

Small Languages, Big Models: A Study of Continual Training on Languages of Norway

David Samuel Vladislav Mikhailov Erik Velldal
Lilja Øvrelid Lucas Charpentier Andrey Kutuzov

University of Oslo, Language Technology Group

davisamu@ifi.uio.no

Abstract

Training large language models requires vast amounts of data, posing a challenge for less widely spoken languages like Norwegian and even more so for truly low-resource languages like Sámi. To address this issue, we present a novel three-stage continual training approach. We also experiment with combining causal and masked language modeling to get more flexible models. Based on our findings, we train, evaluate, and openly release a new large generative language model for Norwegian Bokmål, Nynorsk, and Northern Sámi with 11.4 billion parameters: NorMistral-11B.

1 Introduction

The development of large language models typically requires massive amounts of training data, which benefits wide-spread languages such as English, but poses a significant challenge for less widely spoken languages. Norwegian, with its two written standards Bokmål and Nynorsk, currently has approximately 24B words available in public corpora - about three orders of magnitude less than English.¹ The situation is even more challenging for Northern Sámi, which has only 40 million words currently available.²

To address this data scarcity, we propose a novel approach combining three key elements: efficient

knowledge transfer from existing models, data augmentation with related languages, and targeted up-sampling of lower-resource variants. This method enables us to train an 11.4B parameter model that achieves state-of-the-art performance across Norwegian language tasks while maintaining strong capabilities in Northern Sámi. The three main research contributions of this paper can be summarized as follows:

- 1. Novel training method for data-constrained language models** We propose a three-stage training method for efficient adaptation of existing language models to lower-resource languages. Our results demonstrate that this approach works well for adapting a Mistral model to Bokmål, Nynorsk and Northern Sámi. The model particularly excels at tasks requiring deep linguistic understanding and world knowledge in Norwegian contexts.
- 2. Flexible masked-causal model** We train a general language model that can act as a causal generative model as well as a fully-bidirectional encoder model. This approach allows it to be used as any other generative model while improving its performance at finetuning.
- 3. Fully open training artifacts** We openly release NorMistral-11B under a permissive Apache 2.0 license – <https://hf.co/norallm/normistral-11b-warm> – as well as three smaller 7-billion-parameter models, a new corpus for Northern Sámi and the training code.

In the following sections, we first describe the training corpus of NorMistral-11B in [Section 2](#); then the training and evaluation methodology of this model in [Section 3](#). In [Section 4](#), we then evaluate this model and compare it against other existing models. The following [Section 5](#) then goes into more detail by testing the training choices in our methodology.

¹While Bokmål is the main variety, roughly 15% of the Norwegian population uses Nynorsk. The two varieties are so closely related that they may be regarded as ‘written dialects’, but the lexical differences can be relatively large.

²The Sámi languages are a group of Uralic languages, of which Northern Sámi is the most widely used variant. With the number of speakers estimated to be between 15,000 and 25,000 in total across Norway, Sweden and Finland, it is still considered to be an endangered language. As the Sámi people are recognized as an Indigenous people in Norway, Sámi has status as an official language along with Norwegian.

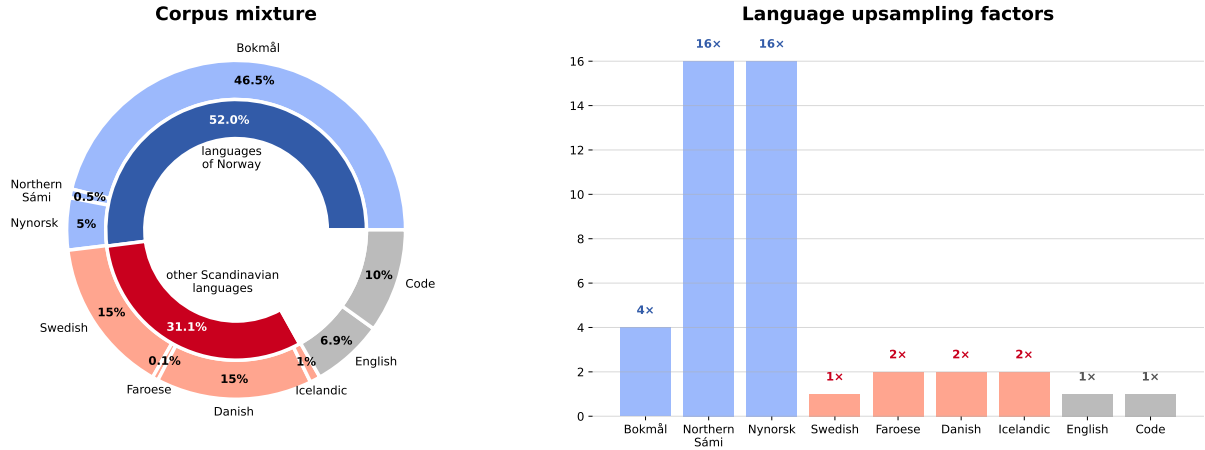


Figure 1: **Training corpus** The left figure shows the proportions of languages in the final corpus mixture, the second plot displays the upsampling factors used to get those proportions.

2 Training corpus

Our goal is to train a model for the official languages of Norway. However, this task is made difficult by the uneven distribution of these languages and the fact that there is only about 25 billion words in these languages available in the publicly accessible corpora.

2.1 Combating the data constraints

24 billion words is about three orders of magnitude less than what is currently available for English language models (Penedo et al., 2024). Assuming the Chinchilla scaling laws (Hoffmann et al., 2022), we could ‘optimally’ train only a 1-billion-parameter model on such a small dataset. However, we are able to train a much larger model due to: ① transferring knowledge from a model already trained on a large English-centric corpus; ② augmenting the corpus with other related Scandinavian languages (Danish, Swedish, Icelandic, and Faroese), as well as English and programming code (Luukkonen et al., 2024); ③ further increasing the size by repeating the data in target languages – this follows the data-constrained scaling laws by Muennighoff et al. (2023), which showed that four repetitions do not have any noticeable negative effects on the regular scaling laws. The resulting corpus of 250B tokens is then ‘optimal’ for the 11.4B parameters of our model (Hoffmann et al., 2022).³

³These measures were also motivated by previous work: for Norwegian, Liu et al. (2024a) did not report improvement for models of more than 3B size; similarly, for Finnish, Luukkonen et al. (2023) reported a decrease in performance when moving from 8B to 13B parameters.

2.2 Combating the uneven distribution

We target Norwegian and Sámi, the two official languages of Norway. Specifically, we target the *Bokmål* written variant of Norwegian with 24 billion words in our corpus, the *Nynorsk* variant with 550 million words, and *Northern Sámi*, which has only 40 million words in our corpus collection. To mitigate the large size differences, we further up-sample the two lower-resource languages (Conneau et al., 2020). To avoid overfitting on many repetitions of the same data, we follow the experimental results in Muennighoff et al. (2023) and repeat the data at most 16 times. This approach yields the final language proportions shown in Figure 1.

2.3 Data sources

Existing corpora We source most of the data from existing publicly available corpora: ① Bokmål and Nynorsk filtered from the public sources with permissive licenses from the Norwegian Colossal Corpus (NCC; Kummervold et al., 2022); ② Bokmål, Nynorsk, Swedish, Danish, and Icelandic from CulturaX (Nguyen et al., 2023); ③ Bokmål and Nynorsk from the HPLT corpus v1.2 (de Gibert et al., 2024); ④ high-quality English from FineWeb-edu (Penedo et al., 2024); ⑤ code from the high-quality part of Stack v2 (Lozhkov et al., 2024); ⑥ Faroese and Northern Sámi from Glot500 (ImaniGooghari et al., 2023); and ⑦ Northern Sámi from the SIKOR free corpus (Gjellatekno and Divvun, 2016).

Web crawl for Sámi The only exception to using existing resources is a part of the Sámi corpus. To obtain more texts for this low-resource language,

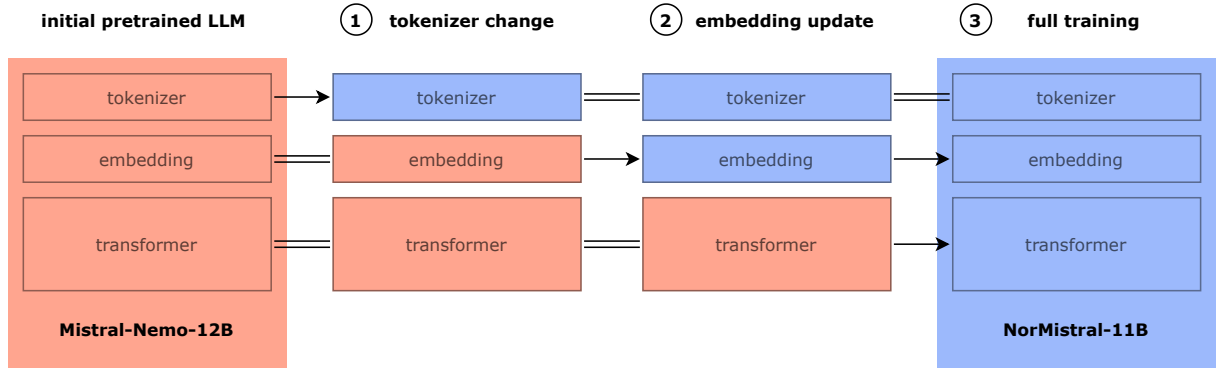


Figure 2: **Three-stage continual pretraining** We propose a novel continual pretraining pipeline consisting of ① creating a new tokenizer optimized for the training corpus, ② realigning the embedding weights to the new tokens, and ③ training the full language model. Arrows symbolize changes between stages, while double-lines represent no changes.

we conducted a web crawl through admissible web pages in Northern Sámi. The crawl was seeded from the external links of the Sámi Wikipedia and continued with a breadth-first search through web-pages that were identified as Northern Sámi using GlotLID (Kargaran et al., 2023) and that allowed crawling according to their Robots Exclusion Protocol. The raw HTML documents were converted into natural text using Trafilatura (Barbaresi, 2021). We have published the whole Sámi collection (fuzzy deduplicated at the document level) online at <https://hf.co/datasets/litg/saami-web>.

3 Training and evaluation of NorMistral

This section describes the training and evaluation pipeline of NorMistral-11B; a continually trained Mistral-Nemo-Base-2407 language model.⁴ The presented methods are evaluated later in Section 5.

3.1 Three-stage continual pretraining

Our aim is to model three lower-resource languages. To achieve this, we rely on models initially trained on more resource-rich languages and continually train them on our corpus. In order to get a model that works efficiently for the target language, we propose a novel three-stage training process, which consists of tokenizer change, embedding update, and full training (Figure 2).

Tokenizer change Before training the language model, we create a new subword tokenizer optimized for the target distribution of languages. While keeping the original tokenizer might not necessarily worsen performance, the main goal of this

Tokenizer	# tokens	NOB	NNO	SME
Mistral-Nemo-12B	131 072	1.79	1.87	2.63
NorMistral-11B	51 200	1.22	1.28	1.82

Table 1: **Tokenizer statistics** The vocabulary size and subword-to-word split ratios of different tokenizers for Bokmål (NOB), Nynorsk (NNO) and Northern Sámi (SME). Lower split ratios result in shorter subword sequences and thus in faster training and inference.

step is to improve the efficiency of training and inference. As evident from Table 1, the new tokenizer produces 30% shorter sequences on average, which translates to more than 30% faster inference time; while requiring 800 million less parameters due to the smaller vocabulary size.

The tokenizer is optimized for the entire training corpus via the greedy byte-pair encoding algorithm (BPE; Gage, 1994). We use the same tokenizer definition as the original Mistral-Nemo-12B.

Embedding update Since all tokens are changed in the previous stage, we need to update the input and output embedding weights next. While it is possible to skip this stage and simply continue training the full model, misaligned embeddings lead to a large initial loss spike, to large (essentially random) gradients for the non-embedding parameters, and thus to catastrophic forgetting (McCloskey and Cohen, 1989). Instead, we follow the tokenizer adaptation method by de Vries and Nissim (2021), aligning the embedding parameters by continually training the language model for 1 000 steps with frozen non-embedding parameters.

⁴hf.co/mistralai/Mistral-Nemo-Base-2407

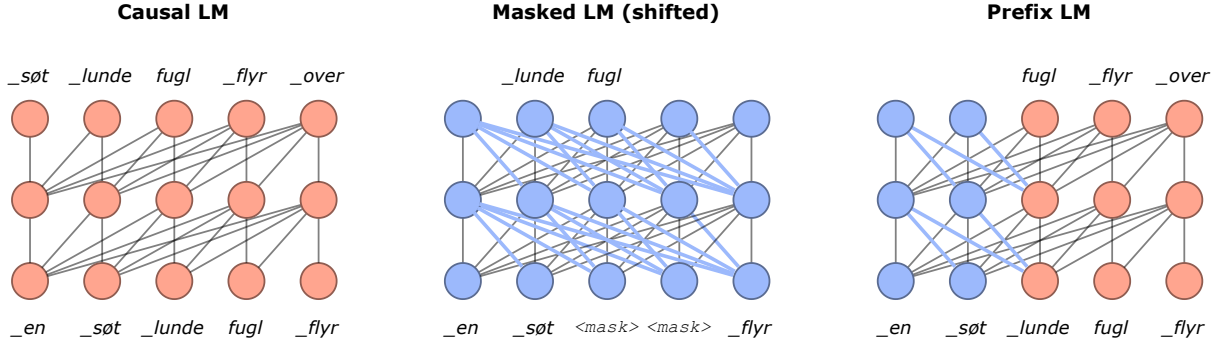


Figure 3: **Inference modes of NorMistral-11B** The hybrid masked-causal pretraining allows the model to be more flexible during inference. It can not only serve as a unidirectional causal language model (left), but also as a fully bidirectional masked language model (middle), or as a partially bidirectional prefix language model (right). The diagrams illustrate possible attention connections.

Full training After realigning the embedding vectors, we continue by unfreezing the remaining parameters and training the full model.

The transformer architecture is inherited from the original Mistral model (Jiang et al., 2023), which is based on the improved Llama architecture (Touvron et al., 2023). This mainly entails: ① pre-normalization with the RMSNorm function for improved training stability (Nguyen and Salazar, 2019; Zhang and Sennrich, 2019), ② SwiGLU activation function for improved expressive power of the feed-forward modules (Shazeer, 2020), ③ rotary positional embeddings for their ability to generalize to longer sequences (Su et al., 2021; Liu et al., 2024b), and ④ grouped-query attention for improved inference efficiency (Ainslie et al., 2023). The remaining architectural details are based on the original transformer design by Vaswani et al. (2017). The hidden dimension is set to 5 120, the intermediate one to 14 336, and there are 40 layers in total. The attention modules have 32 query heads and 8 key & value heads, each of dimension 128. There are 51 200 tokens in the subword vocabulary.

We trained the model on 250 billion tokens, which equates to 60 000 steps of $1\,024 \times 4\,096$ tokens (number of samples \times sequence length). We used the trapezoidal learning-rate schedule with a peak learning rate of $1 \cdot 10^{-4}$, 1 000 warm-up steps and 10 000 decay steps; this schedule allows for further pretraining of this model on more tokens in the future (Hägele et al., 2024). The optimization was performed using AdamW (Loshchilov and Hutter, 2019), with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and weight decay of 0.1. No dropout was applied.

The computations were conducted on 256 AMD

MI250X GPUs and took 8.5 days in total. The model was trained with a reduced bfloat16 precision and the parameters were sharded with model parallelism – pipeline parallelism of 2, tensor parallelism of 2, and a zero-redundancy optimizer (Shoeybi et al., 2020; Rasley et al., 2020; Rajbhandari et al., 2020). The overall theoretical computation cost of the training was $1.7 \cdot 10^{22}$ FLOP/s, with an average of 38% model FLOP/s utilization (MFU) on the actual hardware.

3.2 Hybrid masked-causal language modeling

While causal LMs have recently become very popular, the limited unidirectional text processing limits their learning abilities (Lv et al., 2023) and expressive power (Ewer et al., 2024); especially for finetuning (Devlin et al., 2019; Raffel et al., 2020). Furthermore, it has been recently demonstrated that fully-bidirectional masked models share the same generative abilities, but without limitations of causal models (Samuel, 2024). Following this observation, we train a model that can be flexibly used as a masked or causal language model.

Training objective We combine two training objectives during pretraining, the standard causal language modeling as well as masked next-token prediction (MNTP; Lv et al., 2023; BehnamGhader et al., 2024), a variation of masked language modeling where the next token is predicted rather than the current one (see Masked LM (shifted) in Figure 3). This has been used by Charpentier and Samuel (2024), with evidence of providing better causal modeling quality and increased finetuning performance. We trained with 90% causal LM and 10% MNTP. This ratio is rather conservative – to teach

Benchmark	Language	NorMistral-11B	NorwAI-Mistral-7B	NorwAI-Llama2-7B	normistral-7b-warm	NorGPT-3B	Viking-7B	Viking-13B	Mistral-Nemo-12B
READING COMPREHENSION									
Belebele (0-shot)	Bokmål	56.7	33.4	38.0	37.4	26.8	27.6	28.2	62.8
NorQuAD (1-shot)	Bokmål	76.7	63.0	39.2	64.8	3.0	48.4	57.1	76.5
SENTIMENT ANALYSIS									
NoReC (sentence-level; 16-shot)	Bokmål	90.5	88.6	86.0	84.9	49.7	77.9	79.2	86.9
NoReC (document-level; 1-shot)	Bokmål	91.2	81.2	79.2	82.9	51.5	80.4	86.8	89.2
COMMONSENSE REASONING									
NorCommonsenseQA (0-shot)	Bokmål	59.2	52.4	48.1	48.9	35.7	43.7	49.7	44.0
NorCommonsenseQA (0-shot)	Nynorsk	51.6	43.2	37.9	43.2	29.5	39.0	40.0	33.7
WORLD KNOWLEDGE									
NRK-Quiz-QA (0-shot)	Bokmål	63.7	55.2	52.3	57.9	33.1	44.2	51.0	47.4
NRK-Quiz-QA (0-shot)	Nynorsk	71.9	65.2	64.3	65.9	37.3	51.1	54.8	47.2
NorOpenBookQA (16-shot)	Bokmål	68.2	39.3	38.7	37.7	30.2	35.7	37.3	74.4
NorOpenBookQA (16-shot)	Nynorsk	60.0	36.7	34.4	33.3	38.9	22.2	26.6	73.3
SUMMARIZATION									
NorSumm (0-shot)	Bokmål	40.0	12.1	4.8	17.4	28.6	30.6	33.8	39.5
NorSumm (0-shot)	Nynorsk	30.6	10.6	3.1	9.9	22.4	25.8	27.4	29.4
GRAMMATICAL ERROR CORRECTION									
ASK-GEC (16-shot)	Bokmål	52.6	53.2	51.4	48.7	1.8	51.1	52.4	43.9
LANGUAGE IDENTIFICATION									
Scandinavian LID (16-shot)	Bokmål, Nynorsk, Danish, Swedish	98.6	96.2	93.4	77.5	40.3	68.9	65.6	59.5
TRANSLATION									
Tatoeba (from English; 16-shot)	Bokmål	58.8	58.7	57.9	57.2	1.8	59.7	60.0	49.6
Tatoeba (from English; 16-shot)	Nynorsk	48.0	47.4	47.4	44.7	2.6	45.6	45.6	35.7
Tatoeba (from English; 16-shot)	Northern Sámi	45.5	24.8	26.6	14.9	0.0	4.6	9.8	3.4

Table 2: **Performance of NorMistral-11B** This table compares the performance of NorMistral-11B to the performance of other dense generative models that support Norwegian. All models are evaluated with the same fully-causal in-context-learning setup without any parameter updates. The best results are in bold; higher values are always better. The performance is evaluated by accuracy (NorBelebele, NorCommonsenseQA, NorOpenbookQA & NRK-Quiz-QA), F_1 score (NorQuAD & NoReC), ROUGE-L (Lin, 2004; NorSumm), ERRANT $F_{0.5}$ (Bryant et al., 2017; ASK-GEC), and BLEU (Papineni et al., 2002; Tatoeba). We report the maximum performance score across all prompts. The random guessing baselines are 20% for NorCommonSenseQA, 25% for Belebele and NorOpenBookQA, 28% / 27% for NRK-Quiz-QA NOB / NNO, and 48.52 / 48.4 for NoReC sentence-level / document-level.

the model bidirectional processing without drifting too much from its original training objective.

3.3 Experimental Setup

We compare the performance of NorMistral-11B with publicly available LMs on a broad range of benchmarks. The evaluation is run in k -shot scenarios with $k \in \{0, 1, 16\}$ using `lm-evaluation-harness` (Gao et al., 2024). We report the maximum k for each benchmark, which depends on the availability of a training/development set for demonstration examples and on the average length of these examples.

Baselines We use seven pretrained LMs of comparable size accessed via the `transformers` library (Wolf et al., 2020) as our baselines: NorwAI-Mistral-7B, NorwAI-Llama2-7B, `normistral-7b-warm`, NorGPT-3B (Liu et al., 2024a), Viking-7B, Viking-13B, and Mistral-Nemo-12B.

Benchmarks We consider the following language understanding and generation tasks: ① reading comprehension (NorQuAD; Ivanova et al., 2023 & Belebele; Bandarkar et al., 2024), ② sentiment analysis (NoReC; Velldal et al., 2018), ③ commonsense reasoning (NorCommonsenseQA),⁵ ④ world knowledge (NRK-Quiz-QA & NorOpenBookQA),⁵ ⑤ summarization (NorSumm),⁵ ⑥ grammatical error correction (ASKGEC; Jentoft, 2023), ⑦ language identification (sourced from individual Universal Dependencies 2.14 treebanks; Nivre et al., 2016, 2020), and ⑧ translation (Tatoeba; Tiedemann, 2020).

Prompts We use multiple prompts for each benchmark to account for prompt sensitivity (Lu et al., 2024). The prompts are written by four Norwegian native speakers. The complete set of prompts is omitted due to space constraints.

4 Results

We report the evaluation results in Table 2. Overall, we see a positive indication of NorMistral-11B being a strong Norwegian model as it outperforms other evaluated systems on the majority of tasks.

Comparison to the base model Even though Mistral-Nemo-12B is an English-centric model, it performs well on the Norwegian benchmarks even before any continual pretraining. While we see a clear increase in performance after further training

when evaluated on native Norwegian datasets, there is a notable decrease in performance on Belebele (a well-known multilingual dataset) and NorOpenBookQ (an adaptation of a popular English benchmark). This aspect requires a further study, but overall, we believe that the results clearly show the benefit of three-stage continual pretraining.

Bokmål, Nynorsk and Sámi performance We evaluate the models on all target languages: Bokmål, Nynorsk and Northern Sámi. Relative to other models, the performance gains of NorMistral-11B stay consistent across these three languages.

It is possible to estimate the difference in performance on Nynorsk compared to Bokmål when focusing on NorSumm, a dataset that is perfectly balanced and parallel for the two variants of Norwegian. The substantially higher score for Bokmål indicates that the much smaller amount of Nynorsk in the training corpus (even after upsampling) propagates into the downstream performance.

The results on the English-to-Sámi translation suggest that our model was able to learn aspects of this language even though it made only 0.5% of the training corpus. However, any stronger claim about the level of understanding of Sámi would require a substantially more robust benchmarking suite than what is currently available.

4.1 Using NorMistral in practice

Large language models can be utilized in many different ways. We used the most direct and straightforward one for comparing Norwegian models – in-context learning – but there is a broader spectrum of methods with varying complexity-to-performance trade-offs. We evaluate the most common methods in Table 3 using NorQuAD:

In-context learning This is the most popular method of using large language models, mostly because it does not require any further training (Brown et al., 2020). Using just one sample from the training set as a demonstration can substantially improve the output quality on NorQuAD. More demonstrations can improve the performance further, but at the cost of reduced inference speed.

Quantization In order to reduce the large memory cost of large language models, a popular method is reducing the precision of their parameters. Specifically, we test 8-bit and 4-bit quantization (Dettmers et al., 2022; Dettmers and Zettlemoyer, 2023). There is no noticeable decrease of

⁵Anonymized URL, the resources are under review.

performance on NorQuAD when lowering the precision from the original 16 bits. Note that some GPUs can also increase their throughput at the lowered precision.

Full finetuning The best-performing strategy is to do supervised finetuning of all learnable parameters. This method is also the most difficult to set up, the large memory requirements necessitate distributed training with some model sharding. However, after finetuning, this method clearly outperforms all other ones without any additional cost. Interestingly, when finetuned with partially-bidirectional attention masks (as a prefix LM), the model even exceeds the estimated human performance on NorQuAD – 91.1 F₁ score and 78.1 EM accuracy (Ivanova et al., 2023).

LoRA finetuning Further training NorMistral on a downstream task is more demanding, but it is the preferred way for achieving the best performance – as long as there is a sizeable training set available. Low-rank adaptation (LoRA) reduces the computational cost of finetuning by freezing all original model parameters and training only small low-rank adaptors (Hu et al., 2022). The resulting model is 10 F₁ percentage points better than the best few-shot prompt while running almost 4 times faster because of shorter context lengths. Because of its hybrid pretraining (Section 3.2), NorMistral can also be finetuned as a partially-bidirectional prefix language model, which further improves its performance by 1.4 points without any additional computational cost.

5 Methodological comparisons

We have conducted an initial comparative study of different training methods before settling on the pretraining process from Section 3 and training NorMistral-11B. The results are presented in Table 4, where different models are evaluated on a representative subset of available Norwegian benchmarks: extractive question answering (1-shot NorQuAD), binary sentence-level polarity classification (16-shot NoReC), world knowledge (0-shot NRK-Quiz-QA) and machine translation (16-shot English-to-Bokmål Tatoeba).

Architectural choice There are many promising improvements of the original GPT neural architecture (Radford et al., 2018) – we considered two recent and well-studied architectures: BLOOM by Scao et al. (2022) and Llama by Touvron et al.

Method	F1	EM	Runtime train / eval
0-shot (causal)	59.7	33.5	0 / 6 min
1-shot (causal)	76.7	55.3	0 / 8 min
8-shot (causal)	79.6	60.8	0 / 23 min
0-shot (4-bit, causal)	59.2	33.5	0 / 6 min
0-shot (8-bit, causal)	59.1	33.7	0 / 6 min
Full finetuning (causal)	90.4	79.2	57 / 6 min
Full finetuning (prefix)	92.2	80.3	57 / 6 min
LoRA finetuning (causal)	89.9	77.1	18 / 6 min
LoRA finetuning (prefix)	91.3	79.0	18 / 6 min

Table 3: **Evaluation methods** NorMistral-11B can be flexibly used in many different ways for solving downstream tasks. We compare them on NorQuAD, a dataset for extractive question answering. NorMistral can be finetuned as a standard causal language model and also as a partially bidirectional prefix language model. We also show the total training and evaluation time for each method (run on AMD MI250X GPUs).

(2023), which is also used for training the Mistral models (Jiang et al., 2023). We adopted the training hyperparameters suggested by the respective papers and trained two models with 7 billion parameters on the same Norwegian corpus and with the same Norwegian tokenizer. Table 4 clearly shows that the Llama architecture is preferred for our training corpus and Norwegian benchmarks.

From scratch vs. warm-starting The central research question of this paper is how to train a good large language model for relatively small languages. Here we test our proposed three-stage continual pretraining and compare it against a model trained from scratch. For a fair comparison, we train two 7-billion-parameter models on the same corpus, and with the same architecture and tokenizer. Note that we do not consider existing methods that do not adapt the subword vocabulary – like simple continual training or adapter tuning (Yong et al., 2023) – because they necessarily lead to inefficient inference (Table 1). The results in Table 4 demonstrate that the knowledge transfer from an English-centric model works and the model is able to be adapted to new languages.

Hybrid masked-causal modeling Interestingly, we do not observe an overall increase in perfor-

mance after training with the ‘dual’ training objective, as opposed to the observations by [Charpentier and Samuel \(2024\)](#). However, we believe that this can be explained by continued training – the hybrid masked-causal training is used for a negligible number of steps compared to the fully-causal pre-training of the base Mistral model.

Number of training steps Finally, we compare the performance of model checkpoints saved at different points of training. We can make several observations from the results: ① they confirm the data-scaling laws by [Muennighoff et al. \(2023\)](#) as the model continues to improve even after (at least) four repetitions of the Norwegian data; ② tokenizer adaptation (the first two stages of our training method) is a simple and efficient way of adapting a model to a new language without losing performance; ③ the three-stage continual pretraining does not affect all downstream tasks equally – while it usually leads to monotonical improvement, there are some tasks (NorQuAD) that experience an initial decrease in performance. Further investigation is needed to determine if this drop is significant and if it can be avoided by a more careful switch to a new language distribution at the start of training.

6 Related work

When it comes to creating openly available generative decoder-only models for Norwegian, most of the main efforts are listed in section 3.3 and used in our experiments. However, one other notable mention is NB-GPT-J-6B – a fine-tuned version of the English GPT-J-6B model.⁶ Released by the National Library of Norway in 2022, it was the first GPT-like model trained for Norwegian.

There have also been several efforts on developing smaller transformer models, e.g., based on the BERT encoder architecture and the T5 encoder-decoder architecture. This includes the families of models known as NorBERT ([Kutuzov et al., 2021](#); [Samuel et al., 2023](#)), NorT5 ([Samuel et al., 2023](#)), NB-BERT ([Kummervold et al., 2021](#)), and North-T5, all available in different parameter sizes.⁷

As for Northern Sámi, [Paul et al. \(2024\)](#) has recently experimented with targeting this language. However, their models have not been published nor did they evaluate them on any downstream tasks; we are thus not able to compare them to our model.

⁶<https://huggingface.co/NbAiLab/nb-gpt-j-6B>

⁷<https://huggingface.co/north>

Training method	NorQuAD 1-shot	NoReC 16-shot	NRK 0-shot	Tatoeba 16-shot
TRANSFORMER ARCHITECTURE				
BLOOM	43.6	67.6	44.6	52.2
Llama / Mistral	43.7	80.3	48.2	53.4
CONTINUAL TRAINING				
init. from scratch	43.7	80.3	48.2	53.4
three-stage continual	64.8	84.9	57.9	57.2
HYBRID TRAINING OBJECTIVE				
causal-only	67.0	86.0	59.0	58.8
hybrid masked-causal	69.3	87.5	55.4	58.2
TRAINING STEPS				
0 steps (base model)	76.5	86.9	47.4	49.6
0 steps (adapted tokenizer)	73.5	89.4	44.2	51.4
10,000 steps	69.3	87.5	55.4	58.2
20,000 steps	70.5	89.2	57.7	58.8
30,000 steps	66.2	82.3	59.0	58.5
40,000 steps	68.5	87.0	61.1	58.9
50,000 steps	70.4	88.7	60.2	58.7
60,000 steps	76.7	90.5	63.7	58.8

Table 4: **Comparison of training methods** The methods are compared on NorQuAD with F₁ score, sentence-level Bokmål NoReC with F₁ score, Bokmål NRK-Quiz-QA with accuracy, and on English-to-Bokmål Tatoeba with BLEU.

7 Conclusion

We presented NorMistral-11B, a new large language model for Norwegian Bokmål, Nynorsk, and Northern Sámi. We proposed a novel three-stage continual pretraining approach that efficiently adapts existing models to other languages while maintaining high performance and increasing their inference speed. This approach involves training a new tokenizer, realigning embedding weights, and then training the full model. We also demonstrated the benefits of hybrid masked-causal pretraining, which allows the model to be used flexibly as either a causal or bidirectional model. Our extensive evaluation shows that NorMistral-11B achieves the state-of-the-art performance across a wide range of Norwegian tasks, while also showing promising results for Northern Sámi. This suggests that our approach could be beneficial for developing large language models for other smaller languages. To facilitate further research and development, we have released NorMistral-11B, the three 7B models trained for Section 5, training code, and a new Northern Sámi corpus.

Limitations

No instruction-tuning It is important to note that NorMistral-11B is not instruction-tuned and we did not perform any human preference alignment, nor any addition of guardrails or safety testing; thus, the model cannot be used as an out-of-the-box chatbot. The model is intended to be used by other researchers and potentially further instruction-finetuned on a suitable instruction dataset.

Limitations of the base language model Since NorMistral-11B is continually pretrained on the existing Mistral-Nemo-12B weights, the model is to some extent dependent on the training data of the original Mistral model. The exact composition of this training data is not known, which to some extent limits more detailed studies of this model.

Computational cost As mentioned in [Section 3](#), training NorMistral-11B took more than 52 000 GPU/hours. This is a significant amount. We have not yet estimated the CO₂ footprint of the full training, but it was conducted on the LUMI supercomputer which is powered exclusively with renewable electricity and deployed in one of the most eco-efficient data centers in the world.⁸

Evaluation of Northern Sámi knowledge Finally, our evaluation for Northern Sámi is limited to English-Sámi translation, which is obviously insufficient. Unfortunately, we lack more advanced or diverse benchmarks for low-resource languages like this one. We hope to see further development in this direction by the NLP community.

References

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. [GQA: Training generalized multi-query transformer models from multi-head checkpoints](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.
- Adrien Barbaresi. 2021. [Trafilatura: A web scraping library and command-line tool for text discovery and extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [LLM2vec: Large language models are secretly powerful text encoders](#). In *First Conference on Language Modeling*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [Gpt or bert: why not both?](#)
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [GPT3.int8\(\): 8-bit matrix multiplication for transformers at scale](#). In *Advances in Neural Information Processing Systems*.
- Tim Dettmers and Luke Zettlemoyer. 2023. [The case for 4-bit precision: k-bit inference scaling laws](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

⁸<https://www.lumi-supercomputer.eu/sustainable-future/>

- 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.
- Ethan Ewer, Daewon Chae, Thomas Zeng, Jinkyu Kim, and Kangwook Lee. 2024. [ENTP: Encoder-only next token prediction](#).
- Philip Gage. 1994. [A new algorithm for data compression](#). *The C Users Journal archive*, 12:23–38.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [A framework for few-shot language model evaluation](#).
- 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.
- Giellatekno and Divvun. 2016. [SIKOR North Saami corpus](#).
- Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben allal, Leandro Von Werra, and Martin Jaggi. 2024. [Scaling laws and compute-optimal training beyond fixed training durations](#). In *Workshop on Efficient Systems for Foundation Models II @ ICML2024*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. [An empirical analysis of compute-optimal large language model training](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. [Glot500: Scaling multilingual corpora and language models to 500 languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.
- Sardana Ivanova, Fredrik Andreassen, Matias Jentoft, Sondre Wold, and Lilja Øvrelid. 2023. [NorQuAD: Norwegian question answering dataset](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 159–168, Tórshavn, Faroe Islands. University of Tartu Library.
- Matias Jentoft. 2023. [Grammatical error correction with byte-level language models](#). Master’s thesis, University of Oslo.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *ArXiv*, abs/2310.06825.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. [The Norwegian colossal corpus: A text corpus for training large Norwegian language models](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.
- Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Bryggjeld. 2021. [Operationalizing a national digital library: The case for a Norwegian transformer model](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. [Large-scale contextualised language modelling for Norwegian](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summariza-*

- tion Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Peng Liu, Lemei Zhang, Terje Farup, Even W. Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2024a. [NLEBench+NorGLM: A comprehensive empirical analysis and benchmark dataset for generative language models in Norwegian](#).
- Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. 2024b. [Scaling laws of RoPE-based extrapolation](#). In *The Twelfth International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostafa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2024. [StarCoder 2 and The Stack v2: The next generation](#).
- Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. [How are prompts different in terms of sensitivity?](#) 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 5833–5856, Mexico City, Mexico. Association for Computational Linguistics.
- Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Vaino Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. [Poro 34b and the blessing of multilinguality](#). *ArXiv*, abs/2404.01856.
- Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osmo Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. [FinGPT: Large generative models for a small language](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.
- Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhang Chen, Ji-Rong Wen, and Rui Yan. 2023. [Are we falling in a middle-intelligence trap? an analysis and mitigation of the reversal curse](#). *CoRR*, abs/2311.07468.
- Michael McCloskey and Neal J. Cohen. 1989. [Catastrophic interference in connectionist networks: The sequential learning problem](#). In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.
- Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. [Scaling data-constrained language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- 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](#).
- Toan Q. Nguyen and Julian Salazar. 2019. [Transformers without tears: Improving the normalization of self-attention](#). In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. 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.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ronny Paul, Himanshu Buckchash, Shantipriya Parida, and Dilip K. Prasad. 2024. [Towards a more in-](#)

- clusive ai: Progress and perspectives in large language model training for the sámí language. *ArXiv*, abs/2405.05777.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben 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](#).
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#). *OpenAI Blog*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. [ZeRO: memory optimizations toward training trillion parameter models](#). In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.
- David Samuel. 2024. [BERTs are generative in-context learners](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. [NorBench – a benchmark for Norwegian language models](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenceon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz’alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar’ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nuru-laqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L’opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesh Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Rana, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Francois Lavall’ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramanian, Aur’elie N’ev’eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Nanyang Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly

- Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynek, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrmann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myung-sun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, Patrick Haller, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aaronsiri, Srishiti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xiao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. [BLOOM: A 176b-parameter open-access multilingual language model](#). *ArXiv*, abs/2211.05100.
- Noam M. Shazeer. 2020. [GLU variants improve transformer](#). *ArXiv*, abs/2002.05202.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. [Megatron-LM: Training multi-billion parameter language models using model parallelism](#). Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. [Roformer: Enhanced transformer with rotary position embedding](#).
- Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- 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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. [NoReC: The Norwegian review corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Wietse de Vries and Malvina Nissim. 2021. [As good as new. how to successfully recycle English GPT-2 to make models for other languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Zheng Xin Yong, Hailey Schoelkopf, Niklas Muenighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vasilina Nikoulina. 2023. [BLOOM+1: Adding language support to BLOOM for zero-shot prompting](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Biao Zhang and Rico Sennrich. 2019. [Root Mean Square Layer Normalization](#). In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada.