A Diversity Diet for a Healthier Model: A Case Study of French ModernBERT

Louis Estève[1], **Christophe Servan**[1,2], **Thomas Lavergne**[1], **Agata Savary**[1]

[1]LISN, Paris-Saclay University, CNRS, France; first.last@lisn.fr
[2]AMIAD, Pôle Recherche, France; first.last@polytechnique.fr

Relevant UniDive working groups: WG4

1 Introduction

Natural Language Processing (NLP) has seen over the years a substantial increase in dataset size. Datasets ten years ago, e.g. Universal Dependencies v1.0 (Nivre et al., 2016), had millions of tokens, unlike some modern datasets comprising over ten trillion tokens, such as FineWeb (Penedo et al., 2024) and HPLT (de Gibert et al., 2024).[1] Scalable architectures such as transformers (Vaswani et al., 2017) benefited from this growth as a consequence of neural scaling laws. These laws state that, overall, more weights, training computation, and training data improves model performance (Kaplan et al., 2020). This however comes with an exorbitant cost: managing vast datasets, and using them to train Large Language Models (LLMs), is non-trivial, excessively expensive, and detrimental to the environment (Strubell et al., 2019).

It might be that LLMs benefit from larger datasets due to increased vocabulary diversity. However, naturally-occurring redundancy in data implies that randomly appending new data brings diminishing increases in diversity. In other words, transformers may not be as data-hungry, if we give them diverse pre-training data. Our research question is then: can diversity-driven sampling prevent the overgrowth of pre-training datasets, while preserving or increasing performance? To address this question, we set the following hypotheses:

- H1 A small dataset, with lexical diversity substantially higher than at random, can be sampled efficiently from a large dataset.
- H2 A model trained on a small lexically diverse dataset can be competitive to or outperform one trained on a very large dataset.

This abstract summarizes our paper recently accepted for publication (Estève et al., 2026).

[1]A million is 10^6, while a trillion is 10^{12}.

Dataset	Words	H (↑)
TOPLINE-2400M	2 379M	7.55
RANDOM-400M	401M	7.59
DIVERSE-400M	397M	8.53
RANDOM-240M	242M	7.60
DIVERSE-230M	230M	8.45
RANDOM-150M	150M	7.61
DIVERSE-150M	147M	8.21
BASELINE-100M	105M	7.61

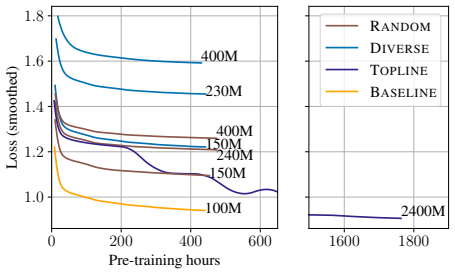

Figure 1: Sampled datasets and matching pre-training processes. Loss is smoothed using a moving average (window size: 2×10^5, series sizes from 1.5M to 6.3M points) applied thrice.

2 Sampling

We test H1 by adapting our previous diversity-driven sampling algorithm (Scholivet et al., 2025). It starts from an initial high quality base corpus of moderate size, and increases it by sampling from an additional corpus, so that diversity of the resulting dataset gradually increases. Shannon-Wiener entropy over word types is used as a diversity measure.

We use this algorithm to create DIVERSE datasets of different sizes, by sampling from a large French corpus of 2,379M tokens, henceforth called TOPLINE-2400M, containing the French part of Wikipedia and from OPUS (Tiedemann, 2012). For each DIVERSE dataset, we create a matching RANDOM dataset of commensurate size (also sampled from TOPLINE-2400M).[2] The pre-training datasets

[2]The sizes are only approximately equal (notably in case of RANDOM-240M vs. DIVERSE-230M), due to creation of

Encoder	PTT	Classification			Sequence labelling		
		PX	XNLI	AMI	WNER	MATIS	MDF
TOPLINE-2400M	1775h	84.7±1.1	75.9±0.6	81.7±2.5	89.8±0.5	92.5±0.7	82.6±0.5
RANDOM-400M	483h	85.9±1.9	75.1±1.7	**83.6±0.7**	90.0±0.4	92.1±1.2	**83.2±0.4**
DIVERSE-400M	483h	85.6±1.5	74.7±0.6	82.6±1.0	90.6±0.9	91.8±1.8	82.0±1.1
RANDOM-240M	483h	82.6±5.2	**76.4±1.0**	82.5±0.8	90.1±0.7	92.8±0.8	82.7±0.3
DIVERSE-230M	483h	**86.3±1.6**	74.9±1.8	82.2±1.5	90.8±0.4	92.8±0.8	82.5±0.2
RANDOM-150M	483h	84.5±1.0	72.3±2.7	82.0±2.0	89.5±0.4	92.5±0.8	82.4±0.3
DIVERSE-150M	483h	85.5±0.7	**75.0±0.9**	82.2±1.9	**90.5±0.8**	93.3±0.3	82.4±0.4
BASELINE-100M	483h	83.2±0.6	74.7±1.4	83.0±0.9	89.5±0.7	92.5±0.3	78.0±4.7

Table 1: Encoder & head fine-tuning. Evaluation on TEST. PTT is pre-training time. PX is PAWS-X (Yang et al., 2019), AMI is Amazon Massive Intent (FitzGerald et al., 2023), WNER is WikiNER (Nothman et al., 2013), MATIS is MultiATIS++ (Upadhyay et al., 2018), MDF is MEDIA (full) (Bonneau-Maynard et al., 2006). Mean ± standard deviation, across five seeds. Bold means $\Delta \geq 1$.

Encoder	PTT	MATIS	MDF
TOPLINE-2400M	1775h	83.2±0.5	58.9±0.5
RANDOM-400M	483h	84.0±0.5	**60.3±0.3**
DIVERSE-400M	483h	84.5±0.3	59.1±0.8
RANDOM-240M	483h	84.4±0.5	59.7±0.6
DIVERSE-230M	483h	**86.4±0.4**	**61.7±0.2**
RANDOM-150M	483h	80.2±0.2	51.9±0.4
DIVERSE-150M	483h	**84.3±0.5**	**61.5±0.4**
BASELINE-100M	483h	77.8±0.2	50.4±0.7

Table 2: Head-only fine-tuning. Evaluation on TEST. PTT is pre-training time. MDF is MEDIA (full), MATIS is MultiATIS++. Mean ± standard deviation, across five seeds. Bold means $\Delta \geq 1$, underline means $\Delta \geq 5$.

are presented in Figure 1. BASELINE-100M is the data present in all datasets. Other datasets append data to that of BASELINE-100M. We see that each DIVERSE dataset has an entropy notably higher than its RANDOM counterpart.

3 Pre-training and fine-tuning

To test H2, for each pair of DIVERSE and RANDOM datasets, we train two ModernBERT encoders, one using the former dataset, one using the latter. For each model we limit the time budget to 483 hours.[3] Pre-training on TOPLINE-2400M, conversely, runs until convergence and takes 1,775 hours. We use the toolkit[4] given by Warner et al. (2025). The models are base-size ModernBERT. Expectedly, the loss profiles of the resulting models (Figure 1) are higher for the models trained on the DIVERSE datasets. All curves other than for TOPLINE-2400M have roughly the same shape, which indicates the similar learning speed. We expect this tendency to persist beyond the 483h time budget (even if we did not train all models until convergence, for the sake of limiting energy consumption). In other words, models trained on diverse data do not seem to learn faster or slower than those trained on randomly sampled data.

To assess the quality of the DIVERSE encoders against that of the RANDOM encoders, we fine-tune all pre-trained models on several classification and sequence labeling tasks in French (evaluated by accuracy), summarized in Table 3. For more details see the original paper (Estève et al., 2026).

4 Fine-tuning results

The results are shown in Tab. 1. A major observation is that both DIVERSE or RANDOM models yield a commensurate performance to TOPLINE-2400M, despite using 6 to 16 times less pre-training data and 3.7 less pre-training time. We find no clear signal, though, as to which of DIVERSE or RANDOM data performs better.

We are, however, particularly interested in the inherent quality of the encoders, independently of fine-tuning effects. Therefore, we perform the same evaluation, but with frozen encoders. Two tasks, MEDIA (full) and MultiATIS++, displayed marked improvements for DIVERSE pre-training datasets over their RANDOM counterpart; we display them in Table 2. We claim that this improvement is indicative of a better potential – at least for these tasks – of DIVERSE encoders for data-constrained scenarii. In particular, in an NLU task (MDF), diversity-driven sampling of 150M tokens leads to equal performances as random sampling of 400M tokens. Note also that, here again, DIVERSE encoders outperform the TOPLINE, despite being considerably more data- and time-efficient.

Future work includes testing this process for auto-regressive models, and investigating the impact of diversities other than lexical (e.g. syntactic).

the RANDOM datasets by batches of data.

[3] Each sampling takes less than 4 hours of CPU time and is not included in the training time budget.

[4] https://github.com/AnswerDotAI/ModernBERT

Task (reference)	Domain	Classes	Size (k)	Measure
PAWS-X (Yang et al., 2019)	Paraphrase classification	2	49 / 2.0 / 2.0	Accuracy
XNLI (Conneau et al., 2018)	NLI	3	393 / 5.4 / 5.0	Accuracy
MASSIVE Intent (FitzGerald et al., 2023)	Intent detection	60	11.5 / 2.0 / 2.9	Accuracy
WikiNER (Nothman et al., 2013)	NER	7	129 / 0.5 / 14.3	F1
MultiATIS++ (Upadhyay et al., 2018)	POS-tagging	131	37.0 / 0.0 / 7.8	F1
Media (full) (Bonneau-Maynard et al., 2006)	NLU	152	13.7 / 1.3 / 3.7	F1

Table 3: Tasks used for fine-tuning. Size refers to the number of entries (by default, sentences), for TRAIN / DEV / TEST. For NER, the number of classes is after BIO expansion.

5 Discussion

We have shown that it is possible to efficiently create small but very DIVERSE pre-training datasets out of a large dataset. Such datasets, when used for pre-training, can yield encoders which have comparable or higher performances to an encoder trained with much more data and for a much longer time. A similar effect can be achieved with just random sampling, however, the potential of DIVERSE encoders is much bigger for some tasks, as shown with head-only fine-tuning experiments.

We believe that the positive impact of diversity sampling on performances might be due to avoiding redundancy in data. Redundancy may, indeed, overvalue the few most frequently occurring data, at the expense of the high quantity of rarely occurring data.

6 Acknowledgements

This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-A0171013834). It was also supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Hélène Bonneau-Maynard, Christelle Ayache, Frédéric Béchet, Alexandre Denis, Anne Kuhn, Fabrice Lefevre, Djamel Mostefa, Mathieu Quignard, Sophie Rosset, Christophe Servan, and Jeanne Villaneau. 2006. Results of the French Evalda-Media evaluation campaign for literal understanding. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, Genes, Italy.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Louis Estève, Christophe Servan, Thomas Lavergne, and Agata Savary. 2026. A Diversity Diet for a Healthier Model: A Case Study of French ModernBERT. In *to appear, Findings of the Association for Computational Linguistics: ACL 2026*, SanDiego, USA. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Jared Kaplan, 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož,

Slovenia. European Language Resources Association (ELRA).

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

Guilherme Penedo, Hynek Kydlíček, Loubna B. Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Manon Scholivet, Agata Savary, Louis Estève, Marie Candito, and Carlos Ramisch. 2025. SELEXINI – a large and diverse automatically parsed corpus of French. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 83–98, Abu Dhabi, UAE. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.