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   Stephan Oepen

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 Northern Sámi. To address this issue, we present a novel three-stage continual training approach that substantially improves the downstream performance together with the inference efficiency for the target languages. Based on our findings, we train, evaluate, and openly release a new 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,[1] currently has approximately 24B words available in our filtered text collection – about three orders of magnitude less than English (Penedo et al., 2024). The situation is even more challenging for Northern Sámi with only 40 million words available.[2]

To address this data scarcity, we propose a novel approach combining three key elements: knowledge transfer from existing models, data augmentation with related languages, and targeted upsampling. This method enables us to train an 11.4B-parameter model that achieves state-of-the-art performance across Norwegian language tasks while obtaining 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. Our model achieves the *state-of-the-art* performance on tasks requiring deep linguistic understanding and world knowledge in Norwegian contexts – while being more than 30% faster than the original Mistral model on Norwegian inputs.

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 embedding model. This approach allows it to be used as any other generative model while allowing future usage as a finetuned encoder model.

3. **Truly open source**   We openly release NorMistral-11B under a permissive Apache 2.0 license – `https://hf.co/norallm/normistral-11b-warm` – as well as three smaller 7B-parameter models and a new corpus for Northern Sámi. The model is trained on fully transparent corpora and evaluated on a robust set of prompts that are included in the paper. The training and evaluation scripts are available at `https://github.com/ltgoslo/norallm`.

---

[1]While 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.

[2]The 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.
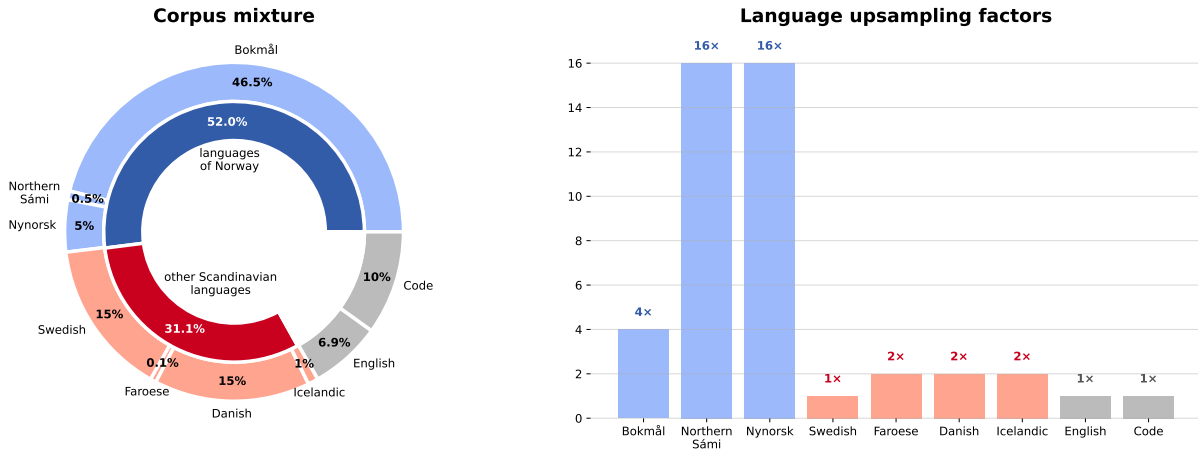
Figure 1: **Language composition of training corpus** The left figure shows the proportions of languages in the final corpus mixture, with the target languages of Norway in blue, related languages in red, and other data sources in gray. The right figure then displays the upsampling factors used to get the aforementioned proportions.

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.Section 6 describes previous works that inspired this paper. Additional appendices then offer further analyses and a detailed description of the evaluation setup.

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 24 billion words in these languages available in the publicly accessible high-quality corpora (see below).

2.1 Combating the data constraints

24B 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 non-unique tokens is then 'compute-optimal' for the 11.4B parameters of our model (Hoffmann et al., 2022). In the previous work, the NorGPT-23B LLM trained on the available Norwegian data by Liu et al. (2024a) did not outperform smaller 3B models. Similarly, for Finnish, Luukkonen et al. (2023) reported a decrease in performance when moving from 8B to 13B parameters on a similarly sized corpus. These observations support our decision not to move beyond the 11B size.

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 0.5 billion words, and *Northern Sámi*, which has only 40 million words in our corpus collection. To mitigate the large size differences, we further upsample 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.
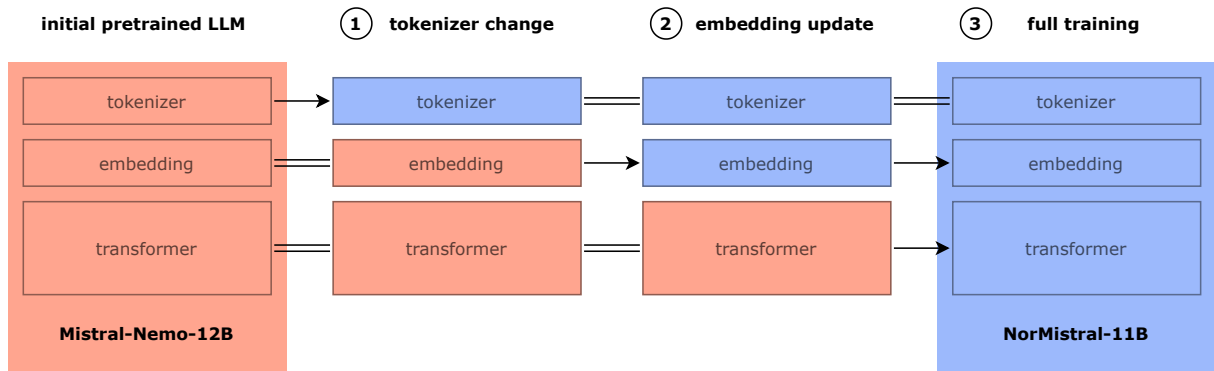
Figure 2: **Three-stage continual pretraining** We propose a novel continual pretraining pipeline consisting of ① creating a new tokenizer optimized for the training corpus, ② realigning the embedding weights to the new tokens, and ③ training the full language model. Arrows symbolize changes between stages, while double-lines represent no changes.

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 *Mímir Core* corpus from de la Rosa et al. (2025), which itself consists of the Norwegian Colossal Corpus (NCC; Kummervold et al., 2022), CulturaX (Nguyen et al., 2023), and the HPLT corpus v1.2 (de Gibert et al., 2024); ② Bokmål, Nynorsk, Swedish, Danish, and Icelandic from *CulturaX* (Nguyen et al., 2023); ③ high-quality English from *FineWeb-edu* (Penedo et al., 2024); ④ code from the high-quality part of *Stack v2* (Lozhkov et al., 2024); ⑤ Faroese and Northern Sámi from *Glot500* (ImaniGooghari et al., 2023); and ⑥ Northern Sámi from the *SIKOR free corpus* (Giellatekno 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, 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 webpages 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 web-crawled texts (fuzzy deduplicated at the document level) online at `https://hf.co/datasets/ltg/saami-web`. In total, it contains about 13 million whitespace-separated words.

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.[3] 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).

Stage 1: 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 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. We measure the inference speed-up on a downstream task in Appendix A, confirming the theoretical benefits.

The tokenizer is optimized for the entire training corpus via the greedy byte-pair encoding algorithm (BPE; Gage, 1994). We use the same tokenizer

---

[3]Available on HuggingFace at `hf.co/mistralai/Mistral-Nemo-Base-2407`

definition as the original Mistral-Nemo-12B: byte-level BPE tokenizer without any Unicode normalization and with a fairly complex pretokenizer regular expression. The pretokenizer splits numbers into individual digits as in Chowdhery et al. (2024). Note that the tokenization is completely lossless and reversible as out-of-vocabulary characters can be split into individual UTF-8 bytes that are always in-vocabulary as atomic tokens.

Stage 2: 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.

The initial token embeddings are transferred from the original embedding matrix (Gu et al., 2018; Wang et al., 2019). Since we use the same tokenizer type as the original Mistral model, many tokens are present in both vocabularies; the embeddings for these are initialized by as direct copies of the original vectors. Tokens not present in the original vocabulary are tokenized (with the original tokenizer) to obtain sub-tokens within the vocabulary; the embedding vectors are then initialized by taking the average of the sub-token embeddings.

Stage 3: 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 (Liu et al., 2024b; Su et al., 2021), and ④ grouped-query attention for improved inference efficiency (Ainslie et al., 2023). The remaining architectural details are based on

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.

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 used 55 000 GPU hours in total – which equals to 8.5 days of runtime on the distributed setup. 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 (Rajbhandari et al., 2020; Rasley et al., 2020; Shoeybi et al., 2020). The overall theoretical computation cost of the training was $1.7 \cdot 10^{22}$ FLOPs, 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
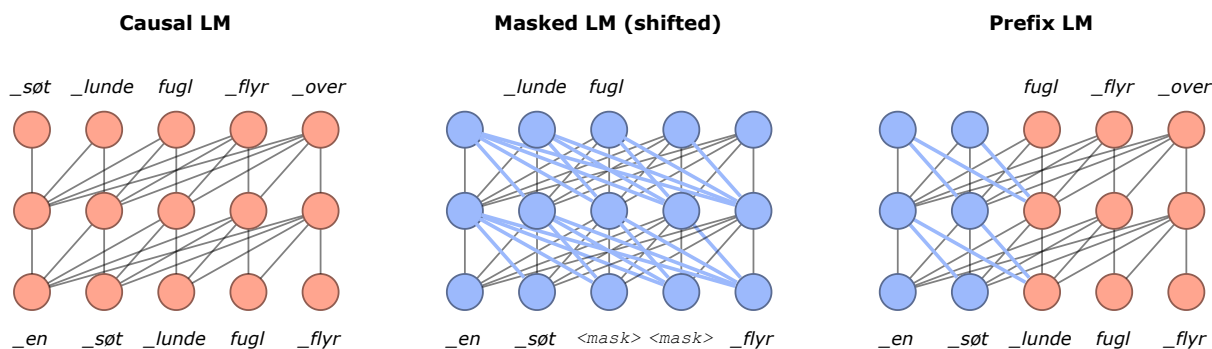
Figure 3: **Inference modes of NorMistral-11B** The hybrid masked-causal pretraining allows the model to be more flexible during inference. It can not only serve as a unidirectional causal language model (left), but also as a fully bidirectional masked language model (middle), or as a partially bidirectional prefix language model (right). The diagrams illustrate possible attention connections.

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 one as well as masked next-token prediction (MNTP; BehnamGhader et al., 2024; Lv et al., 2023), 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 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 using NorEval,[4] an open-source framework for evaluating Norwegian generative LMs built on `lm-evaluation-harness` (Gao et al., 2024). The evaluation is run in k-shot scenarios with $k \in \{0, 1, 16\}$ on ten benchmarks. We report the maximum k for each benchmark across a set of prompts, 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` li-

---

brary (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 The models are evaluated only on datasets created by native speakers. 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; Mikhailov et al., 2025), ④ world knowledge (NRK-Quiz-QA & NorOpenBookQA; Mikhailov et al., 2025), ⑤ summarization (NorSumm; Touileb et al., 2025), ⑥ grammatical error correction (ASK-GEC; Jentoft, 2023), ⑦ language identification (SLIDE; `https://github.com/ltgoslo/slide`), and ⑧ translation (Tatoeba; Tiedemann, 2020). NorEval provides a set of task-specific 4–6 prompts written by Norwegian native speakers, which allows to account for prompt sensitivity (Lu et al., 2024). More details about each task with a complete list of prompts are given in Appendix B.

4 Results

We report the aggregated evaluation results in Table 2 and fine-grained evaluation results in Appendix B. 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

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	**61.0**	54.2	49.7	51.3	34.7	44.9	51.1	46.9
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	77.9	52.3	52.3	49.0	29.5	48.7	47.0	**86.9**
NorOpenBookQA (16-shot)	Nynorsk	77.8	45.6	38.9	41.1	34.4	27.8	36.7	**86.7**
SUMMARIZATION									
NorSumm (0-shot)	Bokmål	**45.0**	12.2	10.7	16.5	33.8	31.9	36.3	44.9
NorSumm (0-shot)	Nynorsk	**32.6**	10.3	10.4	8.6	24.3	25.7	28.8	30.9
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									
SLIDE (16-shot)	Bokmål, Nynorsk, Danish, Swedish	**98.2**	95.7	93.5	98.1	40.3	77.2	84.4	87.3
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	**50.4**	27.5	28.5	18.5	0.0	7.8	11.6	6.5

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 (Belebele, 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), accuracy (SLIDE), 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 48.5% / 48.4% for NoReC sentence-level / document-level.

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) limits the downstream performance on this language variant.

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	F_1	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). We use the two standard metrics for extractive question answering: F_1 score and exact-match accuracy (EM).

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 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 F_1 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 F_1 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 (Scao et al., 2023) and Llama (Touvron et al., 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 Norwegian corpus (the Norwegian Colossal Corpus by Kummervold et al., 2021), and with the same architecture and tokenizer. Note that we do not consider existing methods that do not adapt the

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 F_1 score, sentence-level Bokmål NoReC with F_1 score, Bokmål NRK-Quiz-QA with accuracy, and on English-to-Bokmål Tatoeba with BLEU.

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 performance 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 negligable number of steps compared to the fully-causal pretraining 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

Norwegian language models   There have been several prior efforts on creating language models for Norwegian. 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.[5] Released by the National Library of Norway in 2022, it was the first large generative language model trained for Norwegian.

There have also been several efforts on developing smaller transformer models, e.g., based on the BERT encoder architecture (Devlin et al., 2019) and the T5 encoder-decoder architecture (Raffel et al., 2020). The NorBERT family of models were first released by Kutuzov et al. (2021) and have by now reached their third iteration of releases (Samuel et al., 2023) and come in several different sizes; ranging fron 15M parameters for the XS model to 323M for NorBERT3 Large. Samuel et al. (2023) also introduced the NorT5 family of models, ranging 32M to 808M parameters. Whereas the above-mentioned models where all trained from scratch for Norwegian, Kummervold et al. (2021) trained NB-BERT (base and large) by fine-tuning the pre-trained mBERT model on Norwegian data, also reusing the tokenizer. A similar approach was followed for the North-T5 models.[6]

Language models for Northern Sámi   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.

Continual training techniques   Adaptation of pretrained language models to new domains by continual training has a long history (Gururangan et al., 2020). Our three-stage continual pretraining is designed specifically for adapting language models to a new language – by entirely replacing the original tokenizer, we can get an efficient model (by compressing the textual input into a short sequence of tokens) without the need of any extra parameters. Simple continual pretraining works well performance-wise but the training and inference computation cost is high (Ibrahim et al., 2024). A substantially more efficient approach is to introduce a new tokenizer and replace the embedding layers (first two stages of our approach), as proposed by (Marchisio et al., 2023; de Vries and Nissim, 2021). Similarly, Csaki et al. (2023) only use the first and last stage of our method – they extend the vocabulary by 5 000 new tokens and then train the full model. On the other hand, Kim et al. (2024) pursue a more careful approach, the most similar to our training method. They first extend the subword vocabulary with extra tokens and then meticulously train the new and old parameters in eight subsequent stages.

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 at **https://github.com/ltgoslo/norallm.**

---

[5]`https://huggingface.co/NbAiLab/nb-gpt-j-6B`
[6]`https://huggingface.co/north`

Limitations

Limitations of the base language model   Since NorMistral-11B is continually pretrained on the existing Mistral-Nemo-12B weights, the model is partially 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. Specifically, the original model might have been trained on contaminated data, which could explain its high-scores on well-known evaluation tasks such as Belebele.

Computational cost   As mentioned in Section 3, training NorMistral-11B took more than 55 000 GPU/hours. This is a significant amount. We have not yet estimated the CO_2 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.[7]

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.

Acknowledgments

---

[7]https://www.lumi-supercomputer.eu/sustainable-future/

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?

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi,

Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2024. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

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.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. In *Workshop on Efficient Natural Language and Speech Processing (ENLSP) at NeurIPS 2023*.

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 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.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

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.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models.

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 Brygfjeld. 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 Summarization 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5543–5560, Miami, Florida, USA. Association for Computational Linguistics.

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, Mostofa 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 Hatanpaa, 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, Osma 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, Yuhan 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.

Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid, and Erik Velldal. 2024. It's difficult to be neutral – human and LLM-based sentiment annotation of patient comments. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 8–19, Torino, Italia. ELRA and ICCL.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. 2023. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5474–5490, Toronto, Canada. Association for Computational Linguistics.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Vladislav Mikhailov, Petter Mæhlum, Victoria Ovedie Chruickshank Langø, Erik Velldal, and Lilja Øvrelid. 2025. A collection of question answering datasets for Norwegian. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallinn, Estonia.

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.

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 inclusive AI: Progress and perspectives in large language model training for the sámi 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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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.

Javier de la Rosa, Vladislav Mikhailov, Lemei Zhang, Freddy Wetjen, David Samuel, Peng Liu, Rolv-Arild Braaten, Petter Mæhlum, Magnus Breder Birkenes, Andrey Kutuzov, Tita Enstad, Svein Arne Brygfjeld, Jon Atle Gulla, Stephan Oepen, Erik Velldal, Wilfred Østgulen, Lilja Øvrelid, and Aslak Sira Myhre. 2025. The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*.

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, Ellie 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, and et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model.

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.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Samia Touileb, Vladislav Mikhailov, Marie Ingeborg Kroka, Øvrelid Lilja, and Erik Velldal. 2025. Benchmarking abstractive summarisation: A dataset of human-authored summaries of norwegian news articles. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, Tallin, Estonia.

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, Ł ukasz 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.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. 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 Muennighoff, 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 Vassilina 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.

A   Inference efficiency of three-stage continual pretraining

In order to provide evidence for our claim that three-stage continual pretraining is necessary to increase the inference efficiency, we measure the actual inference speed on downstream tasks. Note that we specifically focus on the first stage of our pretraining recipe – creating a brand new tokenizer for the target domain. Since in-context-lerning evaluation can be done in two modes – classification or generation – we measure the inference speed on both of them. Since the model quality might influence the number of generated tokens, we constrain the generation to only output tokens from the gold answers.

Speedup due to a new tokenizer   We compare the speed of the original language model, Mistral-Nemo-12B, with the speed of our model that was initialized from it, NorMistral 11B. The results in Table 5 show that completely changing the tokenizer results in a noticeable speed up in both tests.

Other evaluated models   For completeness, the inference speed of other models used in this paper are included as well; even though they have different number of non-embedding parameters or even completely different architectures. These additional measurements also show the benefit of replacing the entire vocabulary instead of only extending it with additional tokens.

SENTENCE-LEVEL NOREC (16-SHOT)

Model	Vocabulary	Note	Average length	Time / sample	Slowdown
NorMistral-11B	51 200	*our new Norwegian tokenizer*	522 tokens	0.23 s	1×
Mistral-Nemo-12B	131 072	*original English-centric tokenizer*	640 tokens	0.30 s	1.30×
NorwAI-Mistral-7B	67 993	*extends an English tokenizer*	591 tokens	0.18 s	0.78×
NorwAI-Llama2-7B	67 993	*extends an English tokenizer*	591 tokens	0.15 s	0.65×
NorMistral-7B-warm	32 768	*new Norwegian tokenizer*	569 tokens	0.17 s	0.74×
NorGPT-3B	64 000	*new Norwegian tokenizer*	552 tokens	0.08 s	0.35×
Viking-7B	131 072	*new Nordic tokenizer*	512 tokens	0.14 s	0.61×
Viking-13B	131 072	*new Nordic tokenizer*	512 tokens	0.26 s	1.13×

NORQUAD (8-SHOT)

Model	Vocabulary	Note	Average length	Time / sample	Slowdown
NorMistral-11B	51 200	*our new Norwegian tokenizer*	4 909 tokens	3.10 s	1×
Mistral-Nemo-12B	131 072	*original English-centric tokenizer*	6 171 tokens	4.13 s	1.33×
NorwAI-Mistral-7B	67 993	*extends an English tokenizer*	5 206 tokens	2.28 s	0.73×
NorwAI-Llama2-7B	67 993	*extends an English tokenizer*	5 206 tokens	2.10 s	0.68×
NorMistral-7B-warm	32 768	*new Norwegian tokenizer*	5 012 tokens	2.22 s	0.71×
NorGPT-3B	64 000	*new Norwegian tokenizer*	4 604 tokens	—	—
Viking-7B	131 072	*new Nordic tokenizer*	4 810 tokens	1.93 s	0.62×
Viking-13B	131 072	*new Nordic tokenizer*	4 810 tokens	—	—

Table 5: **Inference speed with different tokenization strategies**   We measure the average sequence length that a model needs to precess per sample, as well as the average processing time per sample. These statistics are measured on a classification task (NoReC) as well as on a generative task (NorQuAD). Some models were not able to process the dataset, either because of not supporting long-enough input sequences or because of out-of-memory errors.

B Evaluation details

We provide a complete description of the evaluation design in this appendix. We provide inference details and prompts as well as full non-aggregated results here. Further information can be found at `https://github.com/ltgoslo/noreval` and `https://github.com/ltgoslo/norallm`.

B.1 Belebele

Belebele is a reading comprehension benchmark for evaluating the natural language understanding of language models (Bandarkar et al., 2024).

Inference setup The model is given a test example formatted according to a prompt template and ranks the answer candidates based on their probabilities. The most probable answer candidate is selected as the resulting answer.

Performance metric There are four possible answers for each passage-question pair. We measure the performance with a simple accuracy.

Prompt templates We used the following five prompt templates from NorEval.

Prompt A:

```
1   Tekst: {$passage}
2   Spørsmål: {$question}
3   A: {$answer_1}
4   B: {$answer_2}
5   C: {$answer_3}
6   D: {$answer_4}
7   Svar: {$prediction:A/B/C/D}
```

Prompt B:

```
1   Bakgrunn: {$passage}
2   Spørsmål: {$question}
3   Svaralternativer:
4   - {$answer_1}
5   - {$answer_2}
6   - {$answer_3}
7   - {$answer_4}
8   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt C:

```
1   {$question}
2   Hvilket av følgende mulige svar er det riktige?
3   A: {$answer_1}
4   B: {$answer_2}
5   C: {$answer_3}
6   D: {$answer_4}
7   Svar: {$prediction:A/B/C/D}
```

Prompt D:

```
1   Svar på følgende spørsmål: {$question}
2   Svaret skal baseres på følgende tekst:
3   {$passage}
4   Velg et svar fra denne listen:
5   - {$answer_1}
6   - {$answer_2}
7   - {$answer_3}
8   - {$answer_4}
9   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt E:

```
1   {$passage}
2
3   {$question}
4
5   A: {$answer_1}
6   B: {$answer_2}
7   C: {$answer_3}
8   D: {$answer_4}
9
10  Er det riktige svaret A, B, C, eller D? {$prediction:A/B/C/D}
```

Full results The complete evaluation results on Belebele (Bokmål) are given in Table 6. Note that the random-guessing baseline on this task achieves accuracy of 25%.

	0-shot				
Prompt template	A	B	C	D	E
NorMistral-11B	45.2	56.7	32.6	**31.1**	22.8
NorwAI-Mistral-7B	29.6	33.4	27.2	24.8	22.9
NorwAI-Llama2-7B	29.6	38.0	26.4	25.9	21.2
NorMistral-7B-warm	22.9	37.4	23.2	27.0	23.0
NorGPT-3B	22.2	26.8	22.9	25.7	22.9
Viking-7B	23.8	27.6	25.4	26.1	22.8
Viking-13B	27.3	27.3	28.2	25.1	22.8
Mistral-Nemo-12B	**60.6**	**62.8**	**38.1**	28.4	**27.0**

Table 6: **Complete results on Belebele question answering (Bokmål)** We show the detailed results for each evaluated model and prompt template. The best results for each column are boldfaced, the overall best result is highlighted in blue.

B.2 NorQuAD

The second benchmark for reading comprehension, NorQuAD by Ivanova et al. (2023), follows the scheme of extractive question-answering from SQuAD (Rajpurkar et al., 2016).

Inference setup The model is given a test example formatted according to a prompt template and generates an answer via the greedy-search decoding strategy.

Performance metrics The performance metrics are exact match (the percentage of predictions that exactly match the gold answer) and F_1-score (the average N-gram overlap between the prediction and the gold answer treated as bag-of-words).

Prompt templates We used the following five prompt templates from NorEval.

Prompt A:

```
1   Tittel: {$title}
2
3   Tekst: {$passage}
4
5   Spørsmål: {$question}
6
7   Svar: {$prediction}
```

590

Prompt B:

```
1   Tittel: {$title}
2
3   Tekst: {$passage}
4
5   Gitt teksten over, hva er svaret på følgende spørsmål? "{$question}"
6
7   Svar: {$prediction}
```

Prompt C:

```
1   Tittel: {$title}
2
3   Tekst: {$passage}
4
5   Svar på følgende: {$question}
6
7   Svar: {$prediction}
```

Prompt D:

```
1   Tittel: {$title}
2
3   Tekst: {$passage}
4
5   Hvordan kan man svare på spørsmålet "{$question}", gitt teksten over?
6
7   Svar: {$prediction}
```

Prompt E:

```
1   Tittel: {$title}
2
3   Tekst: {$passage}
4
5   Gitt teksten over, besvar følgende spørsmål: "{$question}"
6
7   Svar: {$prediction}
```

Full results   The complete evaluation results on NorQuAD (Bokmål) can be found in Table 7, both F_1 scores and exact-match accuracies.

B.3   Sentiment analysis

Sentiment analysis can serve as a good indicator of language understanding when evaluating language models. We use NoReC as a source of manually-annotated data for sentiment analysis (Velldal et al., 2018). While it offers fine-grained 6-class sentiment labels, we simplify the task to binary sentiment analysis, which works more reliably for in-context learning (Mæhlum et al., 2024).

Inference setup   The model is given a test example formatted according to a prompt template and ranks the answer candidates based on their probabilities. The most probable answer candidate is selected as the resulting answer.

Performance metrics   The dataset is slightly unbalanced and so we use the macro-average F_1-score to assess the performance.

Prompt template	0-shot					1-shot				
	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	**35.4**	32.8	**37.9**	16.5	**31.8**	**54.4**	55.3	**53.0**	50.6	52.8
NorwAI-Mistral-7B	28.4	22.0	28.6	8.5	21.6	41.3	40.7	41.5	37.7	42.6
NorwAI-Llama2-7B	23.1	18.0	24.2	7.4	16.7	34.5	39.2	36.4	35.2	37.7
NorMistral-7B-warm	24.8	21.0	23.7	3.2	17.6	37.1	41.9	40.7	36.0	41.3
NorGPT-3B	1.1	1.1	0.2	0.4	0.6	0.0	0.0	0.0	0.0	0.0
Viking-7B	15.0	20.3	16.9	7.6	20.3	28.8	29.9	27.3	26.3	29.7
Viking-13B	19.1	22.5	20.8	11.9	22.5	35.8	35.8	35.8	33.1	35.6
Mistral-Nemo-12B	27.3	**34.3**	29.2	**17.2**	31.6	49.4	**56.4**	49.4	**53.8**	**53.4**

EXACT MATCH

Prompt template	0-shot					1-shot				
	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	**35.4**	32.8	**37.9**	16.5	**31.8**	**54.4**	55.3	**53.0**	50.6	52.8
NorwAI-Mistral-7B	28.4	22.0	28.6	8.5	21.6	41.3	40.7	41.5	37.7	42.6
NorwAI-Llama2-7B	23.1	18.0	24.2	7.4	16.7	34.5	39.2	36.4	35.2	37.7
NorMistral-7B-warm	24.8	21.0	23.7	3.2	17.6	37.1	41.9	40.7	36.0	41.3
NorGPT-3B	1.1	1.1	0.2	0.4	0.6	0.0	0.0	0.0	0.0	0.0
Viking-7B	15.0	20.3	16.9	7.6	20.3	28.8	29.9	27.3	26.3	29.7
Viking-13B	19.1	22.5	20.8	11.9	22.5	35.8	35.8	35.8	33.1	35.6
Mistral-Nemo-12B	27.3	**34.3**	29.2	**17.2**	31.6	49.4	**56.4**	49.4	**53.8**	**53.4**

Table 7: **Complete results on extractive question answering with NorQuAD** We show the detailed results for each evaluated model, few-shot setting and prompt template. The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.3.1 Sentence-level NoReC

The converted dataset with binary sentiment labels can be found at **https://huggingface.co/datasets/ltg/norec_sentence**.

Prompt templates We used the following five prompt templates from NorEval.

Prompt A:

```
1   Tekst: {$text}
2   Sentiment: {$prediction:positiv/negativ}
```

Prompt B:

```
1   {$text}
2   Er denne setningen "positiv" eller "negativ"? {$prediction:positiv/negativ}
```

Prompt C:

```
1   {$text}
2   Hva slags sentiment uttrykker anmelderen? {$prediction:positiv/negativ}
```

Prompt D:

```
1   {$text}
2   Er anmeldelsen "positiv" eller "negativ"? {$prediction:positiv/negativ}
```

Prompt E:

```
1   {$text}
2   Er denne setningen positiv eller negativ? {$prediction:positiv/negativ}
```

Full results The complete evaluation results on sentence-level NoReC are given in Table 8. The random-guessing baseline achieves 48.5% on this task.

Prompt template	0-shot					1-shot					16-shot				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	73.9	68.8	68.8	69.1	68.8	**88.5**$^{\pm 1.3}$	65.4$^{\pm 2.0}$	79.8$^{\pm 1.7}$	76.3$^{\pm 1.8}$	72.4$^{\pm 1.9}$	**90.2**$^{\pm 1.2}$	**91.8**$^{\pm 1.1}$	90.2$^{\pm 1.2}$	**91.1**$^{\pm 1.2}$	**91.6**$^{\pm 1.2}$
NorwAI-Mistral-7B	69.8	55.7	**72.7**	53.3	63.0	76.0$^{\pm 1.8}$	55.1$^{\pm 2.1}$	81.1$^{\pm 1.6}$	77.2$^{\pm 1.7}$	56.1$^{\pm 2.1}$	89.0$^{\pm 1.3}$	86.4$^{\pm 1.4}$	**90.4**$^{\pm 1.2}$	87.3$^{\pm 1.4}$	87.8$^{\pm 1.4}$
NorwAI-Llama2-7B	67.2	54.9	69.1	36.7	58.5	72.0$^{\pm 1.9}$	64.3$^{\pm 2.0}$	65.9$^{\pm 2.0}$	63.6$^{\pm 2.0}$	65.5$^{\pm 2.0}$	88.2$^{\pm 1.3}$	83.5$^{\pm 1.5}$	88.9$^{\pm 1.3}$	82.8$^{\pm 1.6}$	87.5$^{\pm 1.4}$
NorMistral-7B-warm	**75.0**	68.4	61.9	54.7	**69.0**	81.6$^{\pm 1.6}$	69.1$^{\pm 1.9}$	74.6$^{\pm 1.8}$	71.5$^{\pm 1.9}$	69.3$^{\pm 1.9}$	86.6$^{\pm 1.4}$	72.4$^{\pm 1.9}$	85.9$^{\pm 1.4}$	77.2$^{\pm 1.7}$	72.6$^{\pm 1.8}$
NorGPT-3B	72.4	41.2	47.3	**71.4**	67.9	66.0$^{\pm 2.0}$	61.2$^{\pm 2.0}$	61.9$^{\pm 2.0}$	64.3$^{\pm 2.0}$	59.3$^{\pm 2.0}$	58.0$^{\pm 2.0}$	48.9$^{\pm 2.1}$	65.2$^{\pm 2.0}$	48.7$^{\pm 2.1}$	51.1$^{\pm 2.1}$
Viking-7B	70.5	69.0	70.8	59.0	67.4	79.4$^{\pm 1.7}$	70.8$^{\pm 1.9}$	74.4$^{\pm 1.8}$	73.6$^{\pm 1.8}$	55.6$^{\pm 2.1}$	81.8$^{\pm 1.6}$	77.4$^{\pm 1.7}$	73.8$^{\pm 1.8}$	76.2$^{\pm 1.8}$	82.5$^{\pm 1.6}$
Viking-13B	69.1	**69.1**	68.1	50.8	68.1	78.9$^{\pm 1.7}$	69.0$^{\pm 1.9}$	79.9$^{\pm 1.7}$	71.5$^{\pm 1.9}$	69.0$^{\pm 1.9}$	84.0$^{\pm 1.5}$	77.4$^{\pm 1.7}$	83.0$^{\pm 1.6}$	80.4$^{\pm 1.6}$	79.2$^{\pm 1.7}$
Mistral-Nemo-12B	71.9	68.4	68.8	68.4	69.0	84.4$^{\pm 1.5}$	**77.4**$^{\pm 1.7}$	**84.6**$^{\pm 1.5}$	**82.0**$^{\pm 1.6}$	**80.3**$^{\pm 1.6}$	87.0$^{\pm 1.4}$	89.0$^{\pm 1.3}$	88.2$^{\pm 1.3}$	87.5$^{\pm 1.4}$	88.9$^{\pm 1.3}$

Table 8: **Complete results on sentence-level sentiment analysis with NoReC** We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.3.2 Document-level NoReC

The converted dataset with binary sentiment labels can be found at `https://huggingface.co/datasets/ltg/norec_document`.

Prompt templates We used the following five prompt templates from `NorEval` for testing all language models on document-level sentiment analysis:

Prompt A:

```
1   Tekst: {$text}
2   Sentiment: {$prediction:positiv/negativ}
```

Prompt B:

```
1   Tekst: {$text}
2   Er anmeldelsen "positiv" eller "negativ"? {$prediction:positiv/negativ}
```

Prompt C:

```
1   Er polariteten til følgende anmeldelse positiv eller negativ?
2   Anmeldelse: {$text}
3   Anmeldelsen er {$prediction:positiv/negativ}
```

Prompt D:

```
1   Anmeldelse: {$text}
2   Er anmelderen positiv eller negativ? {$prediction:positiv/negativ}
```

Prompt E:

```
1   Anmeldelse: {$text}
2   Vil du oppsummere anmeldelsen som "bra" eller "dårlig"? {$prediction:bra/dårlig}
```

Full results The complete evaluation results on document-level NoReC are provided in Table 9. The random-guessing baseline achieves 48.4% on this task.

	0-shot					1-shot				
Prompt template	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	69.4	67.0	66.6	**68.9**	67.0	$88.4^{\pm0.6}$	$\mathbf{92.5^{\pm0.5}}$	$86.0^{\pm0.6}$	$\mathbf{92.5^{\pm0.5}}$	$87.7^{\pm0.6}$
NorwAI-Mistral-7B	69.1	66.1	67.0	66.5	63.8	$82.8^{\pm0.7}$	$70.4^{\pm0.8}$	$84.6^{\pm0.7}$	$77.1^{\pm0.8}$	$84.6^{\pm0.7}$
NorwAI-Llama2-7B	71.9	39.8	67.0	65.2	65.2	$76.2^{\pm0.8}$	$73.2^{\pm0.8}$	$78.7^{\pm0.8}$	$83.8^{\pm0.7}$	$80.1^{\pm0.7}$
NorMistral-7B-warm	74.8	55.6	67.2	67.4	**67.6**	$84.3^{\pm0.7}$	$73.9^{\pm0.8}$	$84.3^{\pm0.7}$	$75.7^{\pm0.8}$	$73.4^{\pm0.8}$
NorGPT-3B	67.7	52.4	67.0	67.7	67.0	$58.1^{\pm0.9}$	$54.0^{\pm0.9}$	$55.5^{\pm0.9}$	$54.8^{\pm0.9}$	$55.0^{\pm0.9}$
Viking-7B	75.3	56.6	**68.3**	65.9	67.0	$84.5^{\pm0.7}$	$78.4^{\pm0.8}$	$73.9^{\pm0.8}$	$74.5^{\pm0.8}$	$73.4^{\pm0.8}$
Viking-13B	69.0	66.8	67.3	68.3	65.0	$83.2^{\pm0.7}$	$72.5^{\pm0.8}$	$89.2^{\pm0.6}$	$84.5^{\pm0.7}$	$83.2^{\pm0.7}$
Mistral-Nemo-12B	**78.5**	**68.0**	67.0	67.1	67.0	$\mathbf{91.2^{\pm0.5}}$	$89.8^{\pm0.6}$	$\mathbf{90.5^{\pm0.5}}$	$89.8^{\pm0.6}$	$\mathbf{89.3^{\pm0.6}}$

Table 9: **Complete results on document-level sentiment analysis with NoReC** We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.4 NorCommonsenseQA

Accurately predicting the correct answers on this datasets requires different types of commonsense knowledge. The creating of the Norwegian NorCommonsenseQA has been inspired by the English CommonsenseQA dataset (Talmor et al., 2019). The data can be found at `https://huggingface.co/datasets/ltg/norcommonsenseqa`.

Inference setup The model is given a test example formatted according to a prompt template and ranks the answer candidates based on their probabilities. The most probable answer candidate is selected as the resulting answer.

Performance metric There are five possible answers for each question. We measure the performance with a simple accuracy.

Prompt templates We used the following five prompt templates from `NorEval`. The templates are adapted to the Bokmål and Nynorsk versions of this dataset.

Prompt A (Bokmål and Nynorsk):

```
1   Spørsmål: {$question}
2
3   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}/{$answer_5}}
```

Prompt B (Bokmål):

```
1   {$question}
2   Hvilket av følgende mulige svar er det riktige?
3   A: {$answer_1}
4   B: {$answer_2}
5   C: {$answer_3}
6   D: {$answer_4}
7   E: {$answer_5}
8   Svar: {$prediction:A/B/C/D/E}
```

594

Prompt B (Nynorsk):

```
1   {$question}
2   Kva av følgande moglege svar er det rette?
3   A: {$answer_1}
4   B: {$answer_2}
5   C: {$answer_3}
6   D: {$answer_4}
7   E: {$answer_5}
8   Svar: {$prediction:A/B/C/D/E}
```

Prompt C (Bokmål):

```
1   Gitt alternativene under, hva er svaret på følgende spørsmål: {$question}
2
3   Alternativer:
4   - {$answer_1}
5   - {$answer_2}
6   - {$answer_3}
7   - {$answer_4}
8   - {$answer_5}
9
10  Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}/{$answer_5}}
```

Prompt C (Nynorsk):

```
1   Gitt alternativa under, kva er svaret på følgande spørsmål: {$question}
2
3   Alternativ:
4   - {$answer_1}
5   - {$answer_2}
6   - {$answer_3}
7   - {$answer_4}
8   - {$answer_5}
9
10  Svar: {$prediction:A/B/C/D/E}
```

Prompt D (Bokmål):

```
1   {$question}
2   Velg riktig svar blant disse alternativene:
3   - {$answer_1}
4   - {$answer_2}
5   - {$answer_3}
6   - {$answer_4}
7   - {$answer_5}
8
9   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}/{$answer_5}}
```

Prompt D (Nynorsk):

```
1   {$question}
2   Vel rett svar blant desse alternativa:
3   - {$answer_1}
4   - {$answer_2}
5   - {$answer_3}
6   - {$answer_4}
7   - {$answer_5}
8
9   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}/{$answer_5}}
```

Prompt E (Bokmål):

```
1   {$question}
2   A: {$answer_1}
3   B: {$answer_2}
4   C: {$answer_3}
5   D: {$answer_4}
6   E: {$answer_5}
7
8   Er det riktige svaret A, B, C, D, eller E?
9
10  Svar: {$prediction:A/B/C/D/E}
```

Prompt E (Nynorsk):

```
1   {$question}
2   A: {$answer_1}
3   B: {$answer_2}
4   C: {$answer_3}
5   D: {$answer_4}
6   E: {$answer_5}
7
8   Er det rette svaret A, B, C, D, eller E?
9
10  Svar: {$prediction:A/B/C/D/E}
```

Full results The complete evaluation results on NorCommonsenseQA (Bokmål and Nynorsk) are provided in Table 10. For reference, the random-guessing baseline achieves 20% on this task (for both language variants).

Prompt template	Bokmål (0-shot)					Nynorsk (0-shot)				
	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	**61.0**	**56.9**	23.1	**51.8**	**45.3**	**44.2**	**51.6**	**36.8**	**46.3**	**30.5**
NorwAI-Mistral-7B	30.8	49.7	20.3	22.3	28.5	43.2	20.0	23.2	27.4	23.2
NorwAI-Llama2-7B	37.2	54.2	23.2	25.8	33.6	37.9	18.9	17.9	27.4	18.9
NorMistral-7B-warm	30.4	51.3	20.5	21.4	29.2	43.2	18.9	15.8	30.5	20.0
NorGPT-3B	26.4	34.7	22.1	20.1	23.5	29.5	20.0	16.8	25.3	25.3
Viking-7B	26.1	44.9	19.1	20.5	23.2	38.9	21.1	25.3	23.2	23.2
Viking-13B	24.7	51.1	18.1	19.1	24.0	40.0	13.7	24.2	20.0	16.8
Mistral-Nemo-12B	43.4	44.1	**43.7**	38.9	31.7	33.7	33.7	25.3	27.4	25.3

Table 10: **Complete results on commonsense reasoning evaluated on NorCommonsenseQA (Bokmål and Nynorsk)** We show the detailed results for each evaluated model and prompt template. The best results for each column are boldfaced, the overall best result is highlighted in blue.

B.5 NRK-Quiz-QA

This question-answering dataset focuses on knowledge about Norway and its culture. The data can be found at `https://huggingface.co/datasets/ltg/nrk_quiz_qa`.

Inference setup The model is given a test example formatted according to a **Prompt template** and ranks the answer candidates based on their probabilities. The most probable answer candidate is selected as the resulting answer.

Performance metric There is a limited number of possible answers for each question. We measure the performance with a simple accuracy.

Prompt templates   We used the following five prompt templates from `NorEval` for testing all language models on question answering with NRK-Quiz-QA. Note that the examples in this dataset have a variable number of answer options, we show the prompt templates for four options as an example. The templates are adapted to the Bokmål and Nynorsk versions of this dataset.

Prompt A (Bokmål and Nynorsk):

```
1   Spørsmål: {$question}
2
3   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt B (Bokmål):

```
1    {$question}
2
3    Svaralternativer:
4    - {$answer_1}
5    - {$answer_2}
6    - {$answer_3}
7    - {$answer_4}
8
9    Hva er riktig svar?
10
11   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt B (Nynorsk):

```
1    {$question}
2    {$question}
3
4    Svaralternativer:
5    - {$answer_1}
6    - {$answer_2}
7    - {$answer_3}
8    - {$answer_4}
9
10   Kva er rett svar?
11
12   Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt C (Bokmål):

```
1    {$question}
2    A: {$answer_1}
3    B: {$answer_2}
4    C: {$answer_3}
5    D: {$answer_4}
6
7    Er det riktige svaret A, B, C, eller D?
8
9    Svar: {$prediction:A/B/C/D}
```

Prompt C (Nynorsk):

```
1    {$question}
2    A: {$answer_1}
3    B: {$answer_2}
4    C: {$answer_3}
5    D: {$answer_4}
6
7    Er det rette svare A, B, C, eller D?
```

```
8
9    Svar: {$prediction:A/B/C/D}
```

Prompt D (Bokmål and Nynorsk):

```
1    Spørsmål: {$question}
2    A: {$answer_1}
3    B: {$answer_2}
4    C: {$answer_3}
5    D: {$answer_4}
6
7    Svar: {$prediction:A/B/C/D}
```

Prompt E (Bokmål):

```
1    {$question}
2    Velg riktig svar blant disse alternativene:
3    - {$answer_1}
4    - {$answer_2}
5    - {$answer_3}
6    - {$answer_4}
7
8    Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt E (Nynorsk):

```
1    {$question}
2    Vel rett svar blant desse alternativa:
3    - {$answer_1}
4    - {$answer_2}
5    - {$answer_3}
6    - {$answer_4}
7
8    Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Full results The complete evaluation results on NRK-Quiz-QA (Bokmål and Nynorsk) are in Table 11. The random-guessing baseline achieves 28% accuracy on the Bokmål version of this task and 27% on the Nynorsk version.

Prompt template	Bokmål (0-shot)					Nynorsk (0-shot)				
	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	**63.7**	**50.5**	38.6	41.1	**50.6**	**71.9**	**56.5**	**46.4**	41.9	**57.1**
NorwAI-Mistral-7B	55.2	43.4	34.6	34.8	46.6	65.2	50.6	35.8	35.6	53.2
NorwAI-Llama2-7B	52.3	39.2	26.0	30.1	40.3	64.3	44.1	25.3	31.8	44.1
NorMistral-7B-warm	57.9	39.8	27.7	32.5	40.7	65.9	41.3	28.8	32.7	41.1
NorGPT-3B	33.1	28.2	26.3	26.1	27.9	37.3	29.6	25.0	24.7	30.5
Viking-7B	44.3	29.9	26.1	28.8	31.9	51.1	31.2	26.8	30.8	34.7
Viking-13B	51.0	31.8	27.8	30.2	31.6	54.8	34.5	28.0	30.2	31.9
Mistral-Nemo-12B	47.0	46.1	**41.8**	**47.4**	46.6	47.2	43.6	41.4	**45.7**	42.8

Table 11: **Complete results on Norwegian-specific and world knowledge evaluated on NRK-Quiz-QA (Bokmål and Nynorsk)** We show the detailed results for each evaluated model and prompt template. The best results for each column are boldfaced, the overall best result is highlighted in blue.

B.6 NorOpenBookQA

Inspired by the English OpenBookQA (Mihaylov et al., 2018), this task follows the open book exams for testing human understanding of a subject. Correctly answering a question should require multi-step reasoning, common and commonsense knowledge, and rich text comprehension. The data can be found at `https://huggingface.co/datasets/ltg/noropenbookqa`.

Inference setup The model is given a test example formatted according to a prompt template and ranks the answer candidates based on their probabilities. The most probable answer candidate is selected as the resulting answer.

Performance metric There are four possible for answers for each passage-question pair. We measure the performance with a simple accuracy.

Prompt templates We used the following five prompt templates from `NorEval` for testing all language models on question answering with NorOpenBookQA: The templates are adapted to the Bokmål and Nynorsk versions of this dataset.

Prompt A (Bokmål and Nynorsk):

```
1    {$fact}
2    {$question} {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt B (Bokmål):

```
1    Faktatekst: {$fact}
2    Spørsmål til teksten: {$question}
3
4    Svaralternativer:
5    - {$answer_1}
6    - {$answer_2}
7    - {$answer_3}
8    - {$answer_4}
9
10   Hva er riktig svar? {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt B (Nynorsk):

```
1    Faktatekst: {$fact}
2    Spørsmål til teksten: {$question}
3
4    Svaralternativer:
5    - {$answer_1}
6    - {$answer_2}
7    - {$answer_3}
8    - {$answer_4}
9
10   Kva er rett svar? {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt C (Bokmål):

```
1    {$fact}
2    {$question}
3    A: {$answer_1}
4    B: {$answer_2}
5    C: {$answer_3}
6    D: {$answer_4}
7
8    Er det riktige svaret A, B, C, eller D?
9
10   Svar: {$prediction:A/B/C/D}
```

Prompt C (Nynorsk):

```
1  {$fact}
2  {$question}
3  A: {$answer_1}
4  B: {$answer_2}
5  C: {$answer_3}
6  D: {$answer_4}
7
8  Er det rette svare A, B, C, eller D?
9
10 Svar: {$prediction:A/B/C/D}
```

Prompt D (Bokmål and Nynorsk):

```
1  Bakgrunn: {$fact}
2
3  Spørsmål: {$question}
4  A: {$answer_1}
5  B: {$answer_2}
6  C: {$answer_3}
7  D: {$answer_4}
8
9  Svar: {$prediction:A/B/C/D}
```

Prompt E (Bokmål):

```
1  Ta utgangspunkt i følgende fakta når du svarer på spørsmålet: {$fact}
2
3  {$question}
4  Velg riktig svar blant disse alternativene:
5  - {$answer_1}
6  - {$answer_2}
7  - {$answer_3}
8  - {$answer_4}
9
10 Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Prompt E (Nynorsk):

```
1  Ta utgangspunkt i følgande fakta når du svarar på spørsmålet: {$fact}
2
3  {$question}
4  Vel rett svar blant desse alternativa:
5  - {$answer_1}
6  - {$answer_2}
7  - {$answer_3}
8  - {$answer_4}
9
10 Svar: {$prediction:{$answer_1}/{$answer_2}/{$answer_3}/{$answer_4}}
```

Full results   The complete evaluation on NorOpenBookQA (Bokmål and Nynorsk) is in Table 12. Note that randomly guessing the answers achieves 25% on this task.

B.7   Summarization (NorSumm)

NorSumm by Touileb et al. (2025) is a benchmark for abstractive summarization of Norwegian news articles. It offers another perspective on the level of Norwegian language understanding of different language models. An important feature of this dataset is that its Bokmål and Nynorsk variants are parallel.

BOKMÅL

Prompt template	0-shot					1-shot					16-shot				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	44.6	55.7	54.0	67.8	65.8	$45.6^{\pm2.9}$	$74.5^{\pm2.5}$	$68.8^{\pm2.7}$	$76.2^{\pm2.5}$	$70.1^{\pm2.7}$	$\mathbf{51.3^{\pm2.9}}$	$75.8^{\pm2.5}$	$77.9^{\pm2.4}$	$75.8^{\pm2.5}$	$76.5^{\pm2.5}$
NorwAI-Mistral-7B	**47.7**	35.6	34.6	35.9	43.3	$46.3^{\pm2.9}$	$51.7^{\pm2.9}$	$34.9^{\pm2.8}$	$35.2^{\pm2.8}$	$50.0^{\pm2.9}$	$50.0^{\pm2.9}$	$52.3^{\pm2.9}$	$40.9^{\pm2.9}$	$44.3^{\pm2.9}$	$51.7^{\pm2.9}$
NorwAI-Llama2-7B	45.6	32.6	25.8	32.6	43.3	$45.0^{\pm2.9}$	$51.7^{\pm2.9}$	$33.2^{\pm2.7}$	$37.6^{\pm2.8}$	$46.3^{\pm2.9}$	$45.0^{\pm2.9}$	$52.3^{\pm2.9}$	$32.2^{\pm2.7}$	$34.2^{\pm2.8}$	$49.7^{\pm2.9}$
NorMistral-7B-warm	46.6	35.6	28.9	32.6	44.3	$\mathbf{46.6^{\pm2.9}}$	$51.7^{\pm2.9}$	$32.9^{\pm2.7}$	$34.2^{\pm2.8}$	$46.6^{\pm2.9}$	$48.3^{\pm2.9}$	$49.0^{\pm2.9}$	$42.3^{\pm2.9}$	$40.3^{\pm2.8}$	$45.0^{\pm2.9}$
NorGPT-3B	32.6	31.2	24.2	22.5	33.2	$28.5^{\pm2.6}$	$28.2^{\pm2.6}$	$23.8^{\pm2.5}$	$24.2^{\pm2.5}$	$28.5^{\pm2.6}$	$29.5^{\pm2.6}$	$27.9^{\pm2.6}$	$26.5^{\pm2.6}$	$26.8^{\pm2.6}$	$28.2^{\pm2.6}$
Viking-7B	41.9	26.5	20.8	28.5	27.9	$45.3^{\pm2.9}$	$25.8^{\pm2.5}$	$26.8^{\pm2.6}$	$24.5^{\pm2.5}$	$30.9^{\pm2.7}$	$48.7^{\pm2.9}$	$26.5^{\pm2.6}$	$28.2^{\pm2.6}$	$23.8^{\pm2.5}$	$31.9^{\pm2.7}$
Viking-13B	44.6	27.5	21.1	25.5	31.9	$45.6^{\pm2.9}$	$33.2^{\pm2.7}$	$25.2^{\pm2.5}$	$27.2^{\pm2.6}$	$38.6^{\pm2.8}$	$47.0^{\pm2.9}$	$38.9^{\pm2.8}$	$29.9^{\pm2.7}$	$26.2^{\pm2.6}$	$36.9^{\pm2.8}$
Mistral-Nemo-12B	43.6	**60.7**	**58.1**	**71.5**	68.5	$44.0^{\pm2.9}$	$\mathbf{82.6^{\pm2.2}}$	$\mathbf{82.2^{\pm2.2}}$	$\mathbf{82.9^{\pm2.2}}$	$76.2^{\pm2.5}$	$49.7^{\pm2.9}$	$\mathbf{82.9^{\pm2.2}}$	$\mathbf{82.2^{\pm2.2}}$	$\mathbf{85.9^{\pm2.0}}$	$\mathbf{80.9^{\pm2.3}}$

NYNORSK

Prompt template	0-shot					1-shot					16-shot				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	**33.3**	56.7	56.7	56.7	65.6	$28.9^{\pm4.8}$	$70.0^{\pm4.9}$	$68.9^{\pm4.9}$	$71.1^{\pm4.8}$	$68.9^{\pm4.9}$	$\mathbf{40.0^{\pm5.2}}$	$72.2^{\pm4.7}$	$76.7^{\pm4.5}$	$77.8^{\pm4.4}$	$77.8^{\pm4.4}$
NorwAI-Mistral-7B	30.0	37.8	32.2	27.8	38.9	$31.1^{\pm4.9}$	$34.4^{\pm5.0}$	$28.9^{\pm4.8}$	$27.8^{\pm4.7}$	$36.7^{\pm5.1}$	$37.8^{\pm5.1}$	$45.6^{\pm5.3}$	$41.1^{\pm5.2}$	$44.4^{\pm5.3}$	$42.2^{\pm5.2}$
NorwAI-Llama2-7B	25.6	30.0	32.2	27.8	28.9	$28.9^{\pm4.8}$	$37.8^{\pm5.1}$	$32.2^{\pm5.0}$	$32.2^{\pm5.0}$	$36.7^{\pm5.1}$	$27.8^{\pm4.7}$	$37.8^{\pm5.1}$	$32.2^{\pm5.0}$	$32.2^{\pm5.0}$	$38.9^{\pm5.2}$
NorMistral-7B-warm	26.7	28.9	34.4	40.0	36.7	$27.8^{\pm4.7}$	$32.2^{\pm5.0}$	$40.0^{\pm5.2}$	$38.9^{\pm5.2}$	$43.3^{\pm5.3}$	$30.0^{\pm4.9}$	$38.9^{\pm5.2}$	$41.1^{\pm5.2}$	$40.0^{\pm5.2}$	$41.1^{\pm5.2}$
NorGPT-3B	20.0	27.8	34.4	30.0	25.6	$20.0^{\pm4.2}$	$25.6^{\pm4.6}$	$24.4^{\pm4.6}$	$23.3^{\pm4.5}$	$20.0^{\pm4.2}$	$18.9^{\pm4.1}$	$23.3^{\pm4.5}$	$34.4^{\pm5.0}$	$26.7^{\pm4.7}$	$23.3^{\pm4.5}$
Viking-7B	22.2	20.0	18.9	27.8	15.6	$30.0^{\pm4.9}$	$22.2^{\pm4.4}$	$32.2^{\pm5.0}$	$25.6^{\pm4.6}$	$27.8^{\pm4.7}$	$27.8^{\pm4.7}$	$22.2^{\pm4.4}$	$27.8^{\pm4.7}$	$23.3^{\pm4.5}$	$27.8^{\pm4.7}$
Viking-13B	30.0	34.4	17.8	23.3	31.1	$32.2^{\pm5.0}$	$27.8^{\pm4.7}$	$17.8^{\pm4.1}$	$33.3^{\pm5.0}$	$33.3^{\pm5.0}$	$36.7^{\pm5.1}$	$26.7^{\pm4.7}$	$22.2^{\pm4.4}$	$18.9^{\pm4.1}$	$26.7^{\pm4.7}$
Mistral-Nemo-12B	30.0	52.2	52.2	**58.9**	60.0	$\mathbf{35.6^{\pm5.1}}$	$71.1^{\pm4.8}$	$77.8^{\pm4.4}$	$\mathbf{78.9^{\pm4.3}}$	$71.1^{\pm4.8}$	$33.3^{\pm5.0}$	$\mathbf{82.2^{\pm4.1}}$	$\mathbf{82.2^{\pm4.1}}$	$\mathbf{86.7^{\pm3.6}}$	$\mathbf{81.1^{\pm4.1}}$

Table 12: **Complete results on world knowledge evaluated on NorOpenBookQA (Bokmål and Nynorsk)** We show the detailed results for each evaluated model, few-shot setting and prompt template. The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

Inference setup The model is given a test example formatted according to a **Prompt template** and generates the answer via a greedy search decoding strategy.

Performance metric We use Rouge-L as the standard metric for summarization (Lin, 2004). ROUGE-L uses longest common subsequence matching, allowing it to identify matching content even when ordered differently in generated and reference summaries.

Prompt templates We used the following six prompt templates from NorEval for testing all language models on summarization with NorSumm. The templates are adapted to the Bokmål and Nynorsk versions of this dataset.

Prompt A (Bokmål):

```
1   Skriv en oppsummering av følgende artikkel med kun noen få punkter: {$article}
2   Oppsummering: {$prediction}
```

Prompt A (Nynorsk):

```
1   Skriv ei oppsummering av følgande artikkel med berre nokre få punkt: {$article}
2   Oppsummering: {$prediction}
```

Prompt B (Bokmål):

```
1  Oppsummer følgende artikkel med noen få setninger: {$article}
2  Oppsummering: {$prediction}
```

Prompt B (Nynorsk):

```
1  Oppsummer følgande artikkel med nokre få setningar: {$article}
2  Oppsummering: {$prediction}
```

Prompt C (Bokmål):

```
1  {$article}
2  Skriv en kort og presis oppsummering av teksten over. Språket må være klart og lett å forstå.
   ↪  Sørg for å ikke introdusere feil. Oppsummeringen må dekke følgende spørsmål: hvem, hva,
   ↪  hvor, når, og hvorfor er denne saken viktig å vite om. Oppsummeringen må være engasjerende
   ↪  og fremheve nøkkelinformasjon fra artikkelen. Oppsummeringen skal inneholde maksimalt 700
   ↪  tegn, inkludert mellomrom. {$prediction}
```

Prompt C (Nynorsk):

```
1  {$article}
2  Skriv ein kort og presis oppsummering av teksten over. Språket må vere klart og lett å forstå.
   ↪  Sørg for å ikkje introdusere feil. Oppsummeringa må dekkje følgande spørsmål: kven, kva,
   ↪  kor, når, og kvifor er denne saka viktig å vite om. Oppsummeringa må vere engasjerande og
   ↪  framheve nøkkelinformasjon frå artikkelen. Oppsummeringa skal innehalde maksimalt 700 tegn,
   ↪  inkludert mellomrom. {$prediction}
```

Prompt D (Bokmål):

```
1  Gi et kortfattet sammendrag av følgende tekst: {$article} {$prediction}
```

Prompt D (Nynorsk):

```
1  Gje eit kortfatta samandrag av følgande tekst: {$article} {$prediction}
```

Prompt E (Bokmål):

```
1  Lag en kort oppsummering som sammenfatter den følgende teksten i noen få punkter:
2  {$article}
3
4  Oppsummering: {$prediction}
```

Prompt E (Nynorsk):

```
1  Lag ein kort oppsummering som samanfattar den følgande teksten i nokre få punkt:
2  {$article}
3
4  Oppsummering: {$prediction}
```

Prompt F (Bokmål):

```
1  Hele artikkelen:
2  {$article}
3
4  Hovedpunkter: {$prediction}
```

Prompt F (Nynorsk):

```
1  Heile artikkelen:
2  {$article}
3
4  Hovudpunkt: {$prediction}
```

Prompt template	Bokmål (0-shot)							Nynorsk (0-shot)					
	A	B	C	D	E	F		A	B	C	D	E	F
NorMistral-11B	**17.6**	20.5	**34.9**	**45.0**	42.5	**5.7**		**15.4**	17.9	25.4	**32.4**	**32.6**	6.8
NorwAI-Mistral-7B	12.2	0.0	0.0	2.6	0.0	0.0		10.3	0.0	0.0	3.5	0.0	0.0
NorwAI-Llama2-7B	10.7	0.0	0.0	7.8	3.5	0.0		10.4	0.0	0.0	4.4	5.9	0.0
NorMistral-7B-warm	7.8	0.0	0.0	16.5	9.0	0.0		8.6	0.0	0.0	6.9	4.1	0.0
NorGPT-3B	8.8	8.5	31.6	33.8	25.2	2.8		7.4	10.9	21.6	24.3	20.0	4.0
Viking-7B	11.2	5.5	16.5	29.8	31.9	0.0		10.5	3.0	16.4	25.7	24.2	0.0
Viking-13B	11.1	1.7	6.0	23.7	36.3	0.0		9.8	0.0	4.3	19.6	28.8	0.4
Mistral-Nemo-12B	13.4	**26.4**	41.5	35.6	**44.9**	2.9		12.4	**18.2**	**30.0**	30.3	30.9	3.6

Table 13: **Complete results on NorSumm summarization (Bokmål and Nynorsk versions)** We show the detailed results for each evaluated model and prompt template. The best results for each column are boldfaced, the overall best result is highlighted in blue.

Full results The complete evaluation on NorSumm (both Bokmål and Nynorsk variants) is provided in Table 13.

B.8 Grammatical error correction (ASK-GEC)

This tasks tests how do language models understand more low-level features of the Norwegian language. We use the ASK-GEC dataset from Jentoft (2023) that is based on corrected essays of Norwegian language learners.

Inference setup The model is given a test example formatted according to a prompt template; given this input, it then generates the answer via a greedy-search decoding strategy.

Performance metric We use the $F_{0.5}$-score to measure the amount of successfully fixed correction-spans. These spans are heuristically identified by the ERRANT system (Bryant et al., 2017). More details about using this metric for Norwegian grammatical error corrections can be found in Jentoft (2023).

Prompt templates We used the following five prompt templates for grammatical error correction:

Prompt A:

```
1   Tekst: {$text}
2   Korreksjon: {$prediction}
```

Prompt B:

```
1   Tekst: {$text}
2   Rettet versjon: {$prediction}
```

Prompt C:

```
1   Skriv om følgende tekst slik at den blir grammatisk korrekt: {$text}
2   Korreksjon: {$prediction}
```

Prompt D:

```
1   Original versjon: {$text}
2   Korrekturlest og rettet versjon: {$prediction}
```

Prompt E:

```
1   Rett opp grammatiske feil i denne teksten: {$text}
2   Korreksjon: {$prediction}
```

Prompt template	0-shot					1-shot					16-shot				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	2.8	16.6	**39.9**	26.8	**38.8**	38.6	41.8	43.3	**45.2**	44.7	**52.6**	52.3	50.4	51.5	51.4
NorwAI-Mistral-7B	0.0	0.0	22.2	0.0	0.0	37.2	39.1	45.3	42.8	**46.1**	51.9	**52.4**	**52.5**	**53.2**	**52.7**
NorwAI-Llama2-7B	0.0	0.2	13.3	0.0	27.1	38.5	40.5	**46.2**	44.2	45.0	51.1	51.3	51.2	51.4	51.1
NorMistral-7B-warm	2.1	0.0	11.1	**33.6**	18.7	34.4	38.0	42.8	41.2	41.1	48.0	48.2	48.7	48.2	48.5
NorGPT-3B	0.2	0.0	0.2	1.4	0.3	0.4	0.4	0.4	0.4	0.4	0.9	1.3	1.8	1.2	1.5
Viking-7B	2.8	1.2	23.1	0.0	11.4	29.9	37.0	40.6	39.9	38.9	50.7	51.0	50.4	51.2	50.1
Viking-13B	3.1	0.0	37.8	25.1	34.8	**42.6**	**43.5**	45.7	44.8	46.0	52.4	52.0	51.9	52.4	51.8
Mistral-Nemo-12B	**14.7**	**18.6**	36.5	16.9	12.3	38.8	36.8	37.5	38.6	39.6	43.9	43.7	42.7	43.7	43.1

Table 14: **Complete results on grammatical error correction**   We show the detailed results for each evaluated model, few-shot setting and prompt template. The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

Full results   The complete evaluation on ASK-GEC is provided in Table 15.

B.9   Language identification (SLIDE)

We use the Scandinavian language identification and evaluation (SLIDE) from `https://github.com/ltgoslo/slide`. This dataset consists of sentences manually annotated with the language they are written in: Norwegian Bokmål, Nynorsk, Danish or Swedish (we filtered out the examples that are not written in any Scandinavian language). The sentences can be annotated with multiple language labels if applicable.

Inference setup   This task is solved as classification – the label with the highest probability given the prompt (estimated by the evaluated language model) is chosen as the predicted label. The few-shot demonstrations are randomly sampled from the SLIDE validation set.

Performance metric   We test whether a language model is able to correctly predict one of the (potentially multiple) languages a sentence is written in. We thus adopt the *loose accuracy* metric from SLIDE, where a single-label prediction is considered to be correct if is in the set of gold language labels.

Prompt templates   We used the following five prompt templates for testing all language models on grammatical error correction. The few-shot demonstrations are separated by double newlines \n\n.

Prompt A:

```
1   Tekst: {$text}
2   Korreksjon: {$prediction}
```

Prompt B:

```
1   Tekst: {$text}
2   Rettet versjon: {$prediction}
```

Prompt C:

```
1   Skriv om følgende tekst slik at den blir grammatisk korrekt: {$text}
2   Korreksjon: {$prediction}
```

Prompt D:

```
1   Original versjon: {$text}
2   Korrekturlest og rettet versjon: {$prediction}
```

Prompt E:

```
1    Rett opp grammatiske feil i denne teksten: {$text}
2    Korreksjon: {$prediction}
```

Full results The complete evaluation on language identification is given in **??**. Note that the majority baseline on this task is 40.3% loose accuracy and the random baseline is 28.2%.

Prompt template	0-shot					1-shot					16-shot				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
NorMistral-11B	74.0	41.5	**55.7**	62.6	53.1	78.8	$60.8^{\pm0.7}$	$80.6^{\pm0.2}$	$80.6^{\pm0.4}$	58.7	97.9	$95.6^{\pm0.2}$	94.6	$97.0^{\pm0.3}$	$98.2^{\pm0.1}$
NorwAI-Mistral-7B	65.8	54.0	38.8	**69.6**	**66.9**	74.8	39.2	47.9	76.9	64.5	95.1	69.9	90.4	92.0	95.7
NorwAI-Llama2-7B	69.8	37.4	42.9	49.4	59.0	65.4	33.0	40.8	54.5	43.1	93.5	74.1	67.0	77.4	87.5
NorMistral-7B-warm	**87.5**	47.7	42.2	65.4	61.6	$85.7^{\pm0.4}$	40.3	63.5	72.2	$73.2^{\pm0.4}$	$98.1^{\pm0.1}$	92.2	$96.2^{\pm0.3}$	92.1	97.3
NorGPT-3B	36.6	24.0	49.9	43.9	42.6	37.7	35.6	32.4	32.4	32.4	39.0	27.9	40.0	40.3	40.2
Viking-7B	74.4	42.7	41.0	34.7	32.8	46.2	37.1	35.2	39.5	36.0	77.2	47.4	44.3	58.5	52.6
Viking-13B	71.5	**59.5**	41.0	41.9	32.4	55.1	36.4	37.8	43.8	34.1	84.4	62.0	56.8	79.1	65.3
Mistral-Nemo-12B	68.3	41.7	50.3	48.5	40.7	63.8	56.0	74.3	58.6	45.8	85.9	84.6	86.6	87.3	86.1

Table 15: **Complete results on Scandinavian language identification** We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.10 Translation

Inference setup This task is solved as generation via prefix prompting – the model is given a prompt without the `$prediction` suffix and then it autoregressively generates a prediction until outputting a newline. We use simple greedy search to generate the output.

Performance metric We measure the translation quality with SacreBLEU scores (Post, 2018) with signature BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20 as the main metric. We also provide chrF++ scores as an additional metric (Popović, 2017).

B.10.1 English to Bokmål translation

Prompt templates We used the following four prompt templates for testing all language models on translation to Northern Sámi.

Prompt A:

```
1    Engelsk: {$text}
2    Bokmål: {$prediction}
```

Prompt B:

```
1    Oversett følgende setning til Bokmål: {$text}
2    Bokmål: {$prediction}
```

Prompt C:

```
1    Gi en oversettelse til Bokmål for denne setningen: {$text}
2    Bokmål: {$prediction}
```

Prompt D:

```
1    Hva blir "{$text}" på Bokmål?
2    Bokmål: {$prediction}
```

Full results The complete evaluation on translation to Bokmål is given in Table 18 (BLEU scores in the top sub-table and chrF++ scores below).

BLEU SCORES

Prompt template	0-shot				1-shot				16-shot			
	A	B	C	D	A	B	C	D	A	B	C	D
NorMistral-11B	54.9	51.6	43.1	44.5	$58.2^{\pm0.6}$	$\mathbf{58.5^{\pm0.6}}$	$58.2^{\pm0.6}$	$\mathbf{58.1^{\pm0.6}}$	$58.6^{\pm0.6}$	$58.8^{\pm0.6}$	$58.3^{\pm0.6}$	$58.5^{\pm0.6}$
NorwAI-Mistral-7B	**56.1**	31.0	**52.8**	50.9	$58.2^{\pm0.6}$	$58.1^{\pm0.6}$	$58.1^{\pm0.6}$	$57.7^{\pm0.7}$	$58.4^{\pm0.6}$	$58.7^{\pm0.6}$	$58.5^{\pm0.6}$	$58.2^{\pm0.6}$
NorwAI-Llama2-7B	35.5	23.9	23.6	44.8	$55.8^{\pm0.7}$	$57.0^{\pm0.6}$	$57.3^{\pm0.6}$	$56.1^{\pm0.7}$	$57.8^{\pm0.6}$	$57.9^{\pm0.6}$	$57.8^{\pm0.6}$	$57.5^{\pm0.6}$
NorMistral-7B-warm	54.7	53.1	0.0	**51.6**	$55.9^{\pm0.7}$	$56.5^{\pm0.6}$	$56.6^{\pm0.6}$	$56.2^{\pm0.7}$	$57.0^{\pm0.7}$	$57.0^{\pm0.6}$	$57.2^{\pm0.7}$	$56.4^{\pm0.6}$
NorGPT-3B	0.1	0.1	0.1	0.2	$0.0^{\pm0.0}$	$0.2^{\pm0.0}$	$0.1^{\pm0.0}$	$0.4^{\pm0.0}$	$0.2^{\pm0.0}$	$1.8^{\pm0.1}$	$0.7^{\pm0.0}$	$0.9^{\pm0.0}$
Viking-7B	53.4	35.0	42.1	1.4	$54.2^{\pm0.6}$	$58.3^{\pm0.7}$	$57.2^{\pm0.7}$	$56.7^{\pm0.7}$	$58.7^{\pm0.7}$	$59.6^{\pm0.6}$	$59.7^{\pm0.6}$	$59.4^{\pm0.7}$
Viking-13B	24.2	**58.2**	14.9	1.9	$58.6^{\pm0.7}$	$58.3^{\pm0.7}$	$\mathbf{58.7^{\pm0.8}}$	$57.5^{\pm0.7}$	$59.5^{\pm0.6}$	$\mathbf{60.0^{\pm0.6}}$	$59.9^{\pm0.6}$	$59.9^{\pm0.7}$
Mistral-Nemo-12B	44.3	46.1	44.3	45.5	$48.5^{\pm0.6}$	$48.9^{\pm0.6}$	$48.9^{\pm0.6}$	$48.9^{\pm0.6}$	$49.3^{\pm0.6}$	$49.5^{\pm0.6}$	$49.2^{\pm0.6}$	$49.5^{\pm0.6}$

CHRF++ SCORES

Prompt template	0-shot				1-shot				16-shot			
	A	B	C	D	A	B	C	D	A	B	C	D
NorMistral-11B	**71.7**	71.4	67.2	69.3	$\mathbf{73.6^{\pm0.4}}$	$73.8^{\pm0.4}$	$\mathbf{73.8^{\pm0.4}}$	$\mathbf{73.4^{\pm0.4}}$	$73.8^{\pm0.4}$	$74.0^{\pm0.4}$	$73.8^{\pm0.4}$	$73.7^{\pm0.4}$
NorwAI-Mistral-7B	71.1	55.5	**68.8**	**70.8**	$73.0^{\pm0.4}$	$73.2^{\pm0.4}$	$73.2^{\pm0.4}$	$72.6^{\pm0.5}$	$73.4^{\pm0.4}$	$73.7^{\pm0.4}$	$73.6^{\pm0.4}$	$73.4^{\pm0.4}$
NorwAI-Llama2-7B	59.3	40.8	39.9	64.1	$71.3^{\pm0.4}$	$72.1^{\pm0.4}$	$72.4^{\pm0.4}$	$71.2^{\pm0.5}$	$72.7^{\pm0.4}$	$72.9^{\pm0.4}$	$72.9^{\pm0.4}$	$72.6^{\pm0.4}$
NorMistral-7B-warm	69.6	67.5	9.6	66.3	$70.9^{\pm0.4}$	$71.6^{\pm0.4}$	$71.6^{\pm0.4}$	$71.3^{\pm0.4}$	$72.0^{\pm0.4}$	$72.3^{\pm0.4}$	$72.2^{\pm0.4}$	$72.0^{\pm0.4}$
NorGPT-3B	9.8	8.1	9.7	9.5	$4.4^{\pm0.1}$	$4.8^{\pm0.1}$	$4.4^{\pm0.1}$	$6.4^{\pm0.1}$	$4.4^{\pm0.1}$	$12.3^{\pm0.3}$	$7.9^{\pm0.1}$	$8.9^{\pm0.1}$
Viking-7B	70.8	51.8	63.6	12.5	$70.6^{\pm0.4}$	$72.9^{\pm0.7}$	$72.3^{\pm0.5}$	$71.9^{\pm0.5}$	$73.3^{\pm0.5}$	$74.3^{\pm0.4}$	$74.4^{\pm0.4}$	$74.1^{\pm0.5}$
Viking-13B	59.4	**72.8**	46.8	15.5	$73.3^{\pm0.5}$	$\mathbf{74.0^{\pm0.4}}$	$73.6^{\pm0.5}$	$72.6^{\pm0.5}$	$74.2^{\pm0.4}$	$\mathbf{74.6^{\pm0.4}}$	$74.5^{\pm0.4}$	$74.5^{\pm0.4}$
Mistral-Nemo-12B	61.4	64.2	61.7	63.4	$65.8^{\pm0.5}$	$66.5^{\pm0.4}$	$66.4^{\pm0.4}$	$66.6^{\pm0.4}$	$66.8^{\pm0.4}$	$67.0^{\pm0.4}$	$66.9^{\pm0.4}$	$67.0^{\pm0.4}$

Table 16: **Complete results on translation from English to Norwegian Bokmål** We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.10.2 English to Nynorsk translation

Prompt templates We used the following five prompt templates for testing all language models on translation to Nynorsk.

Prompt A:

```
1   Engelsk: {$text}
2   Nynorsk: {$prediction}
```

Prompt B:

```
1   Omsett følgande setning til Nynorsk: {$text}
2   Nynorsk: {$prediction}
```

Prompt C:

```
1   Gje ei Nynorsk omsetjing av denne setninga: {$text}
2   Nynorsk: {$prediction}
```

Prompt D:

```
1    Kva blir "{$text}" på Nynorsk?
2    Nynorsk: {$prediction}
```

Full results   The complete evaluation on translation to Nynorsk is given in Table 18 (BLEU scores in the top sub-table and chrF++ scores below).

BLEU SCORES

	0-shot				1-shot				16-shot			
Prompt template	A	B	C	D	A	B	C	D	A	B	C	D
NorMistral-11B	36.2	6.9	20.1	39.3	$46.3^{\pm1.6}$	$45.4^{\pm1.5}$	$45.2^{\pm1.4}$	$44.7^{\pm1.5}$	$46.5^{\pm1.6}$	$\mathbf{48.0^{\pm1.6}}$	$46.1^{\pm1.6}$	$\mathbf{47.3^{\pm1.6}}$
NorwAI-Mistral-7B	**46.0**	**44.7**	**42.3**	**40.0**	$\mathbf{46.7^{\pm1.6}}$	$\mathbf{46.7^{\pm1.6}}$	$46.6^{\pm1.7}$	$\mathbf{45.9^{\pm1.6}}$	$47.4^{\pm1.8}$	$46.5^{\pm1.8}$	$46.1^{\pm1.6}$	$46.5^{\pm1.7}$
NorwAI-Llama2-7B	43.9	23.2	0.0	28.8	$45.9^{\pm1.7}$	$46.7^{\pm1.8}$	$\mathbf{46.7^{\pm1.7}}$	$45.2^{\pm1.7}$	$\mathbf{47.4^{\pm1.7}}$	$47.2^{\pm1.7}$	$\mathbf{47.3^{\pm1.8}}$	$47.0^{\pm1.8}$
NorMistral-7B-warm	43.5	31.2	15.2	11.7	$43.7^{\pm1.7}$	$44.6^{\pm1.8}$	$43.5^{\pm1.6}$	$43.5^{\pm1.8}$	$43.6^{\pm1.8}$	$44.7^{\pm1.7}$	$43.9^{\pm1.7}$	$44.2^{\pm1.6}$
NorGPT-3B	1.5	2.4	0.8	1.6	$0.1^{\pm0.0}$	$0.2^{\pm0.0}$	$0.1^{\pm0.0}$	$0.4^{\pm0.1}$	$0.2^{\pm0.0}$	$0.7^{\pm0.1}$	$2.6^{\pm0.5}$	$0.8^{\pm0.1}$
Viking-7B	26.7	44.3	16.5	1.9	$45.0^{\pm1.7}$	$44.4^{\pm1.6}$	$43.7^{\pm1.7}$	$42.3^{\pm1.6}$	$44.5^{\pm1.6}$	$43.9^{\pm1.5}$	$44.5^{\pm1.6}$	$45.6^{\pm1.6}$
Viking-13B	42.5	31.6	11.1	1.7	$45.2^{\pm1.7}$	$45.2^{\pm1.7}$	$44.8^{\pm1.7}$	$42.4^{\pm1.6}$	$45.2^{\pm1.7}$	$45.1^{\pm1.6}$	$45.5^{\pm1.7}$	$45.6^{\pm1.7}$
Mistral-Nemo-12B	33.0	33.2	33.9	29.2	$33.6^{\pm1.5}$	$34.7^{\pm1.4}$	$33.9^{\pm1.3}$	$35.1^{\pm1.5}$	$35.6^{\pm1.6}$	$35.4^{\pm1.7}$	$35.3^{\pm1.7}$	$35.7^{\pm1.6}$

CHRF++ SCORES

	0-shot				1-shot				16-shot			
Prompt template	A	B	C	D	A	B	C	D	A	B	C	D
NorMistral-11B	62.1	32.4	53.3	62.2	$65.1^{\pm1.1}$	$64.6^{\pm1.1}$	$64.4^{\pm1.0}$	$63.9^{\pm1.1}$	$65.2^{\pm1.1}$	$\mathbf{66.5^{\pm1.2}}$	$65.7^{\pm1.1}$	$65.9^{\pm1.1}$
NorwAI-Mistral-7B	**64.9**	62.6	**60.8**	**63.4**	$\mathbf{65.4^{\pm1.1}}$	$64.9^{\pm1.1}$	$64.8^{\pm1.1}$	$\mathbf{64.5^{\pm1.1}}$	$\mathbf{65.7^{\pm1.2}}$	$65.1^{\pm1.2}$	$65.0^{\pm1.1}$	$65.1^{\pm1.1}$
NorwAI-Llama2-7B	63.5	41.9	3.4	45.7	$64.1^{\pm1.2}$	$64.8^{\pm1.1}$	$\mathbf{64.8^{\pm1.1}}$	$63.5^{\pm1.1}$	$\mathbf{65.7^{\pm1.2}}$	$65.8^{\pm1.1}$	$\mathbf{65.7^{\pm1.2}}$	$65.2^{\pm1.2}$
NorMistral-7B-warm	62.3	48.1	32.6	29.0	$62.7^{\pm1.3}$	$63.6^{\pm1.3}$	$63.0^{\pm1.2}$	$63.0^{\pm1.3}$	$63.6^{\pm1.2}$	$64.6^{\pm1.1}$	$64.4^{\pm1.1}$	$63.9^{\pm1.2}$
NorGPT-3B	15.5	16.6	12.2	16.9	$4.9^{\pm0.2}$	$4.4^{\pm0.2}$	$3.3^{\pm0.2}$	$7.7^{\pm0.5}$	$4.0^{\pm0.2}$	$9.2^{\pm0.5}$	$15.0^{\pm1.1}$	$8.0^{\pm0.4}$
Viking-7B	56.9	**63.6**	47.8	15.1	$64.1^{\pm1.2}$	$64.4^{\pm1.1}$	$64.6^{\pm1.1}$	$62.8^{\pm1.2}$	$64.5^{\pm1.1}$	$64.3^{\pm1.0}$	$64.5^{\pm1.1}$	$65.2^{\pm1.1}$
Viking-13B	62.9	60.4	44.0	14.9	$64.7^{\pm1.1}$	$64.8^{\pm1.1}$	$64.4^{\pm1.1}$	$61.9^{\pm1.2}$	$64.6^{\pm1.1}$	$64.9^{\pm1.1}$	$65.0^{\pm1.1}$	$65.3^{\pm1.1}$
Mistral-Nemo-12B	55.4	54.2	55.5	53.8	$56.3^{\pm1.1}$	$57.0^{\pm1.1}$	$56.4^{\pm1.0}$	$56.9^{\pm1.1}$	$57.3^{\pm1.2}$	$57.2^{\pm1.2}$	$57.2^{\pm1.2}$	$57.5^{\pm1.2}$

Table 17: **Complete results on translation from English to Nynorsk**   We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.

B.10.3   English to Northern Sámi translation

We source the data from the English-Sámi parallel corpus from Tatoeba (Tiedemann, 2020), specifically the latest v2023-04-12 revision avaiable on HuggingFace at **https://hf.co/datasets/Helsinki-NLP/tatoeba**. We deduplicate this corpus (both on the source and target side) and remove the empty entries – obtaining 53 examples in total.

Prompt templates   We used the following five prompt templates for testing all language models on translation to Northern Sámi.

Prompt A:

```
1    Eaŋgalsgiella: {$text}
2    Davvisámegiella: {$prediction}
```

Prompt B:

```
1    Engelsk: {$text}
2    Samisk: {$prediction}
```

Prompt C:

```
1    Oversett følgende setning til nordsamisk: {$text}
2    Nordsamisk: {$prediction}
```

Prompt D:

```
1    Gi en oversettelse til nordsamisk for denne setningen: {$text}
2    Nordsamisk: {$prediction}
```

Prompt E:

```
1    Hva blir "{$text}" på nordsamisk?
2    Nordsamisk: {$prediction}
```

Full results The complete evaluation on translation to Sámi is given in Table 18 (BLEU scores in the top sub-table and chrF++ scores below).

BLEU SCORES

Prompt template	0-shot A	B	C	D	E	1-shot A	B	C	D	E	16-shot A	B	C	D	E
NorMistral-11B	**31.5**	**31.0**	14.8	**23.2**	14.7	24.8±5.1	**44.1±1.2**	15.7±1.6	20.2±9.4	33.7±8.2	45.5±2.2	48.8±2.4	49.0±2.1	48.8±2.3	**50.4±0.9**
NorwAI-Mistral-7B	21.2	25.0	**24.0**	16.7	**20.7**	**26.6±1.7**	24.7±2.2	24.4±1.6	**24.2±2.4**	25.0±1.1	24.8±2.7	25.1±2.6	26.2±1.8	26.8±2.0	27.5±2.3
NorwAI-Llama2-7B	16.9	10.7	3.3	0.0	15.4	24.7±2.1	23.5±2.7	**24.5±2.1**	22.6±2.5	22.5±1.8	26.6±2.3	27.6±2.0	27.9±1.9	27.6±2.1	28.5±1.6
NorMistral-7B-warm	12.2	5.7	0.0	0.0	0.0	11.2±3.1	10.5±1.9	12.2±3.2	10.5±2.0	8.4±0.5	14.9±2.2	17.2±1.9	17.9±1.5	15.9±1.2	18.5±2.5
NorGPT-3B	0.0	0.0	0.0	0.0	0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Viking-7B	0.0	0.0	0.0	0.0	0.0	0.0±0.0	1.0±2.3	0.7±1.5	0.7±1.6	0.5±1.1	4.6±2.6	6.8±1.8	6.7±2.6	7.8±1.6	4.3±3.9
Viking-13B	0.0	0.0	0.0	0.0	0.0	0.3±0.7	3.3±3.2	2.3±2.1	4.3±4.0	1.8±1.9	9.8±1.1	11.6±1.9	11.5±1.9	11.4±2.0	11.2±1.2
Mistral-Nemo-12B	0.0	0.0	0.0	0.0	0.0	0.0±0.0	1.4±2.0	0.8±1.8	0.9±2.0	0.9±1.9	3.9±2.5	6.5±2.2	4.9±3.2	5.7±3.8	6.1±1.4

CHRF++ SCORES

Prompt template	0-shot A	B	C	D	E	1-shot A	B	C	D	E	16-shot A	B	C	D	E
NorMistral-11B	51.8	**58.8**	49.5	**58.2**	47.3	**55.5±2.4**	**66.7±0.8**	**54.6±1.8**	**56.9±5.4**	63.2±3.7	64.9±1.2	67.2±0.6	67.9±1.0	67.7±1.4	**69.6±0.9**
NorwAI-Mistral-7B	44.9	45.8	**53.9**	50.6	**49.6**	50.7±1.3	50.8±1.5	50.8±0.8	51.5±1.1	50.4±0.7	51.0±1.7	51.7±1.7	52.3±1.7	52.4±1.2	52.8±1.6
NorwAI-Llama2-7B	37.3	30.0	46.8	44.1	39.3	45.6±2.0	44.2±3.4	48.1±2.3	46.6±1.7	44.0±1.4	48.4±1.9	50.2±1.3	50.4±1.6	50.2±1.4	50.4±1.1
NorMistral-7B-warm	27.4	22.5	21.4	26.8	23.8	32.3±2.4	30.9±1.9	32.7±2.2	32.7±1.5	29.9±0.7	38.1±3.7	37.7±3.5	38.9±1.5	37.2±1.5	39.4±2.1
NorGPT-3B	2.7	3.0	3.0	2.8	2.8	2.8±0.1	2.8±0.1	2.8±0.1	2.8±0.1	2.9±0.1	2.9±0.1	2.6±0.1	3.0±0.1	3.0±0.1	3.1±0.1
Viking-7B	10.4	11.5	11.5	11.7	8.1	10.3±0.9	13.0±1.7	12.3±1.5	12.7±1.1	12.4±1.3	18.0±1.3	19.6±0.7	19.1±1.3	19.7±1.0	19.5±1.6
Viking-13B	11.3	10.8	8.0	10.6	6.2	10.9±0.7	14.9±1.4	16.2±0.8	16.1±0.7	15.2±1.5	24.5±2.5	24.5±2.0	24.9±2.2	24.0±2.9	25.2±1.7
Mistral-Nemo-12B	11.5	12.1	12.2	12.4	10.5	14.2±2.0	14.2±0.9	15.3±1.1	14.8±1.3	14.9±0.9	21.9±0.8	22.3±1.7	22.6±1.5	22.4±1.6	23.0±1.5

Table 18: **Complete results on translation from English to Northern Sámi** We show the detailed results for each evaluated model, few-shot setting and prompt template. As the few-shot demonstrations are sampled randomly, we repeat them five times and show the mean accuracy as well as the standard deviation (rendered as superscript). The best results for each column are boldfaced, the overall best result for each few-shot setting is highlighted in blue.