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] Université Paris-Saclay, CNRS, LISN, first.last@lisn.fr
[2]AMIAD, Pôle Recherche, first.last@polytechnique.edu

Abstract

Diversity has been gaining interest in the NLP community in recent years. At the same time, state-of-the-art transformer models such as ModernBERT use very large pre-training datasets, which are driven by size rather than by diversity. This summons to investigate the impact of diversity on pre-training. We do so in this study, with the express intent of reducing pre-training dataset size, while retaining at least comparable performance. We compare diversity-driven sampling algorithms, and we use the best one to pre-train several ModernBERT models on French with a fixed compute budget. We fine-tune and evaluate them on a variety of French benchmarks. We compare them with models pre-trained on randomly sampled data of commensurate size, with the same compute budget. We find that both random and diversity-driven sampling may reduce the pre-training dataset by up to 94% and the pre-training time by up to 73% while maintaining performance. Moreover, in some tasks, the inherent quality of models, estimated via head-only fine-tuning, is up to 10 points higher with diversity sampling than with random sampling.

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) benefitted 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.

We test these hypotheses by (1) proposing two diversity-driven sampling algorithms: one using a heuristic and one based on gradient descent, (2) benchmarking these two algorithms, along with other algorithms found in the literature, and (3) using the best sampling to pre-train ModernBERT (Warner et al., 2025) models, and evaluate them after fine-tuning on a variety of tasks in French.[2]

After presenting previous works (§2), the sampling algorithms we benchmark (§3), and our protocol (§4), we show that increased lexical diversity of the pre-training data substantially reduces the pre-training dataset size while maintaining performance. Moreover, for some tasks, we achieve substantially enhanced performances (§5).

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

[2]Models: https://huggingface.co/collections/cservan/french-modernbert – Code and scripts: https://gitlab.lisn.upsaclay.fr/esteve/transformers.

Findings of the Association for Computational Linguistics: ACL 2026, pages 34168–34181

2 Previous works

For the needs of our study, we need to document what is diversity (and how it is quantified), but also sampling algorithms that relate to our problem, and the model architecture we shall use.

2.1 Diversity

Diversity has been discussed in multiple fields and may be defined as the study of categories in a population, through *variety* which focuses on how many categories are present, *balance* which focuses on how evenly distributed categories are, and *disparity* which focuses on how fundamentally different categories are (Stirling, 1994a,b, 2007; Ramaciotti Morales et al., 2021). Ecology has studied biodiversity as early as the works of Darwin (1888), borrowed from information theory (Hill, 1973; Patil and Taillie, 1982), and further generalised the concept to build upon functional (Chao et al., 2014) and phylogenetic (Chiu et al., 2014) distances between species. It is driven by environmental conversation (Sarkar, 2002), and arguably shows maturity in understanding diversity. As a consequence, economics has built upon the works of ecology (Stirling, 1994a,b, 2007) and studied notably the impact of inequalities (Ceriani and Verme, 2012). Linguistics has assessed the diversity of languages (Greenberg, 1956; Harmon and Loh, 2010). NLP as well has recently studied diversity, although with less consensus and formalisation than in ecology, to the extent that over 100 different measures for diversity have been observed (Estève et al., 2026). Even if focusing on the diversity of input data, there is no consensus on the measurement of diversity, as measures can be based on quantity (Bansal et al., 2021; Mohamed et al., 2022; Bella et al., 2022; Gueuwou et al., 2023; Parrish et al., 2024; Leeb and Schölkopf, 2024; Abdelkadir et al., 2024), BLEU (Awasthi et al., 2022; Burchell et al., 2022), distances (Stasaski et al., 2020; Shi et al., 2021), entropy (Stasaski et al., 2020), human judgement (Lee et al., 2020), overlap (Samardzic et al., 2024; Yadav et al., 2024), type-token-ratio (Liu et al., 2024; Song et al., 2024; Pouran Ben Veyseh et al., 2022), or distances between distributions (Kumar et al., 2022).

As a consequence of the lack of consensus and formalisation in NLP, we choose to measure diversity using entropy (Shannon and Weaver, 1949; Wiener, 1939). Our choice is driven by the fact that the properties of entropy are well-analysed, and it is a cornerstone for fields such as ecology where the notion of diversity is well formalised.

2.2 Diversity-driven sampling

Looking at the broad picture of computer science, sampling a dataset for diversity relates to the knapsack problem (Cacchiani et al., 2022), which consists in sampling a set while maximising a metric (and respecting a size budget). The standard version of this problem requires giving a constant value (or benefit) to each sentence, which means it can at best be a heuristic for diversity sampling.

Closer to data found in NLP applications, the Machine Learning literature has a survey on the topic of data diversity and sampling (Gong et al., 2019) in which they mend the absence of "systematical analysis of the diversification in the machine learning system" (p. 64323). They present, for supervised learning, Determinantal Point Processes (DPP) for producing non-redundant batches out of a fixed dataset, which has the variant k-DPP where k is the requested batch size. DPP has indeed been considered for enhancing diversity (Yang et al., 2021; Hemmi et al., 2022), as avoiding redundancy may foster diversity, and in turn improve the machine learning process.

Sampling a dataset with the intent of favouring diversity relates to the literature on coreset selection, which aims to select "a subset of the most informative training samples" as defined in the surveyby (Guo et al., 2022, p. 181). A coreset is in turn more efficient to train on than the full dataset due to reduced size, while retaining as much performance as possible. Coreset selection has been used in the perspective of selecting points in a space (Agarwal et al., 2005; Perlitz et al., 2023; Nguyen and He, 2025), including with the concept of diversity as pairwise distances (Indyk et al., 2014), which is an instance of disparity in the cross-domain conceptualisation of diversity (Stirling, 2007; Estève et al., 2026). These approaches are at least quadratic to the number of categories time-wise as they require computing all the distances between categories. Coreset selection however does not need to rely on distances, and may thus be subquadratic by focusing on other objects, for instance on confidence, loss or gradient in an already trained model.

Some other approaches consider a specific dataset split to assess which subset of the training split would be a good coreset (San Joaquin et al., 2024). Software has been developed to perform coreset selection (Guo et al., 2022), and coreset

selection has been used in the pre-training process of BERT models (Huang et al., 2022; Attendu and Corbeil, 2023) as well as their fine-tuning (Nguyen and He, 2025; Sharma et al., 2025), providing apparent compromise between performance and efficiency. This has also been tested for LLMs (San Joaquin et al., 2024). An extensive analysis of coreset selection methods in deep learning however found that "although various methods have advantages in certain experimental settings, random selection is still a strong baseline" (Guo et al., 2022, p. 181), pointing to a potentially limited benefit of complex methods given their cost, relative to simpler methods.

In NLP, diversity quantification is also used for generative purposes (Hu et al., 2019; Zhou and Lampouras, 2021; Zhang et al., 2023; Yadav et al., 2024), with an interest in the quality-diversity tradeoff (Zhang et al., 2021), for data augmentation (Song et al., 2024), or instruction tuning (Yang et al., 2025). In active learning, diversity is quantified when synthesizing datasets from the ground up, rather than by sampling (Shi et al., 2021; Xia et al., 2024).

Approaches based on semantic structures present in data (Oren et al., 2021; Gupta et al., 2022) or task-specific objects such as in sentiment analysis where the sentiment of reviews can be required for sampling (Jiang et al., 2023) have been studied. These are motivated by task-specific rather than task-agnostic diversity, which can be for example based on lexical units.

We see mentions of distance-based sampling, as fostering higher distances between studied categories may be a way to improve diversity. This is the case, for instance, in the k-means++ algorithm (Kim, 2020) or the previously mentioned k-DPP sampling (Golobokov et al., 2022).

Chubarian et al. (2021, Algorithm 4) reference the algorithm by Lee et al. (2009) which iterates over a dataset, trying to increase entropy at each data point, by selecting it, unselecting it, or having it replace another selected point.

Given this rich bibliography related to diversity-driven sampling, we proceeded to select algorithms which might match our framework and hypotheses H1-H2. The knapsack problem is NP-hard (Martello et al., 2000) and recent findings indicate that its lower bound is "in subexponential and superpolynomial" (Zhang, 2025, p. 11934), rendering it unfeasible for our needs. k-DPP suffers a comparable issue. To use k-DPP to sample a dataset, we would need to produce a "batch" of the dataset size we want. However, sampling takes at best polynomial time to batch-size k (Derezinski et al., 2019). This is prohibitive if we want a k representing dataset size instead of batch size. Approaches based on distances such as k-means++ can be filtered out, as computing the matrix of distances is quadratic.[3] Diversity-driven approaches in decoding and in active learning are not appropriate either, as we aim to algorithmically sample a dataset rather than create one from scratch. The work of (Yang et al., 2025) has not been tested beyond 10,000 sentences, which is several orders of magnitude lower than our needs. Finally task-specific sampling does not fit our interest in task-agnostic methods.

The approach used by Chubarian et al. (2021) is adapted to scaling and can be used in our study. We will compare it to two other algorithms proposed by us. One of them (§3.1), like Chubarian et al. (2021), works on a local basis, i.e., actions are chosen only by looking at a limited subset of the dataset. The other one accounts for the global nature of diversity and is based on gradient descent (§3.3). All three approaches will be compared to random sampling (Jiang et al., 2023).

2.3 ModernBERT

Warner et al. (2025) proposed ModernBERT, which ports modern transformer improvements from autoregressive models (Radford et al., 2018) to masked architecture (Devlin et al., 2019). The main impact lies in the context extension from 512 to 8,192 tokens, based on a Rotary Positional Embedding (Su et al., 2024), also known as RoPE. The GeGLU layers (Shazeer, 2020) are preferred instead of MLP, which removes bias terms and adds layer normalization after embeddings.

ModernBERT models provide new state-of-the-art performance on various tasks, such as Natural Language Understanding, Information Retrieval, Long-Context Text Retrieval, and Code Retrieval (Warner et al., 2025).

The original ModernBERT was trained on 2 trillion English tokens. The original paper needed 1,879 and 4,076 hours of H100 computation respectively to train BASE and LARGE models.[4]

We aim to reduce these particularly high training costs by using a diversity-driven data sampling

[3] Furthermore, using distances to measure diversity (disparity) has issues, notably due to the high correlation between disparity and vocabulary size (Estève et al., 2024).

[4] "Training Time" in their Table 3, times GPU count.

approach for the pre-training process.

3 Sampling algorithms

We hereby describe sampling algorithms which we will run and compare. For each algorithm, we shall test multiple parameters when possible. In the upcoming descriptions, we use the following notations: v is the global vocabulary size, p_i is the empirical probability of the ith vocabulary entry, and s is a sentence. FULL denotes the input dataset, which comprises n sentences. DIV denotes the output dataset, which comprises n' sentences. Entropy is denoted as H. As a BASELINE, we perform random sampling with $q \in (0, 1]$ as the ratio of the sampled set to FULL.

3.1 Patient add-only method

Our first proposal is Algorithm 1. It takes an initial dataset BASE, which is data the Algorithm starts with to have a reasonable distribution balance-wise. It also takes a maximum output size S, which can be disabled if $S = -1$. In the internal loop (lines 5-14), we consider one candidate sentence s from EXTENSION at once. We filter and normalise it (line 6) to avoid artificial increase in diversity.[5] We check if, when added to DIV*, s increases the entropy H (line 7). If so, we check if this increase in H is higher than for a previously found sentence *best* (line 9). If so, *best* becomes s (line 10). Once sentences improving the entropy H have been found e times, the best is added to the final corpus DIV* (lines 11-12) and we start looking for a new best sentence out of e beneficial sentences (line 13). If the size of DIV* is greater or equal to S, then early stopping is triggered, and DIV* is returned (line 14).

Variable e, for exhaustivity of search, tells us how many sentences increasing diversity to look at before picking the optimal one to append to DIV*. The higher e, the more each *best* sentence may increase diversity. We use an array E of several exhaustivities, one per dataset traversal, sorted by decreasing values, to have a deep search on the first traversal as the data has never been seen before by the algorithm, and less so at each following traversal. Without early stopping, we return DIV* when all exhaustivity levels have been used (line 15). The algorithm has a $O(|\text{EXTENSION}|)$ complexity. It was experimentally tested to create a large diverse automatically parsed corpus of French (Scholivet et al., 2025).

Algorithm 1 Algorithm by Scholivet et al. (2025) to sample data while maximising entropy.

Require: BASE $\subset$ FULL, an initial dataset
Require: EXTENSION = FULL $\setminus$ BASE, a dataset to select from
Require: E, a decreasing array of exhaustivity search parameters (positive integers)
Require: S, maximum size of resulting dataset (-1 if no maximum size)
1: DIV* $\leftarrow$ BASE, the final dataset
2: **for each** e $\in$ E **do**, for each exhaustivity level
3: i $\leftarrow$ 0, counter for current exhaustivity
4: best $\leftarrow \emptyset$, best sentence to append
5: **for each** s $\in$ EXTENSION **do**
6: s $\leftarrow$ normalised(s)
7: **if** H (DIV* $\cup$ s) > H (DIV*) **then**
8: i $\leftarrow$ i + 1
9: **if** H (DIV* $\cup$ s) > H (DIV* $\cup$ best) **then**
10: best $\leftarrow$ s
11: **if** i = e **then**
12: DIV* $\leftarrow$ DIV* $\cup$ best
13: i $\leftarrow$ 0 ; best $\leftarrow \emptyset$
14: **if** $-1 < S \leq |\text{DIV*}|$ **then return** DIV*
15: **return** DIV*

3.2 Impatient add-remove-replace method

Lee et al. (2009) as presented by Chubarian et al. (2021), also propose an algorithm based on iterative dataset traversals. Unlike Algorithm 1, it has no memory of recent favourable sentences. It starts with a BASE of just the sentence with the highest entropy. It then traverses FULL. At each sentence s, it assesses the impact on entropy for three different actions. If s is not already selected, it tests (1) adding s, or (2) replacing a random selected sentence with s. If s is already selected, it tests (3) unselecting s. If the best action improves entropy by at least a margin (dependent on a parameter ϵ), this action is performed. If no update was done during a traversal, the algorithm returns its working dataset as DIV; otherwise it starts another traversal. A pseudocode is given in Chubarian et al. (2021, Algorithm 4).

3.3 Sampling by gradient descent

We here sample by using a macroscopic view of the dataset. We do so with no BASE, using gradient

[5] (Telephone) numbers, HTML/XML tags, URLs, paths, emoticons, punctuation series, phonetic characters, series of alphanumerical tokens, and non-French characters are represented by placeholders, e.g., [NUMBER].

descent, where we learn for each sentence from FULL whether it should belong in DIV. As a result, each sentence is given a partial belonging $r \in [0, 1]$, which gets updated at each gradient descent step.

We want to learn a vector of sentence belonging $b \in [0, 1]^n$. However, learning directly this vector through gradient descent could give values outside of $[0, 1]$. We thus instead learn a real row vector $a \in \mathbb{R}^n$ which we then pass through the sigmoid function $\text{sigm}(x) = (1 + e^{-x})^{-1}$ to obtain the $b \in [0, 1]^n$ row vector.

$$\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \xrightarrow{\text{sigm}} \begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix} \quad (1)$$

We then multiply b by m, the $n \times v$ matrix of the vocabulary for each sentence (thus $m \in (\mathbb{N} \cup \{0\})^{n \times v}$). This yields the $c = bm$ row vector of word frequencies.

$$\begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix} \begin{bmatrix} m_{1,1} & \cdots & m_{1,v} \\ \cdots & \cdots & \cdots \\ m_{n,1} & \cdots & m_{n,v} \end{bmatrix} \rightarrow \begin{bmatrix} c_1 & \cdots & c_v \end{bmatrix} \quad (2)$$

We then use c to compute vector p of vocabulary probabilities where $p_i = c_i (\sum_{j=1}^{v} c_j)^{-1}$

$$\begin{bmatrix} c_1 & \cdots & c_v \end{bmatrix} \rightarrow \begin{bmatrix} p_1 & \cdots & p_v \end{bmatrix} \quad (3)$$

We compute entropy H over $p_1, \cdots, p_v$, which is the main element of the loss function. If we negate it, gradient descent will try to increase entropy. A shortcoming is that sentences may be selected with b_i far from either 0 or 1, but we want in the end each sentence to be unambiguously close to 0 (unselected) or 1 (selected). To solve this, we add a penalty which increases as b_i gets distant from 0 or 1. This penalty is multiplied by α (default to 1000), and increases at each gradient descent step t, against the maximum number of steps T.

$$\mathcal{L} = (-H) + \frac{t}{T} \left(\alpha \frac{1}{n} \sum_{i=1}^{n} b_i (1 - b_i) \right) \quad (4)$$

At the end, we select sentences with $\lceil b_i \rfloor = 1$.

We have now presented the gradient descent approach, but there remains to discuss the initialisation strategies for learned vector a. We first test **random initialisation** where values of a fall within $[-1, 1]$. This entails starting values in b in the approximate range $[0.27, 0.73]$.

We also test **fixed initialisation** where all values of a are initialised at $\tanh(w)$ where $w \in \mathbb{R}$. If $w < 0$, then $b_i < 0.5$ (the start is pessimistic for each sentence), and conversely if $w > 0$, then $b_i > 0.5$ (the start is optimistic for each sentence). Based on empirical results, Figure 1 depicts eleven values for w, in the range $[-0.25, 0.25]$ with a stepping of 0.05. This entails starting values in b in the approximate range $[0.46, 0.54]$.

We last test a **designed initialisation** based on vocabulary dissimilarity between sentences, to benefit starting entropy. We first compute the $n \times n$ matrix $M = -(mm^\intercal)$ of dissimilarity between sentences. We reduce a dimension to have a vector d, where $d_i = n^{-1} \sum_{j=1}^{n} M_{i,j}$. We then apply normalisation, where μ is the mean of d, and σ its standard deviation, to obtain $a_i = 0.1 \times \sigma^{-1} (d_i - \mu)$.

3.4 Comparison of sampling methods

We see in Figure 1 the comparison of the algorithms we described when sampling the union of French treebanks of Universal Dependencies (UD) v2.16 (Nivre et al., 2020). Sampling was done at the sentence level. It is important to account for both n' (lower is better) and H, as comparing the diversity of datasets of incommensurate sizes is hazardous.

We first see that non-random samplings yield entropies markedly higher than random sampling of commensurate output size n', with the exception of the gradient-based approach when using a fixed initialisation with a high w. Among the studied methods, three reach high entropies: Lee et al. (2009), our gradient-based method (using a fixed initialisation with a low w), and our batch-based method (Algorithm 1). We can assess which is best: gradient-based sampling has high memory requirements (it keeps the whole dataset in memory), and our batch-based method yields much smaller datasets than that of Lee et al. (2009). We thus select Algorithm 1, which we denote by "DIVERSE" for our experiments.

4 Pre-training protocol

Our protocol is as follows. We want to compare the quality of encoders trained on RANDOM and DIVERSE data. To do so, we use diversity-driven sampling to create DIVERSE datasets. For each DIVERSE dataset, we create a matching RANDOM dataset of commensurate size. For each pair of DIVERSE and RANDOM datasets, we train two ModernBERT encoders, one using the former dataset, one using the latter. We finish by fine-tuning on downstream tasks to assess the quality of the DIVERSE encoder against that of the RANDOM en-

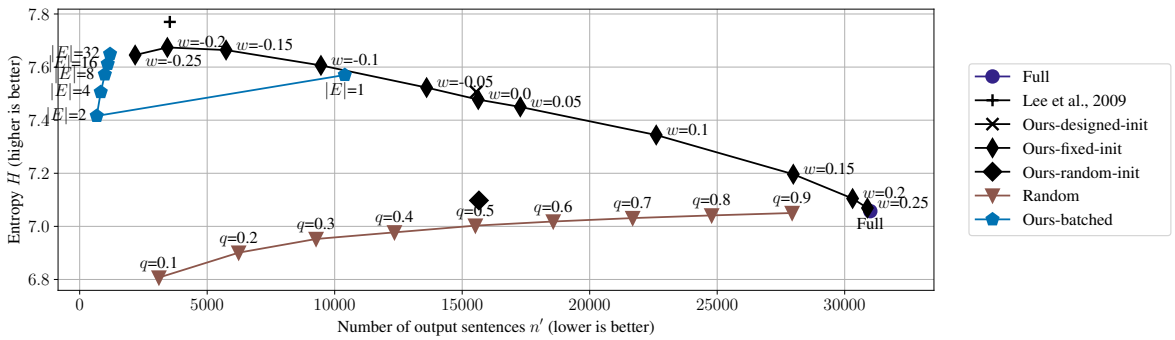

Figure 1: Comparison of sampling algorithms run on UD v2.16 (French). Lee et al. (2009) have the best entropy, but ours-batched is just below, with only a third in n'. For ours-batched, the last value of E is 20, with each preceding value increasing by 10 (the only exception is at $|E| = 1$ where the only value is 1).

coder. We give details in the following subsections.

4.1 Sampling

We use Algorithm 1, explained in the previous section. We first select a FULL dataset. We extract from it the BASE, which will be shared by all RANDOM and DIVERSE models, while the rest of the dataset is the EXTENSION.[6] In our case, the BASE is randomly-selected (approximately) 5% of FULL (and thus the EXTENSION is approximately 95% of FULL). For each DIVERSE sampling configuration, the algorithm selectively adds data from EXTENSION to the BASE to create a DIVERSE dataset. For each DIVERSE dataset it generates, a RANDOM dataset of commensurate size is created by randomly adding data from the EXTENSION to the BASE. The pre-training data used to create the FULL, and in turn train the model, is extracted from the French part of Wikipedia and OPUS (Tiedemann, 2012) corpora, which enables us to propose some open-source models. The FULL is shuffled. Each sampling takes less than 4 hours of CPU time.

4.2 Pre-training datasets

The resulting datasets are presented in Table 1. BASELINE-100M is the BASE present in all datasets, with no data from the EXTENSION. Conversely, TOPLINE-2400M uses the whole FULL.

We see that each DIVERSE dataset has an entropy notably higher than its RANDOM counterpart (i.e. with commensurate size). With an increasing $|E|$ (also due to higher values in E), the sampled data from EXTENSION is more diverse, but also much smaller, meaning that once merged with the BASE, diversity rankings are reversed.

Dataset	Words	H' (↑)	H (↑)
TOPLINE-2400M	2 379M	7.54	7.55
RANDOM-400M	401M	7.57	7.59
DIVERSE-400M	397M	8.80	8.53
RANDOM-240M	242M	7.59	7.60
DIVERSE-230M	230M	8.99	8.45
RANDOM-150M	150M	7.61	7.61
DIVERSE-150M	147M	9.25	8.21
BASELINE-100M	105M	N/A	7.61

Table 1: Sampled datasets: part of EXTENSION (with entropy H') is appended to the BASE (the whole has entropy H). DIVERSE-400M has E only at 1, DIVERSE-230M from 16 to 4 (4 values; originally, it was planned to be from 50 to 20, but an error in settings made it run from 16 to 4) DIVERSE-150M from 170 to 20 (16 values).

4.3 Pre-training process

ModernBERT has a native sequence length of 8,192 tokens and incorporates recent architecture improvements, such as GeGLU layers, RoPE positional embeddings, and alternating local-global attention. The vocabulary of the tokenizer is set to 129K, and includes 1K unused tokens to support downstream applications.

We train one base-size ModernBERT model per dataset from Table 1, using the toolkit[7] given by Warner et al. (2025). Each resulting model has 22 layers, a total parameter count of 200 million, and a hidden size of 768 with a GLU expansion of 2,304. We assume a fixed pre-training compute budget of 483 hours of H100 per model. As a comparison,

[6] The use of a BASE, i.e. forcing the algorithm to start with already selected data, was assessed by Scholivet et al. (2025) to improve sampling for large datasets containing noise.

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

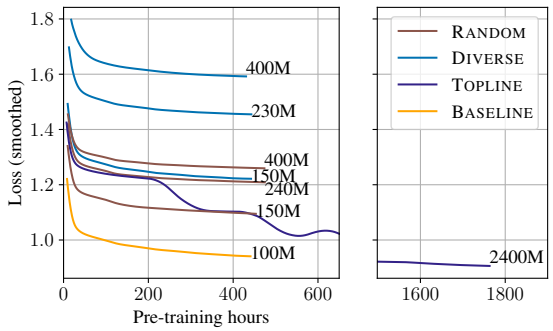

Figure 2: Pre-training process for encoders. Loss is smoothed using a moving average (window size: 2×10^5, series sizes from 1.5M to 6.3M points) applied thrice.

the TOPLINE using FULL took 1,775 hours of GPU time (H100). The data, the pre-training recipe, and the models are openly available.

4.4 RANDOM and DIVERSE encoders

Pre-training on our different datasets has yielded loss profiles which are mainly characterised by whether the training data was RANDOM or DIVERSE (Figure 2).[8] Looking at RANDOM encoders, more data means a higher loss. This is also true for DIVERSE encoders, but their loss remains notably higher than their RANDOM counterpart. About convergence speed, we see that RANDOM encoders all took approximately the same amount of time. DIVERSE encoders had notably higher losses than their respective RANDOM counterparts. The higher losses are likely due to the higher difficulty of training due to diversity, but conversely, it is possible that the encoder better models language and, especially, rare phenomena. If so, we would expect DIVERSE encoders to perform better on downstream tasks.

To evaluate this, and see the impact of our diversity sampling approach on the pre-training data, we aim to perform several classical NLP tasks described in Table 2: classification and sequence labelling tasks (respectively first and second half of the Table). For multilingual datasets, we use the French part. We selected tasks with an increasing number of classes, assuming that it correlates with difficulty. It should be noted that the MEDIA tasks have been considered the most challenging ones by (Béchet and Raymond, 2019).

[8]Henceforth, the names of the datasets also denote the encoders trained upon.

5 Results

We used two models for the classification and sequence labelling tasks. Both use ModernBERT encoders, on which a width-preserving dense layer and a dense layer transforming to the number of classes are added as a decoder. The first uses simply the number of classes in the task. The second expands the list of labels to all the possible tags in the BIO scheme.

5.1 Fine-tuning parameters

Fine-tuning is performed in two ways. The first one fine-tunes both the encoder and the head to see the maximum potential of the pipelines. The second one freezes the encoder and trains the head, to assess encoder quality. Hyperparameters are: learning rate is 4×10^{-4}, weight decay is 1×10^{-3}, max epochs is 32. The optimizer is Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\times10^{-8}$), with an inverse square root scheduler (15% warmup). Batch size is 32, evaluating every 64 steps. We trigger an early stopping if 64 train losses are lower than 1×10^{-3}, or 32 DEV scores not improving. Computation is FP32.

5.2 Fine-tuning results

Table 3 summarizes evaluations of models with both encoders and heads fine-tuned, on the TEST of each task, where rows are grouped by commensurate pre-training dataset sizes. A major observation is that both RANDOM and DIVERSE models yield a commensurate performance to TOPLINE-2400M, despite reducing the pre-training dataset by up to 94% and the training time by up to 73%.[9] These results confirm former findings by (Guo et al., 2022) (§2.2). We find no clear signal, though, as to which of DIVERSE or RANDOM models perform better. We may hypothesize that, for these tasks, the model makes limited use of lexis, or the lexical knowledge gained by DIVERSE pre-training was not impactful.

Note, however, that our main objective is not to boost performances. We are instead particularly interested in the inherent quality of the encoders, independently of fine-tuning effects. Therefore, we perform the same evaluation, but with frozen encoders. This idea resembles probing classifiers, in which different layers of an encoder are used (with no fine-tuning) on input of simple classification

[9]The BASELINE, also performs comparatively well, except on MDF.

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

Table 2: 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.

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 3: Encoder & head fine-tuning. Evaluation on TEST. PTT is pre-training time. PX is PAWS-X, AMI is Amazon Massive Intent, WNER is WikiNER, MATIS is MultiATIS++, MDF is MEDIA (full). 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 4: 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 $\Delta \geq 5$.

tasks, to understand which knowledge they encode, rather than to optimize performances of these tasks.

In the experiment 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 4. For these tasks, relative to BASELINE, the 50M tokens from the EXTENSION received by RANDOM-150M and DIVERSE-150M had notably different consequences on performance. RANDOM-150M saw improvements in average performance of respectively 1.5 and 2.4 points, while DIVERSE-150M saw improvements in average performance of respectively 11.1 and 6.5 points, much higher than its counterpart. This advantage of DIVERSE over RANDOM pre-training data reduces between DIVERSE-230M and RANDOM-240M, and vanishes between DIVERSE-400M and RANDOM-400M. It should also be noted that, for these two tasks, DIVERSE encoders outperform the TOPLINE, despite having considerably less pre-training data and lower number of pre-training steps (Figure 2). Looking back at the diversity of each pre-training dataset (Table 1), we see that this performance advantage is driven more by the diversity of the data selected from the EXTENSION than of the whole pre-training dataset. DIVERSE pre-training datasets with fewer data from EXTENSION have sentences which individually contribute more to diversity, as the sampler parameters were more selective, which may be what contributed to the increased performance of these models, relative to their RANDOM counterparts.

The fact that this performance gain exists solely in head-only fine-tuning is indicative of a fundamental difference between RANDOM and DI-

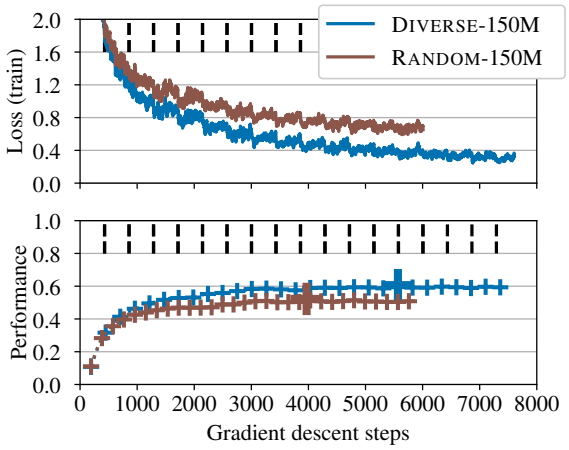

Figure 3: Fine-tuning (head only) for MEDIA (full). Dashed thick lines delimit dataset traversals. Thick crosses denote maximum value.

VERSE encoders. Precisely, it indicates that the DIVERSE encoder contains more task-important lexical knowledge. As, conversely, fine-tuning both the encoder and the head yields commensurate performance for commensurately sized pre-training datasets, we may hypothesize that fine-tuning RANDOM encoders compensates their prior lack of lexical knowledge. This points to the fact that DIVERSE encoders have better potential – at least for these tasks – for data-constrained scenarii.

We plot one of MEDIA's fine-tunings in Figure 3. We see that DIVERSITY provides an appreciable improvement in loss and performance. This observation on MEDIA ports to MultiATIS++ in lesser proportions, but MEDIA is further from being solved than MultiATIS++.

6 Discussion

We have shown that, (H1) it is possible to efficiently create small but very DIVERSE pre-training datasets out of a large dataset, and that (H2) their use tends to have a neutral or positive effect on encoder quality and performance, relative to RANDOM pre-training datasets of commensurate sizes, but also larger datasets. Consequently, it is possible to considerably trim pre-training datasets while retaining performance. Precisely, on the task benefitting the most from lexical diversity (NLU), for the head-only fine-tuning scenario, diversity-driven sampling has allowed to reach the performance of an encoder pre-trained on about 400M RANDOM tokens, with only 150M DIVERSE tokens. Training time was also about a fourth that of the TOPLINE. Future work includes testing this process for auto-regressive models, and investigating the impact of diversities other than lexical (e.g. syntactic).

7 Limitations

This study has the limitation of the potential correlation of lexical diversity to other linguistic diversities, such as morphological, syntactic, or semantic diversity. It may be that the encoders that benefitted from increased lexical diversity were also impacted by changes these other linguistic diversities.

8 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

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. 2005. Geometric Approximation via Coresets. *Combinatorial and Computational Geometry*, 52(MSRI Publications).

Jean-michel Attendu and Jean-philippe Corbeil. 2023. NLU on Data Diets: Dynamic Data Subset Selection for NLP Classification Tasks. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 129–146, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Abhijeet Awasthi, Ashutosh Sathe, and Sunita Sarawagi. 2022. Diverse parallel data synthesis for cross-database adaptation of text-to-SQL parsers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11548–11562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Trapit Bansal, Karthick Prasad Gunasekaran, Tong Wang, Tsendsuren Munkhdalai, and Andrew McCallum. 2021. Diverse distributions of self-supervised tasks for meta-learning in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5812–5824, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frédéric Béchet and Christian Raymond. 2019. Benchmarking benchmarks: introducing new automatic indicators for benchmarking Spoken Language Understanding corpora. In *InterSpeech*, Graz, Austria.

Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Cheema, and Fausto Giunchiglia. 2022. Language diversity: Visible to humans, exploitable by machines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 156–165, Dublin, Ireland. Association for Computational Linguistics.

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.

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics.

Valentina Cacchiani, Manuel Iori, Alberto Locatelli, and Silvano Martello. 2022. Knapsack problems — An overview of recent advances. Part I: Single knapsack problems. *Computers & Operations Research*, 143:105692.

Lidia Ceriani and Paolo Verme. 2012. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10(3):421–443.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Chun-Huo Chiu, Lou Jost, and Anne Chao. 2014. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, 84(1):21–44. Number: 1.

Karine Chubarian, Abdul Rafae Khan, Anastasios Sidiropoulos, and Jia Xu. 2021. Grouping words with semantic diversity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3217–3228, Online. Association for Computational Linguistics.

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.

Charles Darwin. 1888. *The Descent of Man,: And Selection in Relation to Sex*. John Murray, Albemarle Street. Google-Books-ID: NaPu24dY4iAC.

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.

Michal Derezinski, Daniele Calandriello, and Michal Valko. 2019. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Louis Estève, Agata Savary, and Thomas Lavergne. 2024. Vector spaces for quantifying disparity of multiword expressions in annotated text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 110–130, Bangkok, Thailand. Association for Computational Linguistics.

Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2026. A survey of diversity quantification in natural language processing: The why, what, where and how. *Preprint*, arXiv:2507.20858.

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.

Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and

Yi Liu. 2022. DeepGen: Diverse search ad generation and real-time customization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 191–199, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in Machine Learning. *IEEE Access*, 7:64323–64350.

Joseph Harold Greenberg. 1956. The Measurement of Linguistic Diversity. *Language*, 32(1):109. Number: 1.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. DeepCore: A Comprehensive Library for Coreset Selection in Deep Learning. In *Database and Expert Systems Applications*, pages 181–195, Cham. Springer International Publishing.

Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Structurally diverse sampling for sample-efficient training and comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4966–4979, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Harmon and Jonathan Loh. 2010. The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages. *Language Documentation & Conservation*, 4:97–151.

Shinichi Hemmi, Taihei Oki, Shinsaku Sakaue, Kaito Fujii, and Satoru Iwata. 2022. Lazy and Fast Greedy MAP Inference for Determinantal Point Process. *Advances in Neural Information Processing Systems*, 35:2776–2789.

Mark Oliver Hill. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432. Number: 2 Publisher: Ecological Society of America.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Xin Huang, Ashish Khetan, Rene Bidart, and Zohar Karnin. 2022. Pyramid-BERT: Reducing Complexity via Successive Core-set based Token Selection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8798–8817, Dublin, Ireland. Association for Computational Linguistics.

Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S. Mirrokni. 2014. Composable core-sets for diversity and coverage maximization. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '14, pages 100–108, New York, NY, USA. Association for Computing Machinery.

Han Jiang, Rui Wang, Zhihua Wei, Yu Li, and Xinpeng Wang. 2023. Large-scale and multi-perspective opinion summarization with diverse review subsets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5641–5656, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2001.08361.

Yekyung Kim. 2020. Deep active learning for sequence labeling based on diversity and uncertainty in gradient. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 1–8, Suzhou, China. Association for Computational Linguistics.

Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. "diversity and uncertainty in moderation" are the key to data selection for multilingual few-shot transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1042–1055, Seattle, United States. Association for Computational Linguistics.

Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. 2020. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online. Association for Computational Linguistics.

Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. 2009. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, STOC '09, pages 323–332, New York, NY, USA. Association for Computing Machinery.

Felix Leeb and Bernhard Schölkopf. 2024. A diverse multilingual news headlines dataset from around the world. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 647–652, Mexico City, Mexico. Association for Computational Linguistics.

Guisheng Liu, Yi Li, Zhengcong Fei, Haiyan Fu, Xiangyang Luo, and Yanqing Guo. 2024. Prefix-diffusion: A lightweight diffusion model for diverse image captioning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12954–12965, Torino, Italia. ELRA and ICCL.

Silvano Martello, David Pisinger, and Paolo Toth. 2000. New trends in exact algorithms for the 0–1 knapsack problem. *European Journal of Operational Research*, 123(2):325–332.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Binh-Nguyen Nguyen and Yang He. 2025. Swift Cross-Dataset Pruning: Enhancing Fine-Tuning Efficiency in Natural Language Understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 726–739, Abu Dhabi, UAE. Association for Computational Linguistics.

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).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

Inbar Oren, Jonathan Herzig, and Jonathan Berant. 2021. Finding needles in a haystack: Sampling structurally-diverse training sets from synthetic data for compositional generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10793–10809, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alicia Parrish, Susan Hao, Sarah Laszlo, and Lora Aroyo. 2024. Is a picture of a bird a bird? a mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 1–18, Torino, Italia. ELRA and ICCL.

Ganapati P. Patil and Charles Taillie. 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*, 77(379):548–561. Number: 379 Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active Learning for Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877, Singapore. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. MINION: a large-scale and diverse dataset for multilingual event detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, Open AI.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphaël Fournier-S'Niehotta, Rémy Poulain, Lionel Tabourier, and Fabien Tarissan. 2021. Measuring diversity in heterogeneous information networks. *Theoretical Computer Science*, 859:80–115. Publisher: Elsevier.

Tanja Samardzic, Ximena Gutierrez, Christian Bentz, Steven Moran, and Olga Pelloni. 2024. A measure for transparent comparison of linguistic diversity in multilingual NLP data sets. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3367–3382, Mexico City, Mexico. Association for Computational Linguistics.

Ayrton San Joaquin, Bin Wang, Zhengyuan Liu, Nicholas Asher, Brian Lim, Philippe Muller, and Nancy F. Chen. 2024. In2Core: Leveraging Influence Functions for Coreset Selection in Instruction Finetuning of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*

2024, pages 10324–10335, Miami, Florida, USA. Association for Computational Linguistics.

Sahotra Sarkar. 2002. Defining "Biodiversity"; Assessing Biodiversity. *The Monist*, 85(1):131–155. Publisher: Oxford University Press.

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.

Claude Elwood Shannon and Warren Weaver. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Avinash Kumar Sharma, Aisha Hamad Hassan, and Tushar Shinde. 2025. Towards Efficient FinBERT via Quantization and Coreset for Financial Sentiment Analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 70–74, Suzhou, China. Association for Computational Linguistics.

Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024. Scaling data diversity for fine-tuning language models in human alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14358–14369, Torino, Italia. ELRA and ICCL.

Katherine Stasaski, Grace Hui Yang, and Marti A. Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4958–4968, Online. Association for Computational Linguistics.

Andrew Stirling. 1994a. Diversity and ignorance in electricity supply investment. *Energy Policy*, 22(3):195–216.

Andrew Stirling. 1994b. *Power technology choice: putting the money where the mouth is?* PhD, University of Sussex, Sussex.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719. Number: 15 Publisher: Royal Society.

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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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.

Norbert Wiener. 1939. The ergodic theorem. *Duke Mathematical Journal*, 5(1):1–18.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. Explicit over implict: Explicit diversity conditions for effective question answer generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6876–6882, Torino, Italia. ELRA and ICCL.

Xiaohan Yang, Kun Niu, Xiao Li, and Ruijie Yu. 2021. Enhancing Recommendation Diversity Using Determinantal Point Process Forward Inference and Backward Elimination. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 56–60. ISSN: 2575-4955.

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

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Measuring data diversity for instruction tuning: A systematic analysis and a reliable metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18530–18549, Vienna, Austria. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

Xinran Zhang, Maosong Sun, Jiafeng Liu, and Xiaobing Li. 2023. Lingxi: A diversity-aware Chinese modern poetry generation system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 63–75, Toronto, Canada. Association for Computational Linguistics.

Zhidong Zhang. 2025. Lower bound of computational complexity of knapsack problems. *AIMS Mathematics*, 10(5):11918–11938. Cc_license_type: cc_by Primary_atype: AIMS Mathematics Subject_term: Research article Subject_term_id: Research article.

Giulio Zhou and Gerasimos Lampouras. 2021. Informed sampling for diversity in concept-to-text NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2494–2509, Punta Cana, Dominican Republic. Association for Computational Linguistics.