaX is a multilingual dataset that combines mC4 (Xue et al., 2021) and four OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021, 2022) snapshots.² The documents are then deduplicated on the document level and filtered using language filters, URL block lists, and a comprehensive set of metric-based filters (e.g. stopword ratio, perplexity score, word repetition ratio...). In addition, we make use of the French section of Wikipedia³, and French scientific papers and theses from the HALvesting corpus (Kulumba et al., 2024). In total, we gather 265B tokens from Culturax, 4.7B tokens from HALvesting, and 0.5B tokens from Wikipedia. During training, we upsample Wikipedia 10 times, and hence our final pre-training dataset has 275B tokens, compared to 32B which were used during the original CamemBERT and CamemBERTa training.

3.2 Tokenizer

A key improvement in the CamemBERTv2 models is the development of an updated tokenizer. The primary goal was to improve tokenization efficiency by addressing the limitations of the previous version. This includes the introduction of newline and tab characters, as well as support for emojis, which are normalized by

removing zero-width joiner characters and splitting emoji sequences into individual tokens. To improve the handling of numerical data, we opted to split numbers into a maximum of two-digit tokens, which we hypothesize will enhance the model’s ability to process dates and perform simple arithmetic tasks—functions more commonly utilized in encoder models than in generative ones. Additionally, French and English elisions (e.g., l’, lorsqu’) are now treated as single tokens, including the apostrophe. We adopted the WordPiece tokenization algorithm (Devlin et al., 2019), which allows for flexible vocabulary adjustments and the easy addition of new tokens. The vocabulary size was set to 32,768, with around 400 tokens reserved for future expansion to maintain a multiple of 8. We finally train the tokenizer on a subsample of our pre-training dataset that include a subsample of CulturaX and full French Wikipedia and HAL.

3.3 Pre-Training Methodology

The pre-training process for both CamemBERTv2 and CamemBERTav2 models was done in two stages. Initially, both models were trained with a sequence length of 512 tokens, which allowed for faster convergence during the early stages of training. In the second stage, the models were further pre-trained using a sequence length of 1024 tokens to fully capture long-range dependencies and improve performance on tasks requiring extensive context. To create a pre-training dataset for the long sequence training, we further filter our pretraining corpus to have only long documents, while also including short sequences with a 5% chance to ensure the model retains the ability to correctly handle shorter sequences.

Additionally, it was shown by (Antoun et al., 2023) that models trained with MLM require multiple epochs of pre-training to achieve optimal accuracy, due to the Masked Language modeling objective only being able to propagate the loss from the masked tokens. Hence, we train CamemBERTv2 for three epochs over our dataset. We set the token masking rate to 40%, which was found to be optimal by Wettig et al. (2023). In contrast, CamemBERTav2, being based on the more sample-efficient DeBERTaV3 pertaining methodology of replaced-token detection, reaches peak performance after just one epoch, making it significantly more efficient in terms of training time and computational resources. Details about

²Culturax contains the following OSCAR corpora 20.19, 21.09, 22.01, and 23.01, and the version 3.1.0 of mC4

³We use the April 2024 dump

MODEL	GSD		RHAPSODIE		SEQUOIA		FSMB		FTB-NER
	UPOS	LAS	UPOS	LAS	UPOS	LAS	UPOS	LAS	F1
CamemBERT	98.57 \pm 0.07	94.35 \pm 0.15	97.62 \pm 0.08	84.29 \pm 0.56	99.35 \pm 0.09	94.78 \pm 0.12	94.80 \pm 0.16	81.34 \pm 0.63	89.97 \pm 0.50
CamemBERTa	98.55 \pm 0.05	94.38 \pm 0.15	97.52 \pm 0.14	84.23 \pm 0.08	99.44 \pm 0.07	94.85 \pm 0.14	94.80 \pm 0.09	80.74 \pm 0.25	90.33 \pm 0.54
CamemBERTv2	98.60 \pm 0.05	94.18 \pm 0.12	97.62 \pm 0.10	84.09 \pm 0.31	99.37 \pm 0.04	94.80 \pm 0.14	95.05 \pm 0.18	81.49 \pm 0.38	91.99 \pm 0.96
CamemBERTav2	98.54 \pm 0.03	94.35 \pm 0.20	97.70 \pm 0.21	84.30 \pm 0.27	99.42 \pm 0.05	94.61 \pm 0.25	95.19 \pm 0.11	81.32 \pm 0.29	93.40\pm0.62

Table 1: **POS tagging, dependency parsing and NER** results on the test sets of our French datasets. *UPOS* (Universal Part-of-Speech) refers here to POS tagging accuracy, and *LAS* measures the overall accuracy of labeled dependencies in a parsed sentence.

pre-training hyperparameters are available in Table 6

4 Experiments and Results

4.1 Downstream Evaluation

General Domain To evaluate our models we consider a range of French downstream tasks and datasets, including Question Answering (QA) using FQuAD 1.0 (d’Hoffschmidt et al., 2020), Part-Of-Speech (POS) tagging and Dependency Parsing on GSD (McDonald et al., 2013), Rhapsodie (Lacheret et al., 2014), Sequoia (Candito and Seddah, 2012; Candito et al., 2014) in their UD v2.2 formats, and the French Social Media Bank (Seddah et al., 2012). We also assess Named Entity Recognition (NER) on the 2008 FTB version (Abeillé et al., 2000; Candito and Crabbé, 2009) with NER annotations by Sagot et al. (2012). To assess text classification capabilities we evaluate our models on the FLUE benchmark (Le et al., 2019). We re-used the same splits from Antoun et al. (2023), and performed hyper-parameter tuning on all models and datasets with 5 seed variations, except the dependency parsing and part-of-speech tasks where we only validate over 5 seeds using the same sets of parameters.

Domain Specific To assess the models on domain-specific tasks, we include the French subset of the pseudoanonymized dataset for radicalization detection with NER annotations (Riabi et al., 2024) which we refer to as Counter-NER. For biomedical-domain datasets, we evaluate five distinct tasks: EMEA, MEDLINE, CAS1, CAS2, and E3C. EMEA and MEDLINE are part of the QUAERO corpus (Névéol et al., 2014), where EMEA consists of drug leaflets and MEDLINE includes scientific article titles, both annotated with 10 semantic groups from the UMLS. CAS1 and CAS2 are based on the CAS corpus (Grouin et al., 2019), focusing on

pathology and symptoms in the first subtask, while the second subtask involves extracting additional clinical information such as anatomy and treatment. Finally, E3C (Magnini et al., 2020) focuses on clinical cases from scientific articles, using fully annotated texts to identify clinical entities. For consistency, we adopt the dataset splits and hyper-parameters proposed by Touchent and de la Clergerie (2024) for comparison with his model.

4.2 General Domain Results

For general domain tasks, the results show clear performance trends between models:

POS Tagging and Dependency Parsing: As shown in Table 1, all models performed well on Universal POS (UPOS) tagging and dependency parsing, where the updated CamemBERTv2 and CamemBERTav2 model maintaining the previous models’ scores. These results indicate a possible saturation in the benchmark scores for current encoder-based transformer models.

Named Entity Recognition (NER): In general domain NER, evaluated on the FTB dataset, CamemBERTav2 outperformed all other models with an F1 score 93.4% showing a significant improvement over the baseline CamemBERT model (89.97%), while also improving over the MLM-trained CamemBERTv2 model.

Question Answering (QA): For the FQuAD 1.0 dataset (Table 2), CamemBERTav2 achieved the highest F1 score (83.04%) and exact match (EM) score (64.29%), outperforming the other models by a significant margin. The performance gap between CamemBERTv2 and CamemBERTav2 (80.39% vs 83.04%) suggests that the latter’s enhanced pre-training loss and architecture yielded more robust representations for machine comprehension tasks in French.

Model	F1	EM
CamemBERT	80.98 \pm 0.48	62.51 \pm 0.54
CamemBERTa	81.15 \pm 0.38	62.01 \pm 0.45
CamemBERTv2	80.39 \pm 0.36	61.35 \pm 0.39
CamemBERTav2	83.04 \pm 0.19	64.29 \pm 0.31

Table 2: Question Answering results on FQuAD 1.0.

Text Classification: Table 3 presents text classification results across the CLS, PAWS-X, and XNLI tasks from the FLUE benchmark. CamemBERTav2 consistently outperformed other models, achieving top scores in all tasks, with the highest accuracy on the CLS task (95.63%), PAWS-X (93.06%), and XNLI (84.82%). The massive increase in CamemBERTav2’s XNLI scores compared to the previous CamemBERTa model shows that small transformer-based models, that use the sample-efficient RTD objective, can still benefit from increasing the unique token count during pretraining.

Model	CLS	PAWS-X	XNLI
CamemBERT	94.62 \pm 0.04	91.36 \pm 0.38	81.95 \pm 0.51
CamemBERTa	94.92 \pm 0.13	91.67 \pm 0.17	82.00 \pm 0.17
CamemBERTv2	95.07 \pm 0.11	92.00 \pm 0.24	81.75 \pm 0.62
CamemBERTav2	95.63 \pm 0.16	93.06 \pm 0.45	84.82 \pm 0.54

Table 3: Text classification results (Accuracy) on the FLUE benchmark.

4.3 Domain Specific Results

In the evaluation of domain-specific tasks, Table 4, particularly in the medical fields, both CamemBERTv2 and CamemBERTav2 exhibited strong performance. On medical NER tasks, the new models were able to achieve results comparable to domain-specific models, namely CamemBERT-bio, showcasing their ability to handle specialized terminologies and complex entity recognition. Notably, CamemBERTv2 and CamemBERTav2 significantly outperformed their predecessors across all tasks, largely due to the inclusion of scientific and medical articles in their updated pre-training datasets.

In the radicalization NER task, which involves identifying sensitive and domain-specific entities, both models demonstrated large improvements. CamemBERTav2 surpassed the original CamemBERT model by 2 percentage points, while CamemBERTv2 exceeded CamemBERT by over 3 points, further highlighting the

enhancements made in these newer versions. These gains showcase the models’ ability to generalize to challenging, niche domains with specialized vocabularies.

Model	Medical-NER	Counter-NER
CamemBERT	70.96 \pm 0.13	84.18 \pm 1.23
CamemBERTa	71.86 \pm 0.11	87.37 \pm 0.73
CamemBERT-bio	73.96 \pm 0.12	-
CamemBERTv2	72.77 \pm 0.11	87.46 \pm 0.62
CamemBERTav2	73.98 \pm 0.11	89.53 \pm 0.73

Table 4: Summary of NER F1 scores on the domain-specific downstream tasks. Full scores are available in Table 5.

4.4 Discussion

The results from our experiments clearly demonstrate the significant advancements that CamemBERTv2 and CamemBERTav2 bring to French NLP tasks, both in general and domain-specific contexts. In the general domain tasks, CamemBERTav2 consistently outperformed its predecessors, showcasing the effectiveness of the DeBERTaV3 architecture and the RTD objective in handling both contextual and positional representations. The improvements seen in tasks such as NER, QA, and text classification are particularly noteworthy. For example, in the FQuAD 1.0 dataset, the large performance gap between CamemBERTv2 and CamemBERTav2 illustrates the robustness of the latter in understanding complex queries and extracting relevant answers. The enhanced tokenizer, with its improved handling of French-specific linguistic features and expanded vocabulary, also played a key role in these improvements.

Interestingly, while the models achieved high accuracy in POS tagging and dependency parsing tasks, the marginal gains over the original CamemBERT suggest that transformer-based models may be approaching performance ceilings on these specific benchmarks. This observation indicates that future progress in these areas might require new approaches, such as task-specific architectures or training methodologies, rather than further refinements to existing models.

In domain-specific tasks, the inclusion of scientific and medical articles in the pre-training dataset allowed both CamemBERTv2 and CamemBERTav2 to achieve strong results across

specialized fields. Their ability to generalize to biomedical NER tasks, where they performed comparably to models specifically designed for the medical domain, shows the versatility of our updated models. The sizable improvements in the radicalization NER task also reflect the enhanced knowledge embedded in the new models, which is essential for identifying sensitive and rare entities within challenging domains.

These results affirm the value of continual model updates, particularly in addressing the issue of temporal concept drift. As language evolves and new terminologies emerge, updating models with more recent datasets and architectures becomes crucial for maintaining their relevance and utility in real-world applications. Our decision to update the tokenizer to better handle modern language elements like emojis and numerical data further reinforces this point, allowing the models to stay aligned with contemporary communication patterns.

5 Conclusion

In conclusion, the development of CamemBERTv2 and CamemBERTav2 marks a significant advancement in French language modeling, demonstrating improved performance across a variety of general and domain-specific NLP tasks. By leveraging larger and more recent datasets, alongside an updated tokenizer, these models have shown enhanced versatility and robustness, particularly in tasks like NER, QA, and text classification. However, the marginal improvements seen in certain tasks like POS tagging and dependency parsing suggest that these benchmarks may be nearing saturation for current transformer-based models.

Looking ahead, future work should not only focus on refining model architectures and training objectives but also prioritize updating datasets. Temporal concept drift is not solely a model issue—it is also a dataset issue. Many benchmarks currently in use do not reflect the latest linguistic distributions, which can exacerbate the performance gap between models trained on outdated versus modern data. Ensuring that datasets are regularly updated to include contemporary topics, terminologies, and language use is essential for keeping models relevant and maximizing their real-world applicability. Such efforts will ensure that both models and

benchmarks evolve together, addressing temporal drift more effectively and pushing the boundaries of what these systems can achieve.

Acknowledgements

This work was partly funded by Benoît Sagot’s chair in the PRAIRIE institute funded by the French national research agency (ANR as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001. This work also received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 101021607. The authors are grateful to the OPAL infrastructure from Université Côte d’Azur for providing resources and support. This work was also granted access by GENCI to the HPC resources of IDRIS under the allocation 2024-GC011015610. Finally, part of this work was funded by the DINUM through the [AIIAance program](#).

We would also like to thank Nathan Godey, and Arij Riabi for the productive discussions.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. [Towards a cleaner document-oriented multilingual crawled corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.
- Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. [Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus](#). *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9)* 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. [Building a treebank for French](#). In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, Athens, Greece. European Language Resources Association (ELRA).
- Oshin Agarwal and Ani Nenkova. 2022. [Temporal effects on pre-trained models for language processing tasks](#). *Transactions of the Association for Computational Linguistics*, 10:904–921.
- E. Akani, R. Gemignani, and R. Abrougui. 2023. [Enebert: a state-of-the-art language model trained on a corpus of texts generated from the set of dso activities](#). In *27th International Conference on*

- Electricity Distribution (CIRED 2023)*, volume 2023, pages 2903–2907.
- Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2023. Data-efficient french language modeling with camemberta. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.
- Marie Candito and Benoît Crabbé. 2009. [Improving generative statistical parsing with semi-supervised word clustering](#). In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 138–141, Paris, France. Association for Computational Linguistics.
- Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karèn Fort, Djamé Seddah, and Eric De La Clergerie. 2014. Deep syntax annotation of the sequoia french treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Marie Candito and Djamé Seddah. 2012. [Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical \(the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method\) \[in French\]](#). In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.
- Oralie Cattan, Christophe Servan, and Sophie Rosset. 2021. [On the Usability of Transformers-based models for a French Question-Answering task](#). In *Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#). In *ICLR*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. [FQuAD: French Question Answering Dataset](#). *arXiv e-prints*, arXiv:2002.06071.
- Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. [From FreEM to d'AleMBERT: a large corpus and a language model for early Modern French](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.
- R. Gemignani, E. Akani, J.P. Delrieux, and A. Sayouti Souleymane. 2023. [Hape: optimizing customer relation by automatic task distribution using constrained optimization and natural language processing](#). In *27th International Conference on Electricity Distribution (CIRED 2023)*, volume 2023, pages 1764–1768.
- Cyril Grouin, Natalia Grabar, Vincent Claveau, and Thierry Hamon. 2019. [Clinical case reports for NLP](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 273–282, Florence, Italy. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. [Lifelong pretraining: Continually adapting language models to emerging corpora](#). In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 1–16, virtual+Dublin. Association for Computational Linguistics.
- Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. [BARThez: a skilled pretrained French sequence-to-sequence model](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Francis Kulumba, Wissam Antoun, Guillaume Vimont, and Laurent Romary. 2024. [Harvesting textual and structured data from the hal publication repository](#). *Preprint*, arXiv:2407.20595.
- Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. [DrBERT: A robust pre-trained model in French for biomedical and clinical domains](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.
- Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. [Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé](#). In *4e Congrès Mondial de Linguistique Française*, volume 8, pages 2675–2689, Berlin, Germany.

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#). In *International Conference on Learning Representations*.
- Julien Launay, E.I. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. [PAGnol: An extra-large French generative model](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.
- Hang Le, 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 Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. [Flaubert: Unsupervised language model pre-training for french](#). *Preprint*, arXiv:1912.05372.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Ro{bert}a: A robustly optimized {bert} pretraining approach](#).
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. [TimeLMs: Diachronic language models from Twitter](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.
- Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanolli. 2020. [The e3c project: Collection and annotation of a multilingual corpus of clinical cases](#). *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Ryan McDonald, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. [Universal Dependency annotation for multilingual parsing](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.
- Vincent Micheli, Martin d’Hoffschmidt, and François Fleuret. 2020. [On the importance of pre-training data volume for compact language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.
- Martin Müller and Florian Laurent. 2022. [Cedille: A large autoregressive french language model](#). *Preprint*, arXiv:2202.03371.
- Aurélié Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The quæro french medical corpus: a resource for medical entity recognition and normalization. In *Bio text-mining workshop (BioTextM 2014)*, page 7p, Reykjavik, Iceland.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. [Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). *Preprint*, arXiv:2309.09400.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. [Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures](#). *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019*. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.
- Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. [Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks](#). In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.
- Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021. [Can character-based language models improve](#)

downstream task performances in low-resource and noisy language scenarios? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 423–436, Online. Association for Computational Linguistics.

Benoît Sagot, Marion Richard, and Rosa Stern. 2012. [Annotation référentielle du corpus arboré de Paris 7 en entités nommées \(referential named entity annotation of the Paris 7 French TreeBank\) \[in French\]](#). In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 535–542, Grenoble, France. ATALA/AFCP.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Moulleron, and Vanessa Combet. 2012. [The French Social Media Bank: a treebank of noisy user generated content](#). In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Rian Touchent and Éric de la Clergerie. 2024. [CamemBERT-bio: Leveraging continual pre-training for cost-effective models on French biomedical data](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2692–2701, Torino, Italia. ELRA and ICCL.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. [Should you mask 15% in masked language modeling?](#) In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Full Results

A.1 Full Domain Specific NER Results

Dataset	Model	F1
CAS1	CamemBERT	70.72±1.47
	CamemBERTa	71.96±1.38
	Dr-BERT	62.76±1.55
	CamemBERT-Bio	72.28±1.46
	CamemBERTv2	71.18±1.62
	CamemBERTav2	72.87 ±2.29
CAS2	CamemBERT	78.43±1.78
	CamemBERTa	79.06±0.68
	Dr-BERT	76.43±0.49
	CamemBERT-Bio	82.50 ±0.56
	CamemBERTv2	81.87±0.58
	CamemBERTav2	81.85±0.49
E3C	CamemBERT	67.01±2.13
	CamemBERTa	67.01±1.85
	Dr-BERT	56.99±2.40
	CamemBERT-Bio	69.87±1.21
	CamemBERTv2	69.27±0.90
	CamemBERTav2	70.12 ±0.87
EMEA	CamemBERT	73.53±2.04
	CamemBERTa	75.99±0.51
	Dr-BERT	71.33±0.84
	CamemBERT-Bio	76.96±2.00
	CamemBERTv2	76.30±1.00
	CamemBERTav2	77.28 ±0.57
MEDLINE	CamemBERT	65.11±0.56
	CamemBERTa	65.33±0.30
	Dr-BERT	58.90±0.51
	CamemBERT-Bio	68.21 ±0.91
	CamemBERTv2	65.26±0.33
	CamemBERTav2	67.77±0.44
Counter-NER	CamemBERT	84.18±1.23
	CamemBERTa	87.37±0.73
	CamemBERTv2	87.46±0.62
	CamemBERTav2	89.53 ±0.73

Table 5: NER F1 scores on the domain-specific downstream tasks.

B Hyper-parameters

B.1 Pre-training Hyper-parameters

Hyper-parameter	CamemBERTv2 _{base}	CamemBERTv2 _{base}
Number of Layers	12	12
Hidden size	768	768
Generator Hidden size	256	-
FNN inner Hidden size	3072	3072
Attention Heads	12	12
Attention Head size	64	64
Dropout	0.1	0.1
Warmup Steps (p1/p2)	10k/1k	10k/1k
Learning Rates (p1/p2)	7e-4/3e-4	7e-4/3e-4
End Learning Rates (p1/p2)	1e-5	1e-5
Batch Size	8k	8k
Weight Decay	0.01	0.01
Max Steps (p1/p2)	91k/17k	273k/17k
Learning Rate Decay	Polynomial p=0.5	Polynomial p=0.5
Adam ϵ	1e-6	1e-6
Adam β_1	0.878	0.878
Adam β_2	0.974	0.974
Gradient Clipping	1.0	1.0
Masking Probability	20%	40%
Seq. Length (p1/p2)	512/1024	512/1024
Precision	BF16	BF16

Table 6: Hyper-parameters for pre-training CamemBERTa and CamemBERT 2.0.

B.2 Fine-Tuning Hyper-parameters

Task	Learning Rate	LR Sch.	Epochs	Max Len.	Batch Size	Warmup
FQuAD	{3, 5, 7}e-5	cosine	6	1024	{32,64}	{0,0.1}
CLS	{3, 5, 7}e-5	cosine linear	6	1024	{32,64}	0
PAWS-X	{3, 5, 7}e-5	cosine linear	6	148	{32,64}	0
FTB NER	{3, 5, 7}e-5	cosine linear	8	192	{16,32}	{0,0.1}
XNLI	{3, 5, 7}e-5	cosine	10	160	32	0.1
POS	3e-05	linear	64	1024	8	100 steps
Dep. Pars.	3e-05	linear	64	1024	8	100 steps
Counter-NER	{3, 5, 7}e-5	cosine linear	8	512	{16,32}	{0,0.1}
Med-NER	5e-5	linear	3	20	8	0.224

Table 7: Hyperparameter Search During Fine-tuning of CamemBERTv2. All models were trained with FP32