SnakModel

Lessons Learned from Training an Open Danish Large Language Model

Mike Zhang **Max Müller-Eberstein** **Elisa Bassignana**
Rob van der Goot
Aalborg University, Denmark
IT University of Copenhagen, Denmark
Pioneer Center for Artificial Intelligence, Denmark
jjz@cs.aau.dk {mamy, elba, robv}@itu.dk

Abstract

We present SNAKMODEL, a Danish large language model (LLM) based on LLAMA2-7B, which we continuously pre-train on 13.6B Danish words, and further tune on 3.7M Danish instructions. As best practices for creating LLMs for smaller language communities have yet to be established, we examine the effects of early modeling and training decisions on downstream performance throughout the entire training pipeline, including (1) the creation of a strictly curated corpus of Danish text from diverse sources; (2) the language modeling and instruction tuning training process itself, including the analysis of intermediate training dynamics, and ablations across different hyperparameters; (3) an evaluation on eight language and culturally-specific tasks. Across these experiments SNAKMODEL achieves the highest overall performance, outperforming multiple contemporary LLAMA2-7B-based models. By making SNAKMODEL, the majority of our pre-training corpus, and the associated code available under open licenses, we hope to foster further research and development in Danish Natural Language Processing, and establish training guidelines for languages with similar resource constraints.[1]

1 Introduction

The landscape of large language models (LLMs) has seen rapid expansion, with an increasing trend towards open-weight releases for a broader range of languages. Notable English-centric examples include Pythia (Biderman et al., 2023), Vicuna (Zheng et al., 2023), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), Llama2 (Touvron et al., 2023), Llama3 (Dubey et al., 2024), OLMo (Groeneveld et al., 2024), and Phi (Abdin et al., 2024). Simultaneously, recent efforts have extended LLMs to multilingual settings, including models such as mT5 (Xue et al., 2021), Bloom (Le Scao et al., 2023), Aya (Üstün et al., 2024; Singh et al., 2024), RomanSetu (J et al., 2024), and EuroLLM (Martins et al., 2024).

As anglocentric and/or multilingual LLMs have nonetheless continued struggling to adapt to lower-resource settings—especially with respect to pragmatic and sociolinguistic factors (Hershcovich et al., 2022; Cao et al., 2023; Naous et al., 2024; Wang et al., 2024)—there is growing interest in language-specific LLMs, either tailored to a single language (see Related Work; Section 2) or specialized for a small set of similar languages (SiloAI, 2024; Dou et al., 2024). However, the best practices for creating such language-adapted LLMs have yet to be established—especially for smaller language communities with resource limitations with respect to data, compute, or both.

Danish offers a particularly interesting testbed among these smaller languages. As a mid-resource language, which is typologically related to English and has largely overlapping character sets, it has sufficient textual data for LLM adaptation, yet is far from the levels of its neighbors (e.g., Swedish; Ekgren et al., 2024). Additionally, it lacks advanced resources like native instruction-tuning data or human-preference data, making it necessary to use translated datasets for which the downstream effects on model functionality are not yet well understood. Linguistically, Danish has also been shown

These authors contributed equally.
[1]The code and data scripts are available here: https://github.com/nlpnorth/snakmodel/.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 812–825
March 3-4, 2025

to be more challenging to learn for humans than its neighbors due its phonological complexity (Trecca et al., 2021; Christiansen et al., 2023), which results in downstream effects on discourse, such as additional conversational redundancy (Christiansen et al., 2023; Dideriksen et al., 2023).

With the goal to provide the Danish community with a custom-adapted resource, as well as to establish better-grounded guidelines for creating LLMs in languages with similar linguistic characteristics and resource constraints, we present and analyze SNAKMODEL-7B$_{\text{base/instruct}}$, two LLMs designed specifically for the Danish language. Our base model builds upon LLAMA2-7B, which we continuously pre-train on a diverse collection of Danish corpora comprising 350M documents (sentences/paragraphs) and 13.6B words, before tuning it on 3.7M Danish instruction-answer pairs. We evaluate our model against contemporary LLAMA2-7B-based models on the Danish part of the ScandEval benchmark (Nielsen, 2023) that encompasses both language and culture-specific tasks. By releasing not just the related artifacts (final model, intermediate checkpoints, pre-training data, code), but by also analyzing the effects of early decisions in the training and model design process on intermediate training dynamics and downstream performance, we aim to provide resources that are not just relevant for Danish, but for LLM adaptation in general.

Contributions. This work contributes:

- A large, diverse, high-quality collection of Danish corpora, totaling 350M documents with 13.6B words (Section 3). We provide scripts to collect and process the data.
- SNAKMODEL-7B$_{\text{base/instruct}}$, two open-weight 7B-parameter language models continuously pre-trained and instruction-tuned specifically for Danish, for which we release all related artefacts, and extensively analyze the model's intermediate training dynamics (Section 4).
- An evaluation comparing SNAKMODEL-7B$_{\text{instruct}}$ and contemporary Danish models, which analyzes performance with respect to language and cultural tasks (Section 5).
- A consolidation of our findings into recommendations for efficiently training LLMs under similar resource constraints (Section 6).

2 Related Work

Continuously Pre-trained LLMs. Previous work has shown that for both encoder and decoder language models (LM), continuous pre-training is the de facto standard for adapting an LM to a specific domain (Han and Eisenstein, 2019; Alsentzer et al., 2019; Lee et al., 2020; Gururangan et al., 2020; Nguyen et al., 2020) or another language, such as German (LeoLM-Team, 2024), Spanish and Catalan (Ǎguila Team, 2023), Finnish (Luukkonen et al., 2023), Dutch (Rijgersberg and Lucassen, 2023; Vanroy, 2024), Italian (Bacciu et al., 2024), Japanese (Rakuten Group et al., 2024), Basque (Etxaniz et al., 2024), Swedish (AI-Sweden, 2024), Modern Greek (Voukoutis et al., 2024), Norwegian (NORA.LLM-Team, 2024), or multiple languages (Xue et al., 2021; Alves et al., 2024; Üstün et al., 2024; Costa-jussà et al., 2022; Martins et al., 2024; Dou et al., 2024; Nguyen et al., 2024; Aryabumi et al., 2024; Dang et al., 2024).

Open Large Language Models. Recent open language models can be broadly divided into *open-source* LLMs and *open-weight* LLMs. The main difference is that open-weight releases include at least a basic description of the training data, as well as the model weights themselves. For open-source LLMs, instead, the (non-trivial) expectation is to have all resources released, including data, training scripts, evaluation scripts, and model weights. We follow previous endeavors such as Pythia (Biderman et al., 2023), OLMo (Groeneveld et al., 2024), Latxa (Etxaniz et al., 2024), and Meltemi (Voukoutis et al., 2024), and release most sources of our training data, including training and evaluation scripts, as well as the model weights.

Danish Language Resources. In-language resources are the fundamental building block for further training an LLM for the Danish language. There are several open-source toolkits for Danish, including models and datasets (Pauli et al., 2021; Enevoldsen et al., 2021). Additionally, there are several Danish-specific large corpora of raw text, such as DaNewsroom (Varab and Schluter, 2020) and Danish Gigaword (Strømberg-Derczynski et al., 2021). Additionally, Danish subsets can be found in public resources built on crawled web data such as CommonCrawl (Wenzek et al., 2020) and CulturaX (Nguyen et al., 2023). In this work, we collect and combine a variety of sources for wider coverage, before pre-processing them through a joint pipeline.

Danish Large Language Models. Previous endeavors at training LLMs that cover the Danish language include Ciosici and Derczynski (2022), who trained a T5 model (Raffel et al., 2020) for Danish. More recently, within the decoder-only family of models, Munin (Danish-Foundation-Models-Team, 2024) and Viking (SiloAI, 2024) were released. Munin is based on Mistral-7B (v0.1 Jiang et al., 2023) and is further pre-trained on the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) containing 1B words. However, the model seems to underperform compared to its base model counterpart, indicating some form of catastrophic forgetting. Viking is based on LLAMA2-7B, and pre-trained from scratch on a mix of English, Finnish, Swedish, Danish, Norwegian, Icelandic and code (SiloAI, 2024). In this work, SNAKMODEL-7B$_{\text{instruct}}$ is continuously pre-trained for Danish, and outperforms its original checkpoint, as well as all other currently available Danish models with a comparable size.

3 Data & Pre-processing

3.1 Pre-training

Our Danish pre-training data, as shown in Table 1, initially encompassed 927M documents and 24.6B words, as measured by the Unix `wc` command. The data is sourced from diverse platforms, for which we verify appropriate licensing (wherever possible), and include:

Bookshop (`cc-by-4.0`). EU Bookshop text from OPUS (Tiedemann, 2012), as integrated by Skadiņš et al. (2014). It contains well-edited, official EU publications across diverse topics, converted automatically from PDFs.

CC-100 (`UNK`). A cleaned version of a 2018 CommonCrawl dump (Wenzek et al., 2020), reproducing data from Conneau et al. (2020). It consists of web data, filtered using the fastText language classifier (Joulin et al., 2017).

CulturaX (`odc-by + cc0`). mC4 (v3.1.0) combined with accessible OSCAR corpora (Nguyen et al., 2023).

DaNewsroom (`UNK`). Scraped from 19 news outlets (Varab and Schluter, 2020), originally for summarization. We use the full news articles instead of summaries.

DATASET	ORIGINAL		+ FASTTEXT	
	Docs	Words	Docs	Words
Bookshop	8.65M	208M	6.80M	187M
CC-100	344M	7.82B	256M	7.16B
CulturaX	449M	14.8B	333M	13.7B
DaNewsroom	24.2M	391M	11.3M	369M
Dawiki	1.70M	62.4M	1.20M	57.3M
FTSpeech	2.03M	43.3M	1.69M	40.9M
Gigaword	62.0M	1.02B	39.3M	898M
OpenSubtitles	30.2M	207M	19.6M	156M
Reddit	4.50M	73.9M	2.37M	64.0M
Twitter	1.69M	21.9M	406K	6.61M
TOTAL	927M	24.6B	672M	22.6B
+ DEDUPLICATION			**350M**	**13.6B**

Table 1: **Preprocessing Steps.** Data in number of words using `wc` command. In the **Original** column, we already use a pre-defined Danish slice of the dataset. In the **FastText** column, we apply another round of language identification to the data. In the **Deduplication** row, we combine all data and deduplicate it, which results in around 350M documents and 13.6B words for the pre-training process.

Dawiki (`cc-by-sa`). Cleaned Wikipedia data from 01-01-2024 (Attardi, 2015).

FTSpeech (`FT-OD + FT-TV`). A transcription-based corpus from Danish parliamentary data (Kirkedal et al., 2020), used in language modeling due to its large text volume.[2]

Gigaword (`cc0 + cc-by`). Danish Gigaword (Strømberg-Derczynski et al., 2021) covers a range of domains including wiki, books, web, and social media data.

OpenSubtitles (`UNK`). Danish data from OPUS OpenSubtitles (Lison and Tiedemann, 2016; Tiedemann, 2016).[3]

Reddit (`UNK`). Danish Reddit data from ConvoKit (Chang et al., 2020), specifically `Denmark.corpus.zip`.

Twitter (`MIT`). Data from the public Twitter stream,[4] reclassified using our own pipeline due to inaccurate language labels.

To refine the overall concatenated dataset, we implemented a preprocessing pipeline using `fastText` (Joulin et al., 2017)[5] for language iden-

[2]FT-OD and FT-TV refer to Folketing's open data and Folketing TV license.

[3]`http://www.opensubtitles.org/`

[4]`https://archive.org/details/twitterstream`

[5]Using the `lid.176.bin` model with a threshold of 0.6.

tification and `text-dedup` (Mou et al., 2023)[6] for text deduplication. The language identification process eliminated 28% of the documents while retaining 92% of the tokens, indicating that many short documents were removed, where language prediction was less confident. The deduplication step further reduced the corpus by 48% in document count and 40% in token count. We anticipated significant content overlap between CC-100 and CulturaX, which underlines the importance of deduplication in creating a more efficient and representative dataset. These preprocessing steps reduced our dataset to approximately 350M documents with 13.6B words. Following the open LLM approach, we release all scripts used for collecting and processing the data.

3.2 Instruction Tuning

As for most mid-to-low resource languages, Danish (Joshi et al., 2020) currently lacks human-generated instruction tuning data, and instead relies on automatically translated data from English, which itself may be generated by LLMs. From these sources, we select the following three after manually inspecting them for quality:

SkoleGPT (Professionshøjskole, 2024) : A subset of OpenOrca (Lian et al., 2023), which was automatically translated into Danish and filtered for quality, containing 21.6k instruction-output pairs.

Danish OpenHermes (Mabeck, 2024) : A subset of the automatically generated OpenHermes dataset (“Teknium”, 2023), which was automatically translated into Danish. It contains 98.7k instruction-output pairs.

Aya Collection (Singh et al., 2024) : A collection of 44 datasets, which were automatically translated based on instruction templates from fluent speakers. While the underlying Aya Dataset, on which these translations are based, was created by native speakers, the Danish portion of this data contains less than 100 instances, leading us to opt for the translations instead. We use 3.6M instruction-output pairs from the Danish subset of the data.

Together, these data sources sum up to a total of 3.7M instruction-answer pairs, which we train SNAKMODEL-7B$_{\text{base}}$ on in Section 4.2.

[6]`https://github.com/ChenghaoMou/text-dedup`

Parameter	Value
Data Split	
Training data	96.9%
Validation data	3.1%
Training Configuration	
Vocabulary size	32,000
Context length	4,096
Training steps	12,500
Warmup steps	1,250
Number of epochs	1
Global batch size	512
Optimizer Parameters (AdamW)	
β_1; β_2	0.9; 0.95
ϵ	10^{-5}
Peak learning rate	1.5×10^{-5}
Minimum learning rate	5×10^{-8}
Weight decay	0.1
Gradient clipping	1.0

Table 2: **Pre-training Hyperparameters and Configuration Details.** We show the hyperparameter details of SNAKMODEL-7B$_{\text{base}}$ pre-training.

3.3 Evaluation Framework

For evaluation, we use the SCANDEVAL benchmark (Nielsen, 2023) covering eight tasks. The tasks cover named entity recognition (NER; DANSK by Hvingelby et al., 2020), sentiment analysis (SENTI; AngryTweets by Pauli et al., 2021), linguistic acceptability (LA; ScaLA[7]), abstractive summarization (SUMM; Nordjylland-News by Kinch, 2023), commonsense reasoning (CSR; translated HellaSwag by Zellers et al., 2019), and question answering (QA; ScandiQA[8]). The benchmark also include culture-specific datasets, namely Danske Talemåder (TM; Nielsen, 2023), which prompts for meanings behind common proverbs, and a collection of official Danish Citizenship Tests (CT; Nielsen, 2024). Evaluation metrics differ per task, and are indicated as F_1, macro-averaged F_1 (mF_1), micro-averaged F_1 (μF_1), BERTScore (BERTS.; Zhang et al., 2020), and Accuracy (Acc.).

4 Model Training

4.1 Language Modeling Pre-training

Training Details. We continuously pre-train from LLAMA2-7B$_{\text{base}}$ (Touvron et al., 2023). We show configuration and hyperparameter details

[7]Based on the Universal Dependencies dataset from (Kromann and Lynge, 2004).

[8]ScandiQA is a translation of the English MKQA dataset (Longpre et al., 2021) and does not strictly focus on Scandinavian knowledge.

in Table 2. For further pre-training and fine-tuning, we make use of the Megatron-LLM library (Cano et al., 2023), based on the Megatron-LM library.[9] We use the same tokenizer as LLAMA2-7B, byte-pair encoding (BPE; Sennrich et al., 2016) as implemented in the SentencePiece toolkit (Kudo and Richardson, 2018), with a vocabulary size of 32K subwords. As Danish and English share the same Indo-European language family, we assume large overlap in vocabulary subwords. Hence, we do not re-train nor extend the vocabulary.

Hardware and Emissions. SNAKMODEL-7B$_{\text{base}}$ is trained on private infrastructure with one node, containing four NVIDIA A100-PCIe 40GB GPUs. The node has an AMD Epyc 7662 128 Core Processor and 1TB of RAM. Total time of training took 8,928 GPU hours (93 days $\times$ 24 hours $\times$ 4 GPUs) between March–June 2024. The average carbon efficiency was 0.122 $kgCO_2eq/kWh$ during this time in Denmark.[10] This is equivalent to 272.3 $kg\ CO_2\ eq.$ emitted, based on the Machine Learning Impact calculator (Lacoste et al., 2019).[11]

Loss Trajectories. In Figure 1, we show the continuous pre-training process of SNAKMODEL-7B$_{\text{base}}$ in terms of loss curve based on perplexity. The loss shows a declining gain over time. We speculate that the model is close to convergence or that the learning rate is reduced, although previous work has shown that downstream performance can still increase with more training after loss and perplexity have converged (Liu et al., 2023).

Leakage. The training data of LLAMA2-7B is not public. However, since it was released in July 2023 after the ScandEval benchmark, we investigate potential test data leakage by prompting the model for information about the dataset (inspired by Sainz et al., 2023; Balloccu et al., 2024), as well as completions for the first five sentences of each dataset. This process yielded no evidence that the evaluation datasets were included during training.

For SNAKMODEL-7B$_{\text{base}}$, we have access to all training data, such that we can search for 200 random 8-grams from each of our datasets in the raw data. We find that a small amount (6/200) of the tweets from AngryTweets are included in our Twitter sample (without labels). The DANSK NER dataset was completely included (without labels), as it was sampled from Gigaword, and many parts of the ScaLA dataset were also included in its original form in GigaWord and CC100. The code for all leakage tests is included in our code repository.

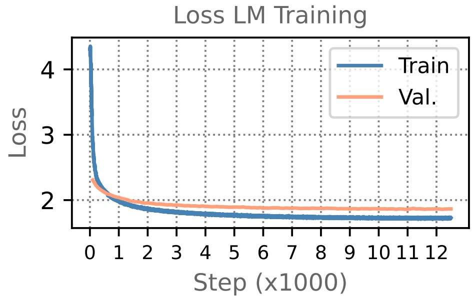


Figure 1: **SNAKMODEL-7B$_{\text{base}}$ Pre-training Behaviour.** We report the stable language model loss during training and validation.

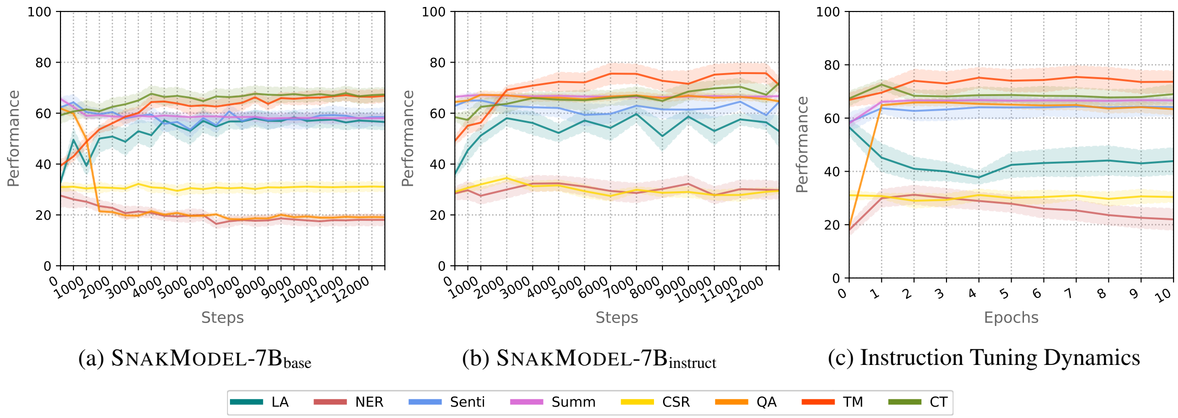
4.2 Instruction Tuning

Starting from SNAKMODEL-7B$_{\text{base}}$, we train our model on the Danish instruction datasets outlined in Section 3.2.

Training Details. For instruction tuning, we opt for the more parameter-efficient low-rank adaptation (LoRA; Hu et al., 2022), to enable faster iterations across multiple ablations (different template formats and base models), and to more easily analyze the intermediate training dynamics (Section 4.3). Nonetheless, we choose a substantially higher-parameter setup than is commonly employed when using LoRA (Hu et al., 2022; Dettmers et al., 2023), in order to approximate full fine-tuning as closely as possible given our computational budget. Specifically, we use rank $r = 128$ adaptation matrices, which are applied to all parameters within the model without quantization (Dettmers et al., 2023). We train for one epoch over our instruction data using the AdamW optimizer with a constant learning rate of 2×10^{-4}, and a global batch size of 64.

Instruction Template. The formatting of instruction-answer pairs is an important design decision with significant downstream impacts (Sclar et al., 2024). For our adaptation context (LLAMA2-7B + Danish), we therefore ablate across three templates: (1) CONCAT, which concatenates instructions and answers; (2) CHAT, which wraps the instruction in special `[INST]`/`[/INST]`

[9] https://github.com/NVIDIA/Megatron-LM.
[10] According to https://app.electricitymaps.com/map.
[11] https://mlco2.github.io/impact.



(a) SNAKMODEL-7B$_{\text{base}}$ (b) SNAKMODEL-7B$_{\text{instruct}}$ (c) Instruction Tuning Dynamics

Figure 2: **SNAKMODEL Training Dynamics** of LM pre-training, instruction tuning, and multi-epoch instruction tuning, as measured on the ScandEval (validation) tasks of linguistic acceptability (LA), named entity recognition (NER), sentiment analysis (SENTI), summarization (SUMM), commonsense reasoning (CSR), question answering (QA), proverb meaning (TM), and citizenship tests (CT).

delimiters following LLAMA2-7B$_{\text{chat}}$[12]; (3) ALPACA, following a multi-line format with instruction/input/answer headers (Wang et al., 2023), which we translate into Danish.

Instruction tuning using the CHAT format leads to the highest overall scores on the validation split of our evaluation benchmark (56.37 avg.). CONCAT performs comparably (55.52 avg.), however we observe that models trained using this template frequently generate continuations to an instruction, instead of an answer. ALPACA performs worst (53.26 avg.), and we observe that when prompting models without correctly terminating the instruction, the CHAT model consistently terminates the instruction on its own (by generating `[/INST]`), while the ALPACA model often struggles to do so.

4.3 Training Dynamics

We next investigate our models' intermediate training dynamics to establish how much language modeling and/or instruction tuning are required to obtain a certain level of performance (evaluated according to Section 3.3), and whether these trajectories differ across task types.[13]

Language Modeling. By tracking the validation performance of the non-instruction-tuned SNAKMODEL-7B$_{\text{base}}$ checkpoints across pre-training, we aim to identify when the English base model begins adapting to Danish. Figure 2a shows performance on the Danish ScandEval tasks from start (LLAMA2-7B$_{\text{base}}$) to finish (SNAKMODEL-7B$_{\text{base}}$). For SENTI, SUMM and CSR, performance remains relatively consistent, while for LA, TM and CT performance gradually increases until 4,000–6,000 steps before converging.

Meanwhile, we see performance decreases for NER and QA, with the latter dropping from 61.9% F1 to around 20% within the first 2,000 steps. We attribute these changes to two respective hypotheses: for NER, answers are enforced to be in JSON-format in ScandEval. As our pre-training data consists exclusively of natural language, the model's output distribution may skew away from tokens such as "{}", required for this task. For QA, we qualitatively observe that SNAKMODEL-7B$_{\text{base}}$ tends to generate continuations to the provided questions, instead of answers. Additionally, it does so in Danish, which may be detrimental to performance, since many answers in QA are English names.

Instruction Tuning. Next, we investigate the effect of applying instruction tuning at different points during Danish pre-training, in order to assess when it starts becoming beneficial. Figure 2b shows the validation performance of intermediate SNAKMODEL-7B$_{\text{base}}$ checkpoints after instruction-tuning, i.e., from LLAMA2-7B$_{\text{base}}$ + INST$_{\text{da}}$ (instruction-tuning on Danish instruction–completion pairs) until our final SNAKMODEL-7B$_{\text{instruct}}$ (fully pre-trained SNAKMODEL-7B$_{\text{base}}$ + INST$_{\text{da}}$). Once again, performance for most tasks

[12]Note that these delimiters are not split by the tokenizer.

[13]The intermediate checkpoints can be found here: https://huggingface.co/NLPnorth/snakmodel-7b-base/tree/main for SNAKMODEL-7B$_{\text{base}}$ and https://huggingface.co/NLPnorth/snakmodel-7b-instruct for SNAKMODEL-7B$_{\text{instruct}}$.

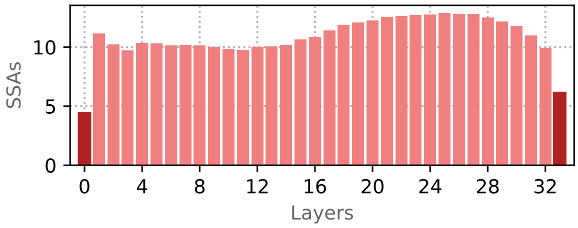


Figure 3: **Layer-wise Weight Divergence of SNAKMODEL-7B$_{\text{base}}$** as measured in total SSAs. Darker bars represent EMB and LMH respectively.

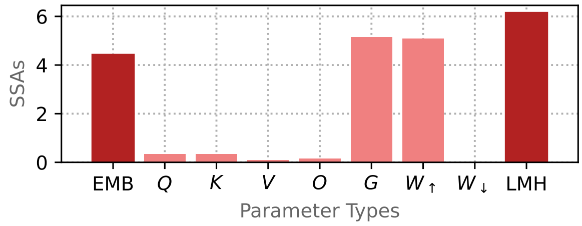


Figure 4: **Parameter-wise Weight Divergence of SNAKMODEL-7B$_{\text{base}}$** as measured in mean SSA. Darker bars represent EMB and LMH respectively.

is surprisingly stable throughout training. We further do not observe the same performance drops for NER and QA as during language modeling pre-training, showing that instruction tuning recovers these original functionalities. Additionally, we observe a general performance increase across the board. In particular, performance for LA, TM, and CT climbs and converges after 2,000–5,000 steps of Danish pre-training, and subsequent instruction-tuning. This indicates that training on less than half of our corpus may already be sufficient to obtain close-to-final performance. Interestingly, the largest performance improvements are observed for benchmark tasks based on Danish data, instead of translations (e.g., LA, TM, CT).

In terms of the training dynamics of instruction tuning itself, Figure 2c shows how one epoch of instruction tuning is already sufficient to obtain most performance gains, including the performance recovery of NER and QA. While there may be some benefit to one or two additional instruction tuning epochs, we believe that at this scale, they can be skipped in favor of efficiency. Since the use of duplicate data across epochs has however also been shown to negatively affect downstream performance (Biderman et al., 2023), we leave the exploration of this trade-off to future work.

Weight Divergence Analysis. Lastly, we take a closer look at changes *within* the model to identify which parameters are most strongly affected by Danish language adaptation. To measure weight divergence, we follow Müller-Eberstein et al. (2024) and measure the principal subspace angles (SSAs; Knyazev and Argentati, 2002) of each parameter before and after adaptatation (0°/90° $\leftrightarrow$ similar/dissimilar). Across layers, Figure 3 shows how there is a slightly higher rate of change towards the penultimate layers of the model. This may be representative of cross-lingual encoding early in the model, and subsequent target language specialization in later layers (Wendler et al., 2024).

Figure 4 provides a more granular view of which parameter types are changing within each layer: Most updates per layer appear to be concentrated in the gate G and up-projection $W_{\uparrow}$ of the SwiGLU feed-forward block (Shazeer, 2020), while the down-projection $W_{\downarrow}$ and self-attention parameters (Q, K, V, O) are relatively unaffected. For the self-attention parameters, we hypothesize that this lack of change could be an effect of the relatively high syntactic similarity of English and Danish, requiring less adaptation for in-sequence dependencies. Interestingly, this pattern is also observed when adapting speech recognition models to under-resourced settings (Müller-Eberstein et al., 2024).

The initial embedding layer (EMB) as well as final language modeling head (LMH) also diverge to a comparable degree as G and $W_{\uparrow}$, which is to be expected given their importance to receiving and generating text in a new language. In terms of token-level changes within EMB and LMH (as measured by the absolute difference of each token row before and after adaptation), we observe larger updates to subwords, which occur both in Danish and other Germanic languages (e.g., "_er", "_ik", "_billion"), while subwords in other scripts appear to be least affected. Overall, our findings indicate that future work may be able to train language-specific models more efficiently by focusing exclusively on the EMB, G, $W_{\uparrow}$ and LMH parameters.

5 Final Results and Analysis

Benchmark Results. Using our final model configurations, we present our results on the test split of the Danish portion of ScandEval in Table 3. We compare SNAKMODEL-7B$_{\text{instruct}}$ against variants built on the same base model, including the original LLAMA2-7B$_{\text{base}}$ and LLAMA2-7B$_{\text{chat}}$. In addition, we train +INST$_{\text{da}}$ variants of these English

TASK → ↓ MODEL	LA (mF_1)	NER (μF_1)	SENTI (mF_1)	SUMM (BERTS.)	CSR (Acc.)	QA (F_1)	TM (Acc.)	CT (Acc.)	AVG.
LLAMA2-7B BASED LLMS									
LLAMA2-7B$_{\text{base}}$	33.43	22.31	61.54	65.50	29.76	63.54	38.69	57.05	46.48
LLAMA2-7B$_{\text{chat}}$	47.42	24.63	62.35	66.15	**32.24**	61.34	46.67	55.18	49.50
LLAMA2-7B$_{\text{base}}$ + INST$_{\text{da}}$	36.10	28.48	62.86	66.43	29.04	64.40	49.10	58.46	49.35
LLAMA2-7B$_{\text{chat}}$ + INST$_{\text{da}}$	43.40	29.70	65.92	65.81	30.95	62.46	57.26	55.59	51.39
VIKING-7B	33.67	17.18	49.48	61.96	25.11	56.29	23.97	34.90	37.82
SNAKMODEL-7B$_{\text{base}}$	**56.28**	19.91	57.42	58.95	30.47	18.52	69.14	60.93	46.45
SNAKMODEL-7B$_{\text{instruct}}$	52.91	**29.76**	**66.70**	**66.61**	29.46	**64.66**	**71.05**	**71.88**	**56.63**$^{\uparrow 10.15}$
MISTRAL-7B BASED LLMS									
MISTRAL-7B-V0.1	38.38	**32.66**	**54.53**	66.47	37.39	**64.55**	64.50	71.56	53.76
MUNIN-7B-ALPHA	53.03	28.71	43.77	67.27	**42.68**	63.44	83.01	77.91	57.48
MUNIN-7B-V0.1DEV0	**57.02**	28.74	50.72	**67.89**	42.17	64.41	**93.45**	**85.82**	**61.28**$^{\uparrow 7.52}$

Table 3: **Results (Test) on the ScandEval Benchmark.** We evaluate LLAMA2-7B$_{\text{base}}$, as well as the chat version against SNAKMODEL-7B$_{\text{instruct}}$ and other 7B models in ScandEval (best results in blue). In the subsequent rows, we test the same LLAMA2-7B tuned the Danish instruction tuning data (+ INST$_{\text{da}}$). In the final rows, we show the Mistral-based models (best results in orange). We evaluate in F_1, macro-averaged F_1 (mF_1), micro-averaged F_1 (μF_1), BERTScore (BERTS.; Zhang et al., 2020), and Accuracy (Acc.).

LLAMA2-7B models on the same Danish instruction datasets as SNAKMODEL-7B$_{\text{instruct}}$, in order to isolate the effect of Danish language modeling pre-training. Finally, we include comparisons to the Viking-7B model (SiloAI, 2024) and similarly-sized models based on the Mistral model suite (Jiang et al., 2023; Danish-Foundation-Models-Team, 2024).

Overall, SNAKMODEL-7B$_{\text{instruct}}$ outperforms all other LLAMA2-7B-based models, including those with access to the same set of Danish instruction-tuning data, with a final average benchmark score of 56.63. The performance improvements over the English model are particularly pronounced for sub-tasks based on natural Danish data, including LA (33.43 → 52.91), TM (38.69 → 71.05), and CT (57.05 → 71.88). While the Mistral-7B-based models outperform SNAKMODEL-7B$_{\text{instruct}}$ by up to 4.65% abs., this approximately matches the base model performance difference between Mistral-7B-v0.1 and LLAMA2-7B$_{\text{base}}$ which spans 7.28%.

Qualitative Behaviors. Since ScandEval scores are largely computed using constrained generation, we would like to highlight some qualitative observations from when models generate text without constraint. First, we find that LLAMA2-7B models fail to generate Danish text consistently, even when explicitly prompted to do so (confirming the findings by Puccetti et al., 2024). Since they nonetheless achieve non-trivial benchmark scores under constrained generation, we hypothesize, that they obtain some Danish language functionality during their original, primarily English pre-training. Our custom LLAMA2-7B models to which we add Danish instruction tuning (+INST$_{\text{da}}$) generate Danish responses (even when prompted in English), highlighting that a relatively small amount of translated Danish instructions is sufficient to bias models towards generating output in a new language. Nonetheless, the fact that SNAKMODEL-7B$_{\text{instruct}}$, which is trained on non-translated Danish text outperforms the models trained on translated data, highlights the importance of curating high-quality native-language data for the adaptation target.

6 Guidance for Future Work

From our final evaluation, as well as our analysis of the training dynamics of SNAKMODEL-7B$_{\text{instruct}}$, we next consolidate some guidance for future work adapting English LLMs to languages with similar linguistic properties and resource constraints.

Data. As we found large overlaps across data sources, as well as large amounts of non-Danish or irrelevant data (Section 3), applying stringent pre-processing standards is important when working with smaller languages—especially when automatic filtering tools may be biased towards larger, related languages (e.g., Swedish).

Training. Our training dynamics analysis (Section 4.3) showed that despite our total 13.6B word pre-training corpus, applying instruction tuning after 2,000–5,000 steps of Danish pre-training (i.e., less than half of the corpus) may already be sufficient to obtain close-to-final performance. For instruction tuning itself, one epoch over translated data appears to be sufficient to amplify instruction-following functionalities in the target language. Nonetheless, training on non-translated target language data is important to improve performance on more culturally specific tasks based on native data (i.e., LA, TM, and CT).

Finally, our weight divergence analysis revealed that most parameter updates are consolidated in the embeddings, feed-forward up-projections, and language modeling head. As English and Danish share a relatively similar syntactic structure, languages with more distinctive typologies may nonetheless exhibit larger changes to the self-attention parameters. For model adaption across a comparable typological distance as English and Danish however, focusing training efforts on the aforementioned parameter types—in addition to employing existing parameter-efficient fine-tuning techniques (e.g., Hu et al., 2022; Dettmers et al., 2023)—may therefore yield even higher efficiency gains.

7 Conclusion

In this work, we introduced the SNAKMODEL suite, which includes a 7B-parameter base and instruction-tuned LLM for Danish, in addition to its pre-training and instruction-tuning data, intermediate checkpoints, and evaluation. By analyzing design decisions related to data curation and training dynamics, we further consolidated guidelines for future work adapting LLMs to new languages, to foster research not just in Danish, but in language communities with similar resource constraints.

Limitations

What Went Wrong and What Decisions We Took. Our training process encountered several challenges across multiple runs. In Run 1, we began by restarting training from the LLAMA2-7B checkpoint using the identical learning rate the original model had been trained on. However, we faced gradient explosion at iteration 2,031, which we attempted to mitigate through gradient clipping. Despite this effort, server crashes at step 3,500 and persistent gradient explosions forced us to halt the run after approximately 46 days, with a final language model loss of ±1.77. For Run 2, we halved the peak learning rate to 1.5×10^{-4} and adjusted other parameters, but gradient explosion recurred at step 1,390, leading us to terminate the run after about 10 days with a final loss of ±1.79. In Run 3, we significantly reduced the peak learning rate to 1.5×10^{-5}, reasoning that as we were continuing pre-training, we should aim for a rate lower than Llama2's final learning rate. This approach has shown effective, with the training reaching iteration 12,500 after approximately 93 days and achieving a language model loss of ±1.72.

Acknowledgments

First, we would like to thank Barbara Plank for allowing us to use her compute hardware for this period of time. Additionally, this work was impossible without the stable High Performance Compute cluster at the IT University of Copenhagen, being able to train a model for ±90 days without a single interruption is extraordinary. Second, we thank Ahmet Üstun for giving us invaluable and concrete comments on hyperparameter setup for continuous pre-training. Last, we would also like to thank the reviewers for their valuable comments. Elisa Bassignana is supported by a research grant (VIL59826) from VILLUM FONDEN. Mike Zhang is supported by a research grant (VIL57392) from VILLUM FONDEN.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219.

AI-Sweden. 2024. Ai-sweden-models/llama-3-8b.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *ArXiv preprint*, abs/2402.17733.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *ArXiv preprint*, abs/2405.15032.

Giuseppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri. 2024. DanteLLM: Let's push Italian LLM research forward! In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4343–4355, Torino, Italia. ELRA and ICCL.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *ArXiv preprint*, abs/2309.16609.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Xingyao Wang, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. 2023. epfllm megatron-llm.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Morten H. Christiansen, Kristian Tylén, Riccardo Fusaroli, Dorthe Bleses, Anders Højen, Fabio Trecca, Christina Dideriksen, and Byurakn Ishkhanyan. 2023. The puzzle of danish.

Manuel R. Ciosici and Leon Derczynski. 2022. Training a t5 using lab-sized resources.

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.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv preprint*, abs/2207.04672.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *ArXiv preprint*, abs/2412.04261.

Danish-Foundation-Models-Team. 2024. Releasing munin 7b alpha - a danish llm.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Christina Dideriksen, Morten H. Christiansen, Mark Dingemanse, Malte Højmark-Bertelsen, Christer Johansson, Kristian Tylén, and Riccardo Fusaroli. 2023. Language-specific constraints on conversation: Evidence from danish and norwegian. *Cognitive Science*, 47(11). Publisher Copyright: © 2023 Cognitive Science Society LLC.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *ArXiv preprint*, abs/2404.03608.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages.

In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Kenneth Enevoldsen, Lasse Hansen, and Kristoffer Nielbo. 2021. Dacy: A unified framework for danish nlp. *ArXiv preprint*, abs/2107.05295.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. 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.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Oliver Kinch. 2023. Nordjylland news summarization. Accessed on 22-08-2024.

Andreas Kirkedal, Marija Stepanovic, and Barbara Plank. 2020. FT speech: Danish parliament speech corpus. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 442–446. ISCA.

Andrew V Knyazev and Merico E Argentati. 2002. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040.

Matthias Trautner Kromann and Stine Kern Lynge. 2004. The danish dependency treebank v. 1.0.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *ArXiv preprint*, abs/1910.09700.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

LeoLM-Team. 2024. Leolm: Igniting german-language llm research.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

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.

Magnus Mabeck. 2024. Danish openhermes. `https://huggingface.co/datasets/Mabeck/danish-OpenHermes`.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

Max Müller-Eberstein, Dianna Yee, Karren Yang, Gautam Varma Mantena, and Colin Lea. 2024. Hypernetworks for Personalizing ASR to Atypical Speech. *Transactions of the Association for Computational Linguistics*, 12:1182–1196.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

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. *ArXiv preprint*, abs/2309.09400.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Dan Saattrup Nielsen. 2024. Danish citizen test. Accessed on 22-08-2024.

NORA.LLM-Team. 2024. Instruction-tuned normistral-7b-warm. Accessed on 17-10-2024.

Amalie Brogaard Pauli, Maria Barrett, Ophélie Lacroix, and Rasmus Hvingelby. 2021. DaNLP: An open-source toolkit for Danish natural language processing. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460–466, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Københavns Professionshøjskole. 2024. Skolegpt instruct. `https://huggingface.co/datasets/kobprof/skolegpt-instruct`.

Giovanni Puccetti, Anna Rogers, Chiara Alzetta, Felice Dell'Orletta, and Andrea Esuli. 2024. AI 'news' content farms are easy to make and hard to detect: A case study in Italian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15312–15338, Bangkok, Thailand. Association for Computational Linguistics.

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.*, 21:140:1–140:67.

Rakuten Group, Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johanes Effendi, et al. 2024. Rakutenai-7b: Extending large language models for japanese. *ArXiv preprint*, abs/2403.15484.

Edwin Rijgersberg and Bob Lucassen. 2023. Geitje: een groot open nederlands taalmodel.

Oscar Sainz, Jon Ander Campos, García-Ferrero Iker, Julen Etxaniz, and Eneko Agirre. 2023. Did ChatGPT cheat on your test?

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer. 2020. Glu variants improve transformer. *ArXiv preprint*, abs/2002.05202.

SiloAI. 2024. Viking 7b/13b/33b: Sailing the nordic seas of multilinguality.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ǎguila Team. 2023. Introducing aguila, a new open-source llm for spanish and catalan.

"Teknium". 2023. OpenHermes dataset. `https://huggingface.co/datasets/teknium/openhermes`.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Fabio Trecca, Kristian Tylén, Anders Højen, and Morten H. Christiansen. 2021. Danish as a window onto language processing and learning. *Language Learning*, 71(3):799–833.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Bram Vanroy. 2024. Fietje 2: An open and efficient llm for dutch.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.