Teuken-7B-Base & Teuken-7B-Instruct: Towards European LLMs

Anonymous ACL submission

Abstract

We present two multilingual Large Language Models (LLMs) designed to embrace Europe's linguistic diversity by supporting all 24 official languages of the European Union. Trained on a dataset comprising around 60% non-English data and utilizing a custom multilingual tokenizer, our models address the limitations of existing LLMs that predominantly focus on English or a few high-resource languages. We detail the models' development principles, i.e., data composition, tokenizer optimization, and training methodologies. The models demonstrate competitive performance across multilingual benchmarks, as evidenced by their performance on European versions of ARC, HellaSwag, MMLU, and TruthfulQA.

1 Introduction

LLMs represents a disruptive technology that has the potential to be applied in numerous applications. To develop LLMs, expertise in various areas is required, starting from large-scale data preprocessing (Penedo et al., 2024), training efficient tokenizers (Petrov et al., 2023; Ali et al., 2024), pre-training the models efficiently on a vast infrastructure spanning thousands of GPUs, instruction tuning the pre-trained models, and properly evaluating them on various downstream tasks (Dubey et al., 2024). Multilingualism as an additional dimension introduces additional considerations to all phases of the model development, such as multilingual data composition and multilingual tokenizer. Therefore, it is crucial that the technology and expertise to build these models is democratized to enable different communities and organizations to employ these models for their use cases.

Many efforts in developing open-source LLMs have been undertaken, such as BLOOM (Scao et al., 2022), LLaMA-3 (Dubey et al., 2024), OLMo (Groeneveld et al., 2024), Aya (Üstün et al., 2024), and Mistral (Jiang et al., 2023).

While these existing efforts represent significant contributions to the community and the advancement of artificial intelligence, there are still two major limitations. First, the current open-source models are predominantly English-centric, limiting their use to a broad multilingual context covering high and low resource languages such as within the European Union. Relying on English-centric models that employ an English-centric tokenizer in a multilingual context introduces severe disadvantages such as lower downstream performance, additional inference costs, and increased training costs for language-specific continued pre-training and fine-tuning for languages besides English (Petrov et al., 2023; Ali et al., 2024). Second, the open-source efforts disclose different levels of granularity regarding sharing details about the development of the models. Even though information regarding the model architecture is usually shared, the decisions/ablation experiments behind certain architectural choices are not always described, which can provide crucial insights relevant to developing custom LLMs. This limitation is even more apparent in the description of the dataset composition, which often provides a coarse overview, hampering the reproduction of the work.

To address the aforementioned limitations, we created a multilingual base model that has been trained on top of all 24 European official languages and the corresponding instruction-tuned model. In particular, we make the following contributions:

- We present Teuken-7B-Base, a European pre-trained LLM that has been trained from scratch for 4 trillion tokens based on all 24 official European languages.
- We present Teuken-7B-Instruct, the is instruction tuned model on top of the base model

Besides these artefacts, we describe our design decisions in detail to facilitate the reproduction of

our work and the development of novel models based on our work. Therefore, our contributions present an essential step towards the democratization of technology within Europe.

2 Related Work

Since the introduction of GPT-3 (Brown et al., 2020), several open-source/open-weights efforts have been undertaken to train LLMs. While the large majority of the work focus on English-centric models (Zhang et al., 2022; Touvron et al., 2023a,b; Groeneveld et al., 2024; Dubey et al., 2024), there have been also efforts training multilingual models.

One of the most prominent examples is BLOOM (Scao et al., 2022), a 176B LLM trained on 46 natural languages. Further examples that specifically address multilingualism are the encoder-decoder models mT5 (Xue et al., 2021), XLM (Lample and Conneau, 2019), XLM-R (Conneau et al., 2020), and the encoder model mBERT (Devlin et al., 2019).

Unlike the previously mentioned efforts, we specifically address 24 official European languages and focus on ensuring that a large fraction of the training data is composed of non-English data, representing a major step towards European LLMs. Concurrent to our work, EuroLLM (Martins et al., 2024), a 1.7B decoder-only LLM that follows the same spirit as our undertaking by addressing all 24 European languages and Salamandra[1] covering 32 languages, have been presented.

3 Pre-Training Data

We used the dataset presented in (Brandizzi et al., 2024) as a base since it i.) covers all 24 official EU languages, ii.) is large, iii.) contains a large amount of non-English data, iv.) comprises various domains, and v.) is filtered based on established practices. In the following, we describe the composition of our final training comprised of web-crawled and curated data.

For the the web crawled part, we sampled our dataset based on the 60 filtered Common Crawls WET dumps (cf. Appendix A.1) provided by Brandizzi *et al.* (Brandizzi et al., 2024). The raw WET dumps have been filtered based on established heuristics, ensuring that the pre-training corpus is of high quality (Brandizzi et al., 2024). Additionally, we employed the recently released Fine-Web EDU dataset (Penedo et al., 2024) as significant performance improvements for models trained on the dataset have been reported. It should be noted that Fine-Web EDU is only available in English. We still integrated it and investigated the cross-lingual performance of the corresponding model checkpoint (see Section 5.2). To ensure a multilingual composition, we up-sampled all languages except English, which we down-sampled. Specifically, we employed the content for all languages except English Overall, the web crawled part makes up 86.55% (cf. Table 42) our pre-training dataset.

In addition to web crawled data, we included all curated datasets from (Brandizzi et al., 2024). Part of the curated datasets are domain-specific data such as academics, finance, and patents ensuring the diversity of our training dataset.

Our composed training dataset contains 4 trillion tokens, of which 86.55% (cf. Table 42) originates from web data, and the remaining 13.45% represent is curated data As illustrated in Figure 1 and Figure 2, 41.70% of the tokens stem from English content and due to the inclusion of German, French, and Spanish, we approach around two-thirds of the total tokens.

4 Multilingual Tokenization and Fertility Impact on Model Efficiency

In multilingual natural language processing (NLP), it is crucial to train balanced multilingual tokenizers (Petrov et al., 2023; Ali et al., 2024) to avoid increased training and inference costs and latency during inference for non-English queries. Furthermore, it prevents the model from learning long-range dependencies in limited context windows (Vaswani et al., 2017).

Therefore, we developed a custom multilingual tokenizer, closely following Ali et al. (2024), that is optimized for all 24 official European languages. It aims to reduce excessive text fragmentation, a phenomenon termed high "fertility", and refers to the average number of tokens generated per word.

Fertility (F) is defined as the ratio of the total number of tokens (T) to the total number of words (W) in a text, as shown in the following:

$$F = \frac{T}{W}$$

We conducted a fertility analysis on 2,000 sentences from the FLORES-200 (Costa-jussà et al., 2022) dataset to compare tokenizers. Because the dataset is translated across languages, i.e., the analysis is conducted on semantic equivalent content, it

[1] https://huggingface.co/BSC-LT/salamandra-7b

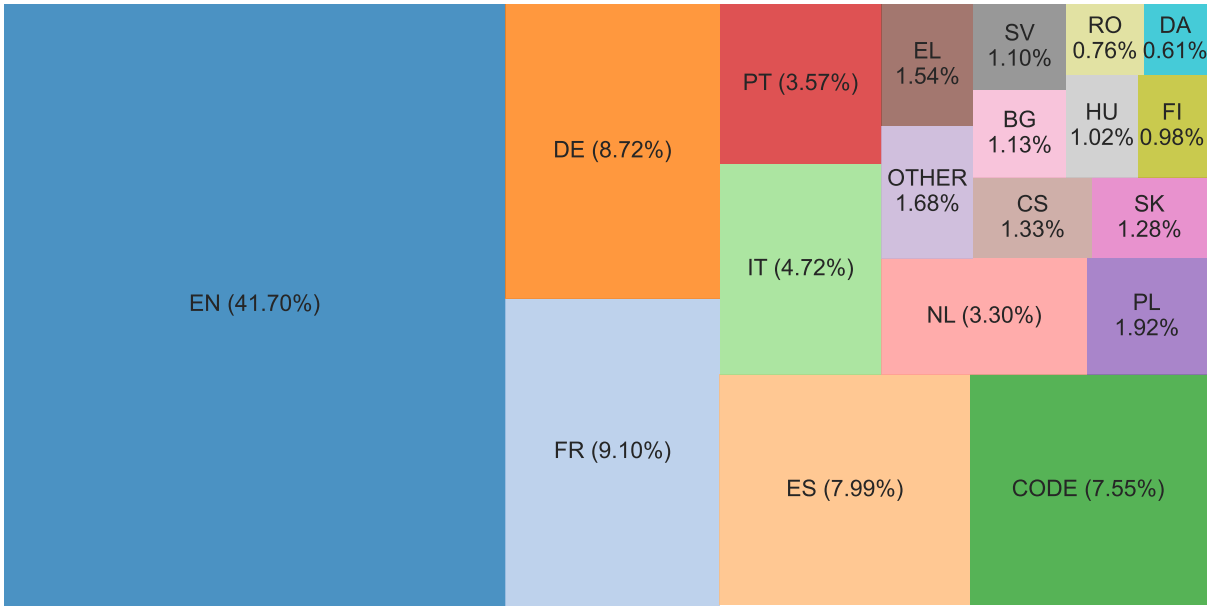


Figure 1: Language distribution of the tokenized dataset, comparing the presence of English and other European languages.

Figure 2: Breakdown of the "OTHER" category.

provides a reliable basis for evaluation. A comparison with other widely-used tokenizers is presented in Figure 3

Our custom tokenizer demonstrates that for 19 out of the 24 languages, fertility values are similar or lower than those of related tokenizers. This effect is especially pronounced in languages with complex morphology or long word structures, such as Finnish, German, and Hungarian.

Lowering fertility enables longer queries and documents to be processed without exceeding the context window. This is particularly advantageous in tasks that require the processing of legal or medical documents, where maintaining the integrity of long documents is essential for accurate understanding.

5 Base Model

In the following, we describe the model architecture (Section 5.1) and training (Section 5.2). Additionally, we describe in the appendix the used training framework (Appendix A.7) and the training infrastructure (Appendix A.8).

5.1 Model Architecture

Our model is a 7B transformer-based decoder-only model. Table 7 provides an overview of our model architecture. We want to highlight that our architectural choices are derived from internal ablation studies and findings from related work. Our models have a sequence length of 4096 tokens and employ Rotary (Su et al., 2024) positional embeddings that are employed to train state-of-the-art models (Dubey et al., 2024). To accelerate inference and reduce memory requirements, we employed grouped-query attention (Ainslie et al., 2023). An entire overview of our architectural choices is presented in Table 7.

These and other design decisions were guided by medium-scale (Chinchilla-optimal (Hoffmann et al., 2022) training of a 2.6B parameter model) ablation runs for various training-related hyperparameters. Our goal with these ablations was to find improvements in the compute-equivalent setting while also confirming whether proposed modifications transfer to our codebase, which is not necessarily the case (Narang et al., 2021). Appendix A.2 describes the conducted ablation experiments in detail, Table 6 summarizes the results, and Figures 6 and 7 contain loss curves for the various

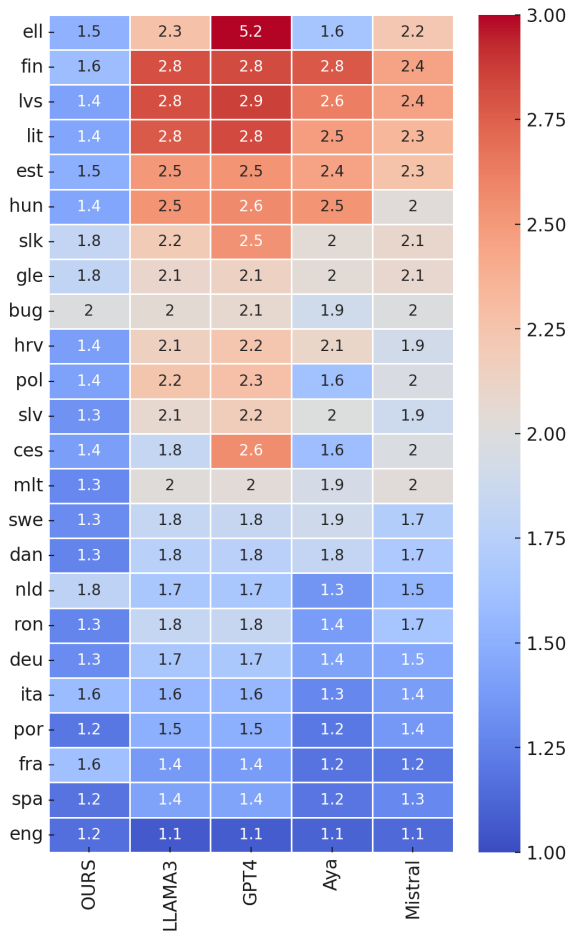


Figure 3: Fertility across the official 24 European languages.

experiments.

5.2 Training

Using the causal language modelling training objective, we trained our model on 4T tokens covering all 24 European languages as described in Section 3, employed AdamW as an optimizer and used a cosine learning rate schedule starting with a learning rate of 3e-5, increasing it to the maximum learning rate of 3e-4 within the first 10,000 steps, and decaying it afterwards. During the training, we took two additional design decisions motivated by recent findings from related work.

First, Penedo *et al.* (Penedo et al., 2024) showed that training based on educational content further improves the performance of LLMs. Therefore, after 2.85T tokens, we conducted an ablation where we trained one model up to 3T tokens based on FineWeb-EDU (Penedo et al., 2024) (which is only available in English), and a second model with the initial data composition (see Section 3). As shown in Figure 4, the model trained based on education content obtained a performance % across all tasks and languages except for HellaSwag, where the results were comparable. In the appendix, we show the performance development for English (Figure 8), French (Figure 9), German (Figure 10), Finish (Figure 11) and Estonian (Figure 12).

Second, we want to emphasize that the initial goal has been to train our base model up to 3T tokens, which are beyond compute-optimal based on the scaling laws presented by Hoffmann *et al.* (Hoffmann et al., 2022) for our model size. Because recent works show that it is beneficial to further the pre-train model, e.g., LLaMA 3 has been trained up to 15T tokens, we decided to train the model on an additional 1T token. Consequently, we faced the challenge of defining the corresponding learning rate schedule that had been originally defined for training up to 3T tokens. We ablated different settings to continue pre-training inspired by (Ibrahim et al., 2024) as described in Appendix A.4. As a result, we further decreased the learning rate from 10% of the maximum learning rate to 8.3% within the following 1T tokens.

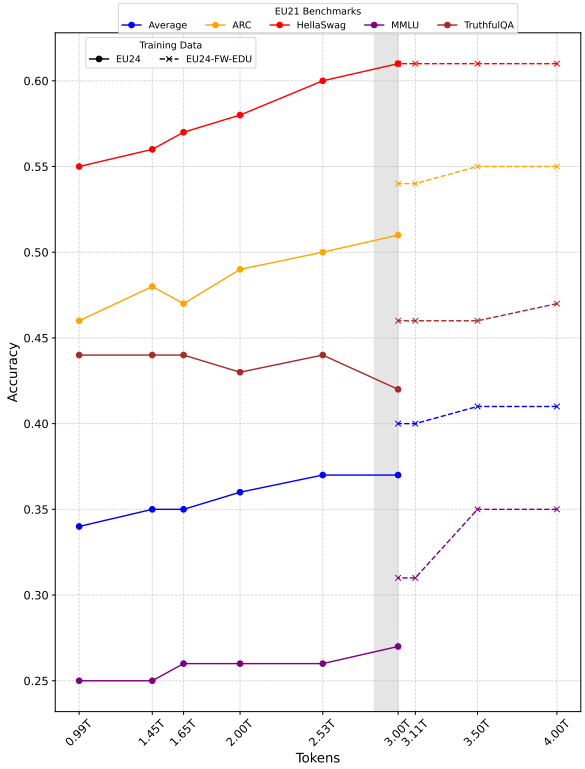


Figure 4: Downstream performance of the base model from 0.99T to 4T tokens across 21 European languages. The grey area highlights the ablation comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens.

6 Instruction Tuned Model

In this section, we describe the creation of the instruction tuned model, including the dataset composition (Section 6.1) and the instruction tuning (Section 6.2).

Additionally to "Ours (Instruct)", we ablated another model variant, which contains examples for all 24EU Languages, as shown in Table 41. To our surprise and as visible in Figure 14, both variants are comparable in performance on MT-Bench-X. However, "Ours (Instruct)" performs slightly better on the EU21 benchmark for selected languages. Hence, we present the variant "Ours (Instruct)", which focuses on English and German for instruction-tuning.

6.1 Data

In the work of (Weber et al., 2024), the authors show that employing (parallel) multilingual instruction tuning datasets improves cross-lingual instruction following the capabilities of the model. Therefore, we composed a multilingual dataset of English and German data. The dataset comprises several subsets, as listed in Table 40. Our dataset is not entirely parallel, as suggested by (Weber et al., 2024) due to the lack of parallel data.

English Our objective was to balance the dataset, aiming to include a similar number of English and German examples. As shown with the subset selection in Table 40, we included only high-quality English data. To reduce the dataset size, the following multi-step filtering process was executed. First, we computed reward scores for all English examples using Starling-RM-7B-alpha[2]. Second, to increase the models' coding ability and multi-turn chat capability, we include all multi-turn examples, the complete code alpaca‘ subset, and the portion of the LMSYS-Chat dataset subset characterized as high-quality by Zheng et al. (2024). The remaining high-quality subsets — Open Orca (Mukherjee et al., 2023) and both versions of the Evol instruct dataset (Xu et al., 2024) — are then sampled to balance the German and English dataset proportion. For this, we select the highest-scoring examples to ensure that each subset contributed an equal amount of high-quality data.

German We incorporated several datasets for the German data selection to ensure broad linguistic coverage. We included the German and English subsets of the Bactrian-X dataset, as Weber et al. (2024) demonstrated that semantically parallel data enhances cross-lingual model performance. In the same spirit, we include a proportion of the Alpaca, Evol-Instruct, and ShareGPT datasets for which either the instructions or also the responses are translated to German by Chen et al. (2023).

6.2 Post-Training

For instruction-tuning the model, we extended FastChat[3] for multilingual system prompts and translated the system prompt "*A chat between a human and an artificial intelligence assistant. The assistant gives helpful and polite answers to the human's questions.*" for the five evaluation languages. We employed standard instruction tuning by calculating the loss only for the assistant responses of a multi-turn conversation. We trained the model on 8xH100 GPUs for 2.5 days for three epochs.

7 Results

In the following, we present the results of our base and instruction-tuned models, with a particular focus on multilingual evaluation. Our evaluation targets the official European languages, as this aligns with the core objective of our research.

Section 7.1 describes our evaluation set-up, Section 7.2 presents our results across all 21 investigated languages, while Section 7.3 focuses on the performance of our models in the six widely spoken languages: English, German, French, Italian, Spanish, and Portuguese, and Section 7.4 presents our findings on the remaining 15 languages. Additionally, Section 7.5 highlights the performance of various multilingual models on the largest common language subset. Finally, Section 7.6 presents the instruction following capabilities of Teuken-7B-Instruct on the recently published multilingual benchmark MT-Bench-X (Weber et al., 2024).

7.1 Evaluation Set-Up

Evaluation Datasets The evaluation was conducted using ARC (Clark et al., 2018)/EU21-ARC (Thellmann et al., 2024) (25-shot for science-based questions), HellaSwag (Zellers et al., 2019)/EU21-HeSw (Thellmann et al., 2024) (10-shot for commonsense reasoning), TruthfulIQA (Lin et al., 2022)/EU21-TQA (Thellmann

[2] https://huggingface.co/berkeley-nest/Starling-RM-7B-alpha

[3] https://github.com/lm-sys/FastChat/tree/main/fastchat

et al., 2024) (6-shot pseudo-shot for generating truthful answers), and MMLU (Hendrycks et al., 2021)/EU21-MMLU (Thellmann et al., 2024) (5-shot for broad knowledge), which are available in 21 of the 24 official European languages, for which we report mean accuracy.

Models We evaluated our models against related multilingual models with up to 8B parameters that have been pre-trained on causal language modelling from scratch and models that result from an instruction/fine-tuning of the pre-trained model. We did not include models that represent the distillation of larger models to ensure a comparable setting. As a result, we evaluated against Aya-23 8B, Bloom-7B1, Bloomz-7B1, Meta-Llama-3.1-8B, Meta-Llama-3.1-8B Chat, Salamandra-7B, Salamandra-7B Instruct, Pharia-1-LLM-7B, Pharia-1-LLM-7B-C-A (Control-Aligned), Occiglot-7B-EU5, and Mistral. Table 1 provides an overview of the number of languages and tokens that the models have been trained on. Note that the Mistral models were not trained as multilingual models and are intended to be used as English-only models. However, they contain multilingual to some extent, as described in (Jiang et al., 2024). Due to its strong performance, we included Mistral in our evaluations.

Model	Languages	Tokens
Aya-23 8B	23	N/A
Bloom-7B1	46	350B
Bloomz-7B1	46	354B
Meta-Llama-3.1-8B	8	15T
Mistral-7B-v0.3	1	N/A
Salamandra-7B	35	7.8T
Pharia-1-LLM-7B	7	4.7T
Ours (Base)	24	4T

Table 1: Evaluated models and the number of languages and training tokens the models have been trained on.

7.2 Performance on 21 European Languages

Table 2 and Table 9 present the multilingual results of the instruction-tuned and based models for all 21 European languages. Table 2 illustrates that Meta-Llama-3.1-8B-Instruct leads with an average accuracy of 56.3% across all benchmarks, followed by our instruction-tuned model and the Salamandra model. The results are encouraging, considering that our model has been trained on significantly fewer tokens than both other models (see Tables 1). We hypothesize that the training on the FineWeb-EDU dataset has contributed significantly to this aspect since it has shown to provide similar results compared to different composed datasets while requiring significantly fewer tokens (Penedo et al., 2024). Noteworthy, it improves cross-lingual performance even though the dataset is in English (see Figure 4). On benchmark-level, Meta-Llama-3.1-8B-Instruct excels in EU21-MMLU (57.6%) while Salamandra-7B-Instruct is particularly strong in EU21-ARC (58.9%) and EU21-HeSw (63.7%) and Mistral-7B-Instruct-v0.3 obtains the best result on EU21-TQA (54.8%). Again, our model is competitive with the best-performing model in all tasks except for EU21-MMLU despite being trained on significantly fewer tokens.

Besides the average performance across languages, we also investigated the robustness across languages. Figure 5 shows that our model and Salamandra-7B-Instruct and Bloomz-7B1 are significantly more robust than the other models on all benchmarks except EU21-TQA where several models obtain robust performance across languages, ensuring that the model performs comparably across languages.

7.3 Performance on Top-6 European Languages

Table 3 focuses on the performance of instruction tuned models on six widely spoken European languages: English, German, French, Italian, Spanish, and Portuguese. These languages are well-represented in training data and are commonly used in evaluating multilingual models. Meta-Llama-3.1-8B-Instruct leads with 62.3% average performance across tasks, excelling in EU21-HeSw (67.7%) and EU21-ARC (64.8%), followed closely by Mistral-7B-Instruct-v0.3 with 61.3%, obtaining the best performance in EU21-TQA (56.8%) and EU21-ARC (65.4%). Occiglot-7B-eu5-Instruct which is a fine-tuned Mistral model outperforms the all models on EU21-HeSw followed by Pharia-1-LLM-7B-C-A. Ours (Instruct) performs competitively in EU21-HeSw 67.5% and EU21-ARC (63.2%), but shows limitations in EU21-MMLU (45.7%), particularly in domain-specific tasks.

7.4 Performance on Exclusive European Languages

Table 4 compares the models across 15 less commonly evaluated European languages (Romanian,

Model	Avg.	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B-Instruct	**.563**	.563	.579	.532	**.576**
Mistral-7B-Instruct-v0.3	.527	.530	.538	**.548**	.491
Salamandra-7B-Instruct	.543	**.595**	**.637**	.482	.459
Aya-23-8B	.485	.475	.535	.476	.455
Occiglot-7B-eu5-Instruct	.475	.484	.519	.471	.428
Pharia-1-LLM-7B-C-A	.417	.396	.438	.469	.366
Bloomz-7B1	.358	.316	.354	.461	.302
Ours (Base)	.496	.550	.615	.469	.349
Ours (Instruct)	.543	.581	.624	.543	.425

Table 2: Results on multilingual benchmarks for 21 European languages with instruction-tuned models.

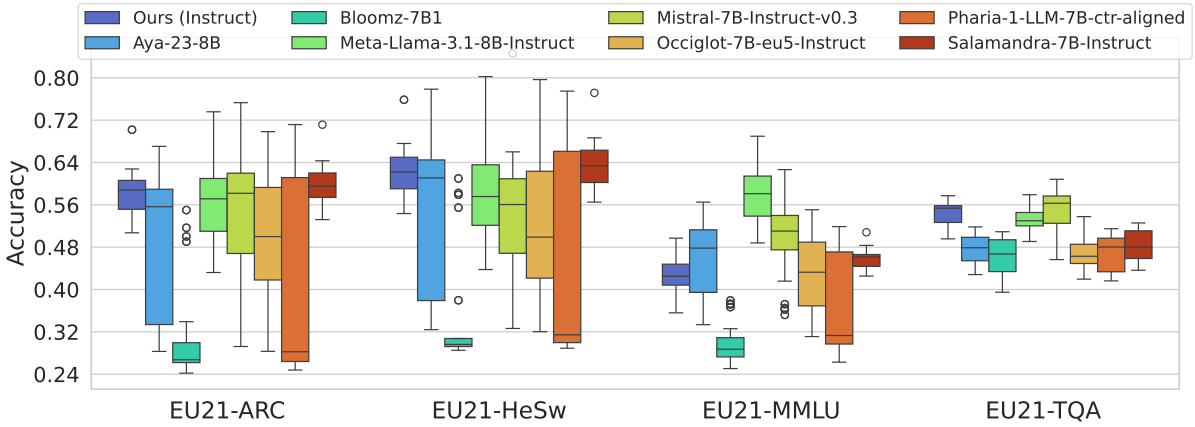


Figure 5: Models' performance across 21 languages for different benchmarks.

Czech, Danish, Greek, Estonian, Finnish, Hungarian, Lithuanian, Latvian, Dutch, Bulgarian, Polish, Slovak, Slovenian, and Swedish). Meta-Llama-3.1-8B-Instruct leads with an average accuracy of 53.8%, excelling in EU21-MMLU (55.4%) closely followed by Salamandra-7B-Instruct (53.5%) obtains the strongest results on EU21-ARC (57.6%) and EU21-HeSw (61.9%). Ours (Instruct) is competitive on average with 53.1% and benchmark-level. The model is among the two best-performing models except for EU21-MMLU and outperforms all other models on EU21-TQA (54.7%).

7.5 Comparison on Common Languages

Table 8 compares the performance of "Ours (Instruct)" models with Aya-23-8B, Bloomz-7B1, and Salamandra-7B-Instruct across ten common European languages: Czech, Dutch, English, French, Greek, Italian, Polish, Portuguese, Romanian, and Spanish. Ours (Instruct) achieves 56.0%, closely matching Aya-23-8B's top score of 56.3%, and leads in EU21-TQA (54.1%), outperforming Aya (48.4%) and Salamandra (47.6%), while remaining competitive in EU21-ARC (61.0%) and EU21-HeSw (65.3%).

7.6 Multilingual Instruction Following Evaluation

For evaluating the multilingual instruction-following capabilities of Teuken-7B-Instruct, we utilized MT-Bench-X that employs an LLM as a judge (Weber et al., 2024) across all five available evaluation languages: English, German, French, Italian, and Spanish. The results are presented in the Appendix A.6.2. We compare the cross-lingual performance of our model with Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 and Salamandra-7B-Instruct in the following. While the creative writing skills (Writing, Roleplay) and question-answering skills (Stem, Humanities) of our model show good performance and are comparable for German to Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3, the mathematical and coding skills are inferior across languages. This performance gap can be explained by the fact that our model has not been optimized for these capabilities.

Model	Avg.	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B-Instruct	**.623**	.648	.677	.535	**.632**
Mistral-7B-Instruct-v0.3	.613	**.654**	.670	**.568**	.560
Aya-23-8B	.574	.614	.687	.470	.526
Occiglot-7B-eu5-Instruct	.583	.646	**.712**	.458	.518
Pharia-1-LLM-7B-C-A	.580	.643	.696	.497	.485
Salamandra-7B-Instruct	.565	.643	.685	.455	.477
Bloomz-7B1	.433	.449	.502	.426	.354
Ours (Base)	.521	.599	.667	.445	.371
Ours (Instruct)	.574	.632	.675	.532	.457

Table 3: Results on multilingual benchmarks for 6 Languages (English, German, French, Italian, Spanish, Portuguese) for instruction-tuned models.

Model	Avg.	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B-Instruct	**.538**	.529	.540	.530	**.554**
Mistral-7B-Instruct-v0.3	.492	.480	.485	.540	.463
Salamandra-7B-Instruct	.535	**.576**	**.619**	.493	.451
Aya-23-8B	.450	.420	.473	.479	.427
Occiglot-7B-eu5-Instruct	.432	.419	.441	.476	.392
Pharia-1-LLM-7B-C-A	.352	.296	.334	.458	.319
Bloomz-7B1	.328	.263	.295	.475	.281
Ours (Base)	.486	.531	.594	.478	.341
Ours (Instruct)	.531	.560	.603	**.547**	.413

Table 4: Results on multilingual benchmarks across the 15 exclusive European languages (Romanian, Czech, Danish, Greek, Estonian, Finnish, Hungarian, Lithuanian, Latvian, Dutch, Bulgarian, Polish, Slovak, Slovenian, and Swedish for instruction-tuned models.

Compared to *Salamandra-7B-Instruct*, our model demonstrates superior cross-lingual capabilities, as shown in Figure 14a. Only, in EN-Extraction and FR- and ES-Reasoning it outperforms our model.

8 Conclusion & Future Work

In this work, we presented the development of two multilingual large language models, "Ours (Base)" and "Ours (Instruct)", tailored to support the linguistic diversity of Europe by encompassing all 24 official EU languages. Through the use of a custom multilingual tokenizer and a dataset prioritizing non-English content, our models address limitations found in existing multilingual models, particularly their English-centric bias. Our results demonstrate competitive performance across multiple benchmarks, including ARC, HellaSwag, MMLU, and TruthfulQA, while training on significantly fewer tokens. The presented multilingual models, focusing on European languages, are an excellent starting point for further pre-training or fine-tuning each European language. We have taken great care to ensure that all languages are represented in the tokenizer, promoting the inclusivity and applicability of our models in diverse linguistic contexts. We shared insights regarding our model development, often not described in detail, to support future development aimed at developing multilingual models.

Our ongoing efforts aim to improve the models' capabilities further, focusing on math and coding capabilities by further training/fine-tuning our models. We emphasized the competitiveness of our model's results despite training on fewer tokens. Therefore, we will focus on further improving the training datasets' data quality to ensure efficient training, resulting in decreased resource requirements. Finally, we aim to train and share larger LLMs covering an even larger set of European languages beyond the 24 official European languages, making the technology available to an even larger community.

9 Limitations

Developing LLMs and particularly multilingual LLMs is a complex undertaking which involves several trade-offs during the different development phases (data composition, tokenize development, pre-training, instruction-tuning, and evaluation) Despite the essential contribution of our work towards European LLMs, it has several limitations.

First, we did not investigate multilingual scaling laws due to limited resources. This should be conducted in a separate effort, and our results present an excellent reference.

Second, our model has also been trained on educational content (FineWeb-EDU) available in English. Our results revealed that training on educational content in English improves the cross-lingual performance of the model. However, the performance varies across languages. We assume that employing educational content for other (low-resource) languages will strongly benefit these languages.

Third, we decided not to particularly focus on coding and math content, i.e., our model should not be employed in a context where these capabilities are required even though it has these capabilities to an extent.

Fourth, our model has a context window of 4096 tokens, currently limiting the use of the model in settings where long documents should be processed.

Finally, due to the limited resources, our model has been trained on significantly fewer tokens than related models, which might explain why our model performs poorly in specific tasks.

Because we open source our models, the limitations 2-5 can be addressed by further training/instruction tuning the models dependent on the specific needs. Since our model relies on a multilingual tokenizer that ensures low fertility across languages, further training can be conducted efficiently.

10 Ethical & Broader Impact

Considering the enormous potential LLMs to be applied in various use cases, limiting them to a few languages introduces disadvantages such as lower downstream performance, increased continued training/fine-tuning and inference costs for communities not represented by the selected languages. Our work makes a significant step towards democratizing this technology within Europe, which represents a diverse linguistic context. Furthermore, our strong focus on multilingualism encompasses the different cultural aspects reflected in natural language.

Because training LLMs requires vast resources, it is crucial to use them responsibly. By employing a multilingual tokenizer, multilingual documents are not unnecessarily highly fragmented, ensuring efficient training and reduced carbon footprint. Consequently, future work further training our models can be conducted efficiently across the investigated languages.

We acknowledge that a model trained in multiple languages can be misused, such as generating harmful or misleading content in any of the 24 languages. Exploiting this can lead to the spread of disinformation, manipulation, or even malicious actions.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, pages 4895–4901. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Nicolo' Brandizzi, Hammam Abdelwahab, Anirban Bhowmick, Lennard Helmer, Benny Jörg Stein, Pavel Denisov, Qasid Saleem, Michael Fromm, Mehdi Ali, Richard Rutmann, Farzad Naderi, Mohamad Saif Agy, Alexander Schwirjow, Fabian Küch, Luzian Hahn, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Dennis Wegener, Nicolas Flores-Herr, Joachim Köhler, and Johannes Leveling. 2024. Data processing for the opengpt-x model family. *Preprint*, arXiv:2410.08800.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 *NeurIPS*.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

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, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. 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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh,

Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Google. 2020. T5 v1.1.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen

Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *NeurIPS*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pretrain large language models. *Trans. Mach. Learn. Res.*, 2024.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. *arXiv preprint arXiv:2405.09373*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2024. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36.

Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5:341–353.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Preprint*, arXiv:1901.07291.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *ArXiv e-prints*, pages arXiv–1607.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL (1)*, pages 3214–3252. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. 2021. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Guilherme Penedo, Hynek Kydlícek, 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. *CoRR*, abs/2406.17557.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *NeurIPS*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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. *Journal of machine learning research*, 21(140):1–67.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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

Sam Shleifer, Jason Weston, and Myle Ott. 2021. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards cross-lingual llm evaluation for european languages. *Preprint*, arXiv:2410.08928.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin 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 *ACL (1)*, pages 15894–15939. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? *CoRR*, abs/2402.13703.

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. 2024. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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 *NAACL-HLT*, pages 483–498. 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 *ACL (1)*, pages 4791–4800. Association for Computational Linguistics.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher

Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Appendix

A.1 Data

In this section, we outline the used Common Crawl data, detailing the cutoff dates and the data collection period. The dumps span multiple years, with varying distributions of weeks per year (cf. Table 5), which is critical for understanding the temporal coverage of the training data.

Year	Week
2014	42
2015	14, 48
2016	22, 44
2017	13, 47, 51
2018	5, 9, 13, 17, 22, 26, 30, 34, 39, 43, 47, 51
2019	4, 9, 13, 18, 22, 26, 30, 35, 39, 47, 51
2020	5, 10, 16, 24, 29, 34, 40, 45, 50
2021	4, 10, 17, 21, 25, 31, 39, 43, 49
2022	5, 21, 27, 33, 40, 49
2023	6, 14, 23, 40, 50

Table 5: List of dumps by year and week.

The data spans from 2014 to 2023, with earlier years (2014-2016) containing fewer dumps but often representing a larger dataset per dump. The latest available data cutoff is from 2023 (week 50), giving the model a comprehensive span of nearly a decade of web data.

This periodization ensures that the model has exposure to a wide temporal range of web content, with a balanced emphasis on both earlier, higher-density dumps and more recent, frequent dumps, providing a robust training corpus.

A.2 Training Ablation Experiments

We present here the training ablations and our methodology in more detail. Table 6 summarizes the main results, while Figures 6 and 7 display training loss curves of the various ablations against the baseline ("original" in the plots). The loss curves are throughput-normalized where appropriate.

We conducted the experiments on 2.6B scale, and focused on 3 downstream-tasks: (1) ARC Easy (Clark et al., 2018), (2) HellaSwag (Zellers et al., 2019), and (3) LAMBADA (Paperno et al., 2016). Because the models often achieved better-than-random performance in these evaluation tasks and because they test different abilities of the model, they were deemed a good-enough proxy measure of general downstream improvements.

Consistent throughput across runs is critical in comparing in the compute-equivalent setting. While guaranteeing consistent throughput across runs was not possible due to external factors (such as nondeterministic job layout on the supercomputing cluster, or varied I/O throughput due to a shared parallel file system), we believe the variance to be small enough (up to a percent) not to significantly affect our findings. Due to encountering errors as well as resource constraints, we had to cut several ablations short. Because we believe our findings to be meaningful nonetheless, we include unfinished ablations as well.

We assume that every training iteration takes the average amount of time and thus apply the same proportional normalization to every step. This assumption is fine for the conducted ablations because there is no theoretical throughput variance between steps. We normalize across time with regard to the baseline to compare in the compute-equivalent setting using the following formula:

$$t_i(s) = s\frac{1}{\bar{T}_i},$$

$$s'_i = \text{time-normalize}_i(s) = s\frac{t_i(s)}{t_{\text{baseline}}(s)},$$

where t is a time (e.g., in seconds), i is one of the ablation experiments (including the baseline), s is the iteration (step) to normalize time until, $\bar{T}$ are averaged throughput values (e.g., iterations per second), and s'_i is a time-normalized step for experiment i. This normalization allows us to compare the training and validation loss at each point in time and express that a method with lower loss at the same point in time as the baseline is more compute-efficient up to that point in time with regard to that loss.

The following ablations were conducted:

1. **SwiGLU:** replace the first layer of the MLP part of the Transformer with a T5-style (i.e., without biases) (Raffel et al., 2020; Google, 2020) Swish-activated (Ramachandran et al., 2017) gated linear unit layer (Dauphin et al., 2017; Shazeer, 2020).
2. **Untied in/out embedding:** learn separate weights for the input embedding and output "unembedding" layers (Google, 2020).
3. **No `Linear` biases:** remove all bias terms in `Linear` layers (Google, 2020). The ablation

ran until 33 000 steps and was not evaluated due to being deemed too far from completion.

4. **No GPT-like weight init:** whether to scale weight initialization in layers that a residual path leads into as a function of depth (Radford et al., 2019). The ablation ran until 39 000 steps and was not evaluated due to being deemed too far from completion.
5. **Head scaling:** multiply each Attention head's output by a learned scalar factor (Shleifer et al., 2021).
6. **No dropout:** disable dropout (Srivastava et al., 2014) in all layers (Google, 2020). The latest checkpoint available for evaluation was at 21 000 steps. While it showed a strongly monotonic improvement over the baseline, it is especially hard to interpret improvement in training loss in the dropout vs. no dropout setting. We evaluated this change even though it was far from finishing training because its improvements were so drastic.
7. **RMSNorm:** replace LayerNorm normalization layers (Lei Ba et al., 2016) with root mean square layer normalization layers (Zhang and Sennrich, 2019).
8. **NoPE:** no position embedding (Kazemnejad et al., 2024); completely remove position embeddings.
9. **ALiBi:** replace the Rotary position embedding (Su et al., 2024) with the Attention with linear biases position embedding (Press et al., 2021). Because we were unable to use an optimized ALiBi kernel for this ablation, throughput-normalization was not considered out of fairness.
10. **GQA (2 groups):** replace multi-head Attention with grouped-query Attention (Ainslie et al., 2023) with 2 groups (i.e., 2 key/value heads).
11. **Adan (4× base LR):** replace the AdamW optimizer with the Adan optimizer (Xie et al., 2024), using the baseline's learning rate multiplied by 4. The increased learning rate was chosen based on previous small-scale experiments.
12. **2×/4× base LR:** use the baseline's learning rate multiplied by 2 or 4. The ablation with a factor 4 increase did not run until completion; its checkpoint used for evaluation was saved at 48,000 steps, 5 100 steps before the end of training, and was deemed "close enough" to completion to provide a fair evaluation comparison, especially due to its noticeable training loss improvements.

We decided to implement most of the "free lunch" improvements and some neutral results based around current research results at that point in time while disregarding some findings that were deemed too experimental and/or risky.

Notably, we decided to use neither Adan nor 4× the learning rate despite these changes yielding the best results. Instead, to minimize risks, we opted to use the same learning rate as Llama-2 (Touvron et al., 2023b). Due to our inexperience at training at this scale, we were worried about possible convergence problems if we had chosen these important parameters incorrectly.

In summary, the chosen changes are: SwiGLU, no biases, no dropout during pre-training, RMSNorm, GQA (with 2 groups). GQA was chosen despite its comparatively worse results to optimize for inference in the post-training setting.

A.3 Impact of Educational Content

Figures 8-12 demonstrate the development of the model downstream performance between 0.99T and 4T tokens for English, French, German, Finnish, and Estonian. The grey area in the figures highlights the ablation in which we compare the performance of the EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens.

A.4 Continued Pre-Training

Because we decided to further train our base model from 3T tokens to 4T tokens, we needed to define a new learning rate schedule. We ablated three different learning rate schedules *Full-Rewarmup*, *Partial-Rewarmup*, and *Continue Decay*. In the Full-Rewarmup setting, we defined a cosine learning rate schedule with the same minimal and maximal learning as for the initial pre-training phase. In the Partial-Rewarmup, we set the maximum learning rate to a quarter of the maximum learning rate of the initial pre-training phase. Finally, in the Continue-Decay setting, we decreased the learning rate to 8.3% of the maximum learning rate within the following 1T tokens, providing the lowest training loss (Figure 13). Therefore, we decided to employ this setting to train our model further.

A.5 Evaluation

The tables 9-12 present the results of our base model compared to related base models.

Change	ARC Easy	HellaSwag	LAMBADA	Interpretation
-	0.535	0.355	0.503	0
SwiGLU	0.527	0.361	0.507	+
Untied in/out embedding	0.524	0.355	0.498	-
No `Linear` biases (33k st.)	-	-	-	+?
No GPT-like weight init (39k st.)	-	-	-	-
Head scaling	0.527	0.356	0.493	-
No dropout (21k st.)	*0.492*	*0.334*	*0.414*	+?
RMSNorm	0.530	0.358	0.502	0
NoPE	0.516	0.351	0.486	-
ALiBi	0.527	0.349	0.486	-
GQA (2 groups)	0.513	0.346	0.459	?
Adan (4× base LR)	0.544	**0.374**	**0.522**	+?
2× learning rate	0.540	0.369	0.514	+
4× learning rate (48k st.)	***0.545***	*0.371*	*0.517*	+

Table 6: Selected evaluation results. Bold is best, underlined is better than baseline. Italic means the run was evaluated before finishing. Ablations without values were not evaluated. The rightmost column contains a subjective interpretation/recommendation, where "+", "-", "0" indicate a positive, negative, and neutral interpretation, respectively."?" indicates it is difficult to make a conclusive statement.

A.5.1 Toxicity

We evaluate our models compared to other instruction-tuned models on the PolygloToxicityPrompts benchmark (Jain et al., 2024). We report results from $\text{PTP}_{\text{SMALL}}$. Our evaluation settings are the same as those detailed in Section 3.4 of (Jain et al., 2024), except the repetition parameter K which is set to $K = 1$ in our setup. We further report toxicity and profanity metrics provided by the Perspective API (Lees et al., 2022). For both attributes $i \in \{\text{Profanity}, \text{Toxicity}\}$ we report the EMPIRICAL PROBABILITY denoted EP_i as well as the AVERAGE denoted A_i as defined in Section 3.4 of (Jain et al., 2024). Tables 34 to 39 show the results by language as well as aggregated, for the languages German, English, French, Italian and Spanish.

A.6 Instruction Tuning

In the following, we provide more insights regarding our instruction tuned model. Particularly, we give an overview of the training datasets that we employed for instruction tuning the model (Appendix A.6.1), and present additional evaluation results (Appendix A.6.2).

A.6.1 Instruction Tuning Datasets

Table 40 provides an overview of the employed instruction tuning datasets.

A.6.2 Instruction Tuning Evaluation Results

In Figure 14, we present the evaluation results of MT-Bench-X (Weber et al., 2024). For the discussion of the results, we refer to Section 7.6.

A.7 Software

We selected our training framework based on efficiency in terms of throughput (TFLOP/s) and maintenance. Therefore, we decided to train our models based on a fork of Megatron-LM (Korthikanti et al., 2023) that supports various scalability features ensuring efficient training of transformer-based decoder-only models. Implementation contributions include a tensor-model-parallel SwiGLU layer (Ramachandran et al., 2017; Shazeer, 2020; Raffel et al., 2020; Google, 2020) and fused RMSNorm (Zhang and Sennrich, 2019) integration. In particular, we made use of 3D parallelism, i.e., data, tensor, and pipeline parallelism. Additionally, we used ZeRO (Korthikanti et al., 2023) to reduce memory requirements further by sharding the optimizer state.

A.8 Training Infrastructure

We trained our models for 812.321 GPU hours on a supercomputer, which comprises 936 compute nodes, each containing 4× NVIDIA A100 (40 GB) GPUs connected via NVLink3 intra-node and through Mellanox HDR200 InfiniBand (IB) inter-node.

Hyper-Parameter	Value
Training Objective	CLM
Activation Function	SwiGLU
Seq Length	4096
Position Embeddings	Rotary
Num Layers	32
Hidden Size	4096
FFN Hidden Size	13440
Num Attention Heads	32
Head Dim	128
Group Query Attention	yes
Num Query Groups	2
Normalization	RMSNorm
Learning rate	3e-4
Min learning rate	3e-5
Disable bias in linear	yes
Hidden dropout	0.0
Attention dropout	0.0
Optimizer	AdamW
Beta1	0.9
Beta2	0.95
Data-type	bf16

Table 7: Hyper-Parameter Configuration

We utilized a maximum of 1024 GPUs and a minimum of 32 GPUs for our training runs, achieving perfect to near linear scaling behavior. Ablations were conducted to optimize the training runs for the system and network configuration in terms of parallelization, memory, and GPU utilization. The model employed tensor and pipeline parallelism of 2, along with sequence parallelism and optimizer sharding. State-of-the-art methods, such as flash attention and mixed-precision training, were also enabled, leading to maximum utilization of the hardware.

When scaling beyond 64 GPUs, the training became more susceptible to hardware failures. We encountered issues with high-bandwidth NCCL communications triggering hardware failures initiated by the IB port restarting. To enhance the robustness and fault tolerance of the training runs against node and port failures, we used environment variables for extending timeouts such as `NCCL_IB_TIMEOUT` and `UCX_RC_TIMEOUT`, along with `NCCL_ASYNC_ERROR_HANDLING=1`. Furthermore, model checkpoints were saved at regular intervals and whenever errors or exit signals of the scheduler (Slurm) were detected.

The training runs were actively monitored using TensorBoard and custom cluster monitoring tool aiding to better visualization and debugging opportunities. We also implemented an automatic job submission strategy to combat hardware failures and queue times.

A.9 License

The base model will be released under a permissive license and the instruction model will be released under a research license.

A.10 Use of AI Assistants

We acknowledge the use of ChatGPT-4 to generate code for the Figures 1-5, and Figures 8-13 that we carefully checked and adapted.

Model	Average	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Aya-23-8B	.563	.593	.661	.484	.512
Salamandra-7B-Instruct	.557	.620	.664	.476	.468
Bloomz-7B1	.394	.373	.419	.456	.327
Ours (Instruct)	.560	.610	.653	.541	.437

Table 8: Instruct model results on multilingual benchmarks across the 10 common languages (Czech, Dutch, English, French, Greek, Italian, Polish, Portuguese, Romanian, and Spanish). The tables include mean accuracy across languages.

Model	Average	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B	.548	.554	.588	.495	.556
Salamandra-7B	.523	.589	.637	.449	.417
Mistral-7B-v0.3	.505	.513	.534	.472	.501
Occiglot-7B-eu5	.464	.470	.511	.448	.426
Pharia-1-LLM-7B-control	.409	.393	.433	.456	.353
Bloom-7B1	.348	.319	.355	.464	.256
Ours (Base)	.496	.550	.615	.469	.349

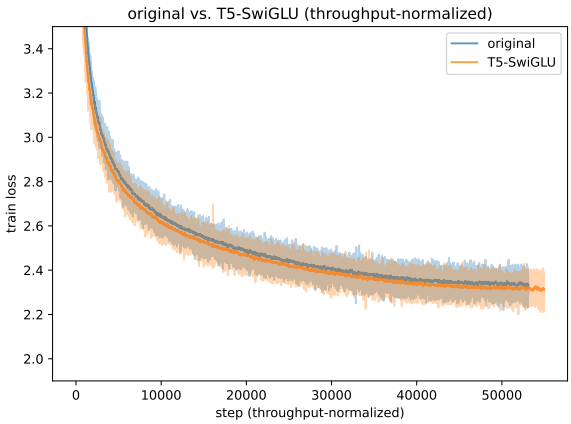
Table 9: Results on multilingual benchmarks for 21 European languages with base models.

Model	Average	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B	.601	.634	.687	.474	.607
Mistral-7B-v0.3	.580	.632	.662	.453	.571
Occiglot-7B-eu5	.568	.630	.702	.424	.514
Pharia-1-LLM-7B	.559	.635	.686	.453	.461
Salamandra-7B	.545	.631	.685	.429	.435
Bloom-7B1	.406	.448	.501	.416	.260
Ours (Base)	.521	.599	.667	.445	.371

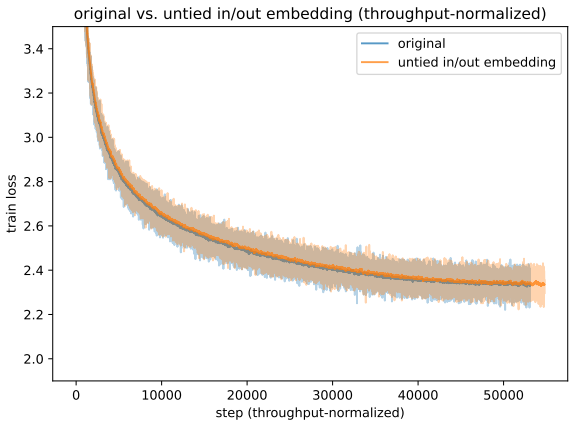
Table 10: Results on multilingual benchmarks for 6 Languages with base models.

Model	Average	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Meta-Llama-3.1-8B	.527	.522	.549	.503	.535
Salamandra-7B	.514	.572	.617	.457	.410
Mistral-7B-v0.3	.475	.466	.482	.479	.473
Occiglot-7B-eu5	.422	.406	.435	.458	.391
Pharia-1-LLM-7B	.349	.296	.332	.457	.310
Bloom-7B1	.325	.267	.296	.483	.254
Ours (Base)	.486	.531	.594	.478	.341

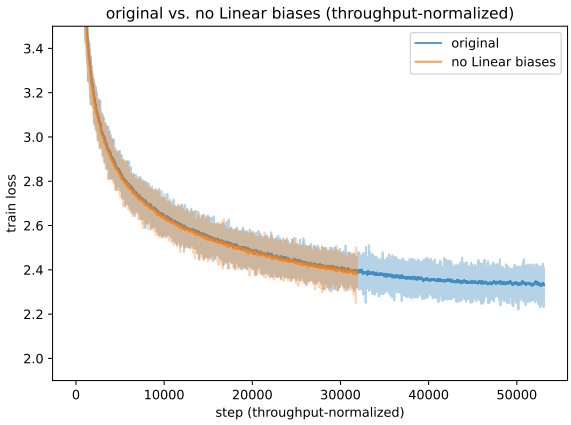
Table 11: Results on multilingual benchmarks across 15 exclusive European languages with base models.



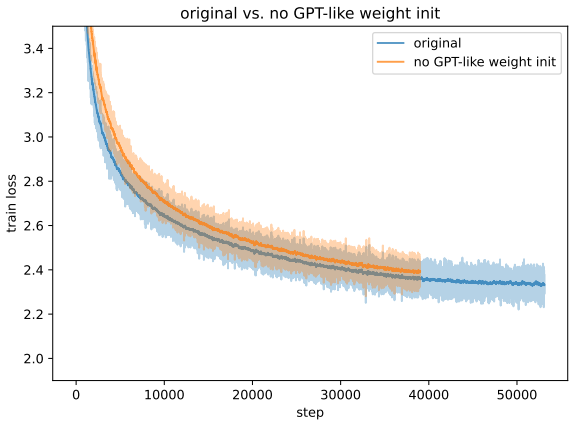
(a) "SwiGLU" ablation (time-normalized)



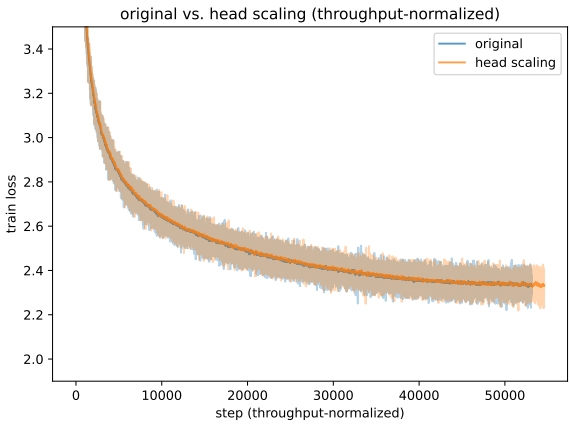
(b) "Untied in/out embedding" ablation (time-normalized)



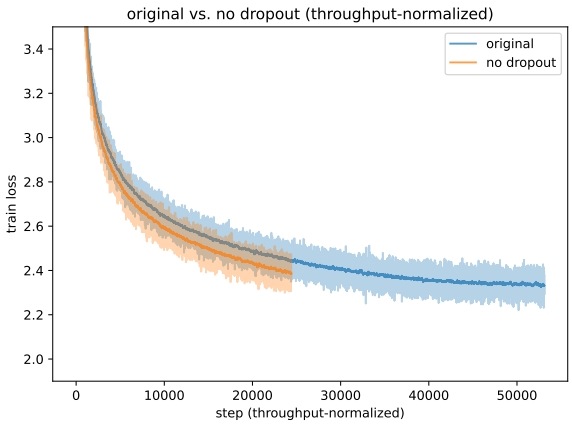
(c) "No `Linear` biases" ablation (time-normalized)



(d) "No GPT-like weight init" ablation

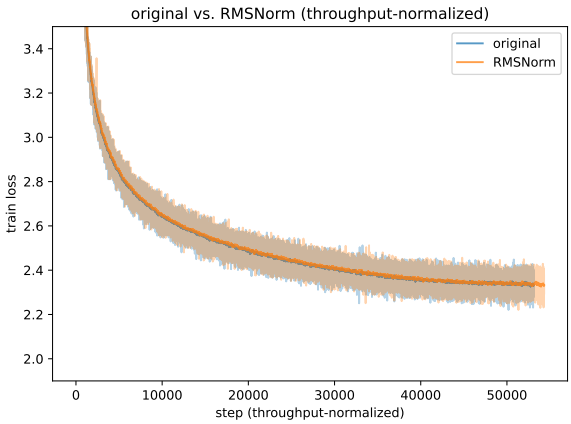


(e) "Head scaling" ablation (time-normalized)

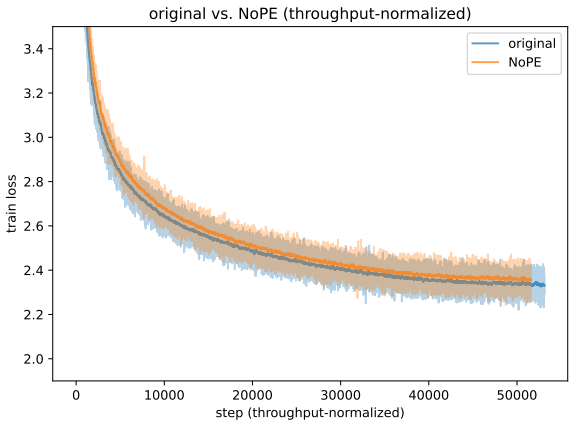


(f) "No dropout" ablation

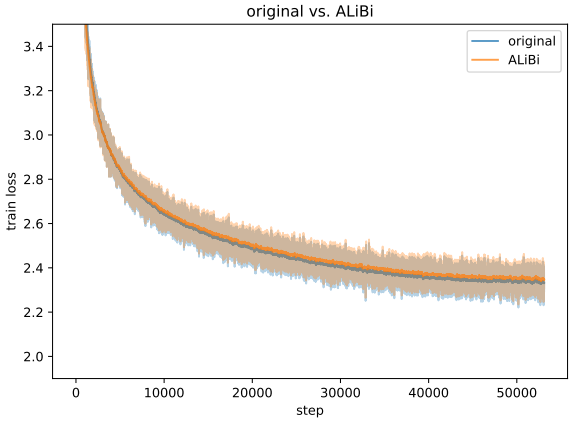
Figure 6: Training loss curves using (a) SwiGLU, (b) untied input and output embeddings, (c) no `Linear` biases, (d) no GPT-like weight initialization, (e) per-head scaling factors, and (f) no dropout. All plotted against the baseline.



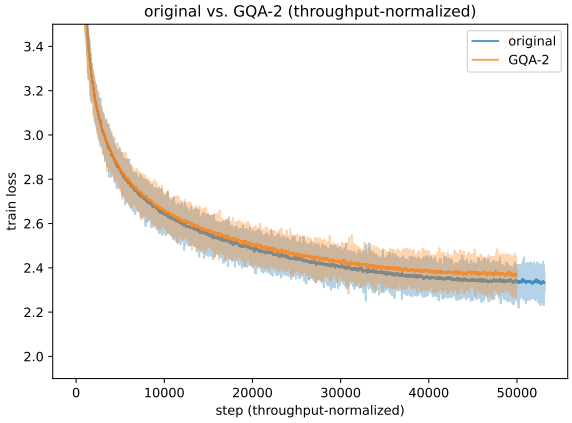
(a) “RMSNorm” ablation (time-normalized)



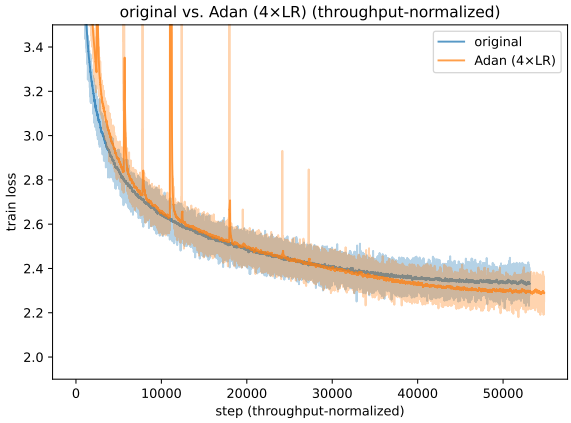
(b) “NoPE” ablation (time-normalized)



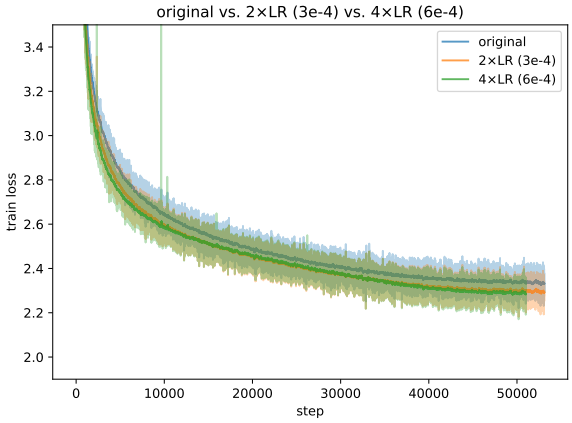
(c) “ALiBi” ablation (*not* time-normalized out of fairness)



(d) “GQA (2 groups)” ablation (time-normalized)



(e) “Adan” ablation (time-normalized)



(f) “2×/4× base LR” ablation

Figure 7: Training loss curves using (a) RMSNorm, (b) no position embedding, (c) Attention with linear biases position embedding, (d) grouped-query Attention with 2 groups, (e) the Adan optimizer, and (f) 2× or 4× the base learning rate. All plotted against the baseline.

Model	Average	EU21-ARC	EU21-HeSw	EU21-TQA	EU21-MMLU
Salamandra-7B	.536	.612	.664	.447	.423
Bloom-7B1	.376	.376	.419	.452	.258
Ours (Base)	.512	.578	.643	.466	.360

Table 12: Base model results on multilingual benchmarks across the 10 common languages (Czech, Dutch, English, French, Greek, Italian, Polish, Portuguese, Romanian, and Spanish). The tables include mean accuracy across languages.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Mistral-7B-Instruct-v0.3	.563	.591	.572	.496	.593
Meta-Llama-3.1-8B-Instruct	.547	.554	.544	.546	.546
Salamandra-7B-Instruct	.545	.595	.642	.449	.493
Mistral-7B-v0.3	.538	.563	.568	.488	.531
Meta-Llama-3.1-8B	.534	.540	.559	.516	.520
Salamandra-7B	.519	.586	.638	.389	.463
Occiglot-7B-eu5-Instruct	.464	.470	.480	.369	.538
Occiglot-7B-eu5	.448	.464	.474	.344	.511
Aya-23-8B	.422	.386	.421	.403	.477
Pharia-1-LLM-7B-ctr	.335	.264	.305	.270	.500
Pharia-1-LLM-7B-ctr-aligned	.333	.253	.310	.263	.507
Bloom-7B1	.332	.259	.295	.259	.517
Bloomz-7B1	.326	.260	.292	.251	.500
Ours (Instruct)	.527	.565	.625	.364	.555
Ours (Base)	.494	.535	.608	.323	.509

Table 13: Task accuracies for the Bulgarian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.565	.577	.571	.581	.529
Meta-Llama-3.1-8B	.553	.566	.583	.556	.508
Mistral-7B-Instruct-v0.3	.550	.580	.545	.510	.563
Aya-23-8B	.536	.569	.611	.499	.467
Salamandra-7B-Instruct	.536	.591	.624	.461	.467
Mistral-7B-v0.3	.528	.564	.540	.520	.487
Salamandra-7B	.512	.587	.627	.405	.429
Occiglot-7B-eu5-Instruct	.466	.484	.485	.435	.463
Occiglot-7B-eu5	.456	.461	.477	.436	.449
Bloomz-7B1	.334	.267	.296	.287	.488
Pharia-1-LLM-7B-ctr	.332	.275	.307	.302	.444
Pharia-1-LLM-7B-ctr-aligned	.330	.270	.310	.306	.433
Bloom-7B1	.327	.280	.298	.259	.472
Ours (Instruct)	.540	.588	.618	.423	.530
Ours (Base)	.494	.557	.601	.353	.464

Table 14: Task accuracies for the Czech language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.576	.571	.605	.596	.533
Mistral-7B-Instruct-v0.3	.564	.582	.589	.532	.555
Meta-Llama-3.1-8B	.562	.556	.616	.569	.506
Salamandra-7B-Instruct	.541	.601	.663	.462	.436
Mistral-7B-v0.3	.535	.561	.580	.542	.457
Salamandra-7B	.522	.594	.659	.425	.410
Occiglot-7B-eu5-Instruct	.490	.500	.541	.452	.466
Occiglot-7B-eu5	.472	.484	.529	.448	.429
Aya-23-8B	.441	.404	.476	.440	.444
Pharia-1-LLM-7B-ctr-aligned	.364	.319	.358	.364	.416
Pharia-1-LLM-7B-ctr	.358	.325	.354	.339	.412
Bloomz-7B1	.330	.268	.306	.306	.440
Bloom-7B1	.322	.270	.306	.253	.461
Ours (Instruct)	.548	.598	.650	.447	.496
Ours (Base)	.489	.553	.636	.349	.418

Table 15: Task accuracies for the Danish language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.591	.610	.621	.615	.520
Mistral-7B-Instruct-v0.3	.584	.620	.609	.540	.568
Meta-Llama-3.1-8B	.577	.589	.632	.590	.495
Occiglot-7B-eu5-Instruct	.575	.625	.691	.514	.469
Pharia-1-LLM-7B-ctr-aligned	.567	.618	.669	.482	.498
Aya-23-8B	.564	.589	.645	.519	.501
Occiglot-7B-eu5	.561	.615	.679	.512	.439
Mistral-7B-v0.3	.552	.595	.598	.551	.465
Pharia-1-LLM-7B-ctr	.546	.616	.659	.450	.457
Salamandra-7B-Instruct	.543	.614	.644	.466	.451
Salamandra-7B	.531	.601	.645	.432	.445
Bloomz-7B1	.340	.299	.307	.309	.442
Bloom-7B1	.323	.294	.310	.262	.427
Ours (Instruct)	.564	.621	.644	.451	.542
Ours (Base)	.507	.577	.634	.367	.452

Table 16: Task accuracies for the German language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Aya-23-8B	.541	.556	.622	.478	.506
Salamandra-7B-Instruct	.537	.574	.633	.425	.514
Meta-Llama-3.1-8B-Instruct	.530	.510	.540	.536	.535
Meta-Llama-3.1-8B	.519	.504	.553	.513	.504
Salamandra-7B	.518	.570	.633	.387	.480
Mistral-7B-Instruct-v0.3	.379	.309	.376	.362	.470
Mistral-7B-v0.3	.377	.321	.383	.364	.440
Occiglot-7B-eu5-Instruct	.350	.302	.354	.311	.434
Occiglot-7B-eu5	.350	.289	.354	.315	.440
Pharia-1-LLM-7B-ctr-aligned	.325	.251	.289	.275	.483
Pharia-1-LLM-7B-ctr	.323	.247	.290	.275	.481
Bloom-7B1	.323	.254	.287	.254	.496
Bloomz-7B1	.322	.243	.285	.273	.488
Ours (Instruct)	.522	.573	.621	.356	.539
Ours (Base)	.486	.540	.611	.315	.479

Table 17: Task accuracies for the Greek language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Mistral-7B-Instruct-v0.3	.706	.753	.846	.627	.597
Meta-Llama-3.1-8B-Instruct	.692	.736	.802	.690	.541
Meta-Llama-3.1-8B	.662	.715	.819	.661	.452
Mistral-7B-v0.3	.654	.727	.829	.635	.426
Occiglot-7B-eu5-Instruct	.624	.698	.797	.551	.449
Pharia-1-LLM-7B-ctr-aligned	.623	.712	.775	.519	.486
Aya-23-8B	.617	.670	.779	.565	.454
Salamandra-7B-Instruct	.609	.711	.772	.508	.446
Pharia-1-LLM-7B-ctr	.607	.702	.770	.503	.454
Occiglot-7B-eu5	.605	.681	.790	.543	.404
Salamandra-7B	.586	.695	.771	.453	.423
Bloomz-7B1	.496	.550	.610	.372	.452
Bloom-7B1	.449	.546	.606	.257	.389
Ours (Instruct)	.618	.702	.759	.497	.514
Ours (Base)	.563	.674	.749	.415	.413

Table 18: Task accuracies for the English language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.618	.641	.673	.627	.530
Mistral-7B-Instruct-v0.3	.602	.655	.654	.548	.552
Meta-Llama-3.1-8B	.597	.632	.679	.606	.470
Occiglot-7B-eu5-Instruct	.591	.656	.729	.520	.458
Pharia-1-LLM-7B-ctr-aligned	.578	.648	.693	.480	.492
Occiglot-7B-eu5	.576	.636	.712	.516	.438
Mistral-7B-v0.3	.574	.628	.651	.561	.456
Salamandra-7B-Instruct	.566	.643	.687	.483	.450
Aya-23-8B	.566	.608	.677	.523	.456
Pharia-1-LLM-7B-ctr	.552	.636	.681	.453	.437
Salamandra-7B	.549	.634	.686	.434	.443
Bloomz-7B1	.478	.517	.582	.379	.434
Bloom-7B1	.437	.509	.573	.260	.407
Ours (Instruct)	.565	.628	.667	.448	.520
Ours (Base)	.513	.591	.665	.354	.442

Table 19: Task accuracies for the Spanish language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Salamandra-7B-Instruct	.510	.541	.576	.441	.480
Salamandra-7B	.482	.530	.574	.407	.419
Meta-Llama-3.1-8B-Instruct	.472	.443	.452	.501	.492
Meta-Llama-3.1-8B	.464	.444	.463	.487	.464
Mistral-7B-Instruct-v0.3	.369	.311	.338	.372	.456
Mistral-7B-v0.3	.364	.293	.342	.389	.431
Occiglot-7B-eu5-Instruct	.352	.290	.334	.322	.461
Occiglot-7B-eu5	.347	.287	.333	.327	.442
Aya-23-8B	.345	.286	.327	.334	.434
Pharia-1-LLM-7B-ctr-aligned	.318	.257	.295	.288	.434
Pharia-1-LLM-7B-ctr	.316	.258	.297	.286	.423
Bloom-7B1	.315	.256	.288	.248	.466
Bloomz-7B1	.311	.263	.285	.271	.425
Ours (Instruct)	.499	.507	.544	.405	.540
Ours (Base)	.453	.484	.536	.342	.452

Table 20: Task accuracies for the Estonian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Salamandra-7B-Instruct	.512	.532	.602	.439	.476
Meta-Llama-3.1-8B-Instruct	.505	.491	.511	.527	.491
Meta-Llama-3.1-8B	.493	.483	.523	.512	.454
Salamandra-7B	.491	.540	.603	.415	.404
Mistral-7B-Instruct-v0.3	.410	.359	.404	.416	.463
Mistral-7B-v0.3	.403	.358	.406	.426	.424
Occiglot-7B-eu5-Instruct	.374	.328	.371	.351	.446
Occiglot-7B-eu5	.367	.317	.368	.354	.431
Aya-23-8B	.356	.299	.348	.349	.428
Pharia-1-LLM-7B-ctr-aligned	.325	.275	.305	.304	.417
Pharia-1-LLM-7B-ctr	.324	.281	.300	.289	.427
Bloom-7B1	.313	.258	.301	.247	.446
Bloomz-7B1	.307	.267	.296	.269	.395
Ours (Instruct)	.511	.529	.590	.414	.511
Ours (Base)	.466	.511	.587	.343	.424

Table 21: Task accuracies for the Finnish language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.612	.638	.667	.624	.521
Mistral-7B-Instruct-v0.3	.606	.647	.660	.549	.567
Meta-Llama-3.1-8B	.593	.620	.675	.598	.478
Occiglot-7B-eu5-Instruct	.581	.656	.722	.511	.435
Pharia-1-LLM-7B-ctr-aligned	.575	.643	.698	.477	.482
Mistral-7B-v0.3	.573	.627	.652	.562	.451
Aya-23-8B	.564	.610	.687	.519	.441
Occiglot-7B-eu5	.563	.643	.713	.504	.393
Salamandra-7B-Instruct	.555	.635	.681	.461	.444
Pharia-1-LLM-7B-ctr	.551	.636	.684	.458	.428
Salamandra-7B	.532	.624	.683	.430	.391
Bloomz-7B1	.461	.500	.579	.367	.396
Bloom-7B1	.431	.501	.573	.262	.387
Ours (Instruct)	.576	.623	.676	.448	.558
Ours (Base)	.519	.596	.665	.357	.457

Table 22: Task accuracies for the French language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.544	.524	.558	.566	.530
Meta-Llama-3.1-8B	.533	.522	.558	.547	.506
Salamandra-7B-Instruct	.526	.553	.592	.444	.514
Mistral-7B-Instruct-v0.3	.521	.515	.504	.488	.576
Salamandra-7B	.502	.541	.589	.400	.476
Mistral-7B-v0.3	.492	.492	.496	.501	.479
Occiglot-7B-eu5-Instruct	.434	.424	.422	.398	.491
Occiglot-7B-eu5	.426	.412	.417	.403	.470
Aya-23-8B	.378	.299	.347	.374	.493
Bloomz-7B1	.336	.268	.294	.272	.509
Pharia-1-LLM-7B-ctr-aligned	.334	.265	.297	.297	.478
Pharia-1-LLM-7B-ctr	.333	.265	.294	.292	.478
Bloom-7B1	.332	.276	.290	.244	.517
Ours (Instruct)	.518	.538	.580	.402	.554
Ours (Base)	.470	.521	.568	.299	.489

Table 23: Task accuracies for the Hungarian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.616	.635	.649	.614	.565
Mistral-7B-Instruct-v0.3	.590	.625	.627	.545	.565
Occiglot-7B-eu5-Instruct	.589	.647	.708	.519	.483
Meta-Llama-3.1-8B	.588	.624	.655	.592	.483
Pharia-1-LLM-7B-ctr-aligned	.575	.627	.679	.480	.515
Occiglot-7B-eu5	.575	.629	.702	.517	.451
Aya-23-8B	.565	.596	.663	.513	.488
Mistral-7B-v0.3	.561	.601	.618	.556	.471
Salamandra-7B-Instruct	.558	.626	.658	.472	.478
Pharia-1-LLM-7B-ctr	.554	.616	.670	.457	.473
Salamandra-7B	.536	.610	.656	.432	.448
Bloomz-7B1	.368	.339	.380	.326	.428
Bloom-7B1	.363	.345	.390	.263	.453
Ours (Instruct)	.566	.604	.649	.448	.562
Ours (Base)	.506	.573	.640	.356	.457

Table 24: Task accuracies for the Italian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Salamandra-7B-Instruct	.527	.554	.589	.444	.519
Salamandra-7B	.503	.557	.580	.394	.481
Meta-Llama-3.1-8B-Instruct	.470	.447	.442	.498	.495
Meta-Llama-3.1-8B	.465	.444	.447	.482	.486
Aya-23-8B	.396	.318	.378	.390	.499
Mistral-7B-Instruct-v0.3	.370	.292	.337	.365	.485
Mistral-7B-v0.3	.366	.289	.339	.374	.463
Occiglot-7B-eu5-Instruct	.352	.283	.328	.326	.472
Occiglot-7B-eu5	.349	.277	.326	.327	.466
Pharia-1-LLM-7B-ctr-aligned	.322	.248	.298	.286	.456
Bloom-7B1	.322	.261	.293	.259	.476
Pharia-1-LLM-7B-ctr	.319	.246	.300	.283	.447
Bloomz-7B1	.318	.254	.291	.267	.459
Ours (Instruct)	.509	.518	.558	.392	.566
Ours (Base)	.471	.506	.557	.317	.502

Table 25: Task accuracies for the Lithuanian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Salamandra-7B-Instruct	.514	.533	.565	.433	.526
Salamandra-7B	.502	.525	.570	.406	.506
Meta-Llama-3.1-8B-Instruct	.469	.432	.438	.488	.518
Meta-Llama-3.1-8B	.468	.432	.443	.482	.515
Mistral-7B-Instruct-v0.3	.375	.295	.326	.352	.525
Mistral-7B-v0.3	.373	.292	.329	.368	.502
Occiglot-7B-eu5-Instruct	.366	.294	.320	.326	.522
Occiglot-7B-eu5	.358	.281	.321	.324	.507
Aya-23-8B	.358	.283	.324	.338	.486
Pharia-1-LLM-7B-ctr-aligned	.334	.267	.294	.294	.480
Bloomz-7B1	.330	.261	.291	.276	.494
Pharia-1-LLM-7B-ctr	.328	.260	.296	.280	.473
Bloom-7B1	.327	.259	.293	.255	.503
Ours (Instruct)	.512	.507	.557	.408	.577
Ours (Base)	.477	.484	.548	.349	.527

Table 26: Task accuracies for the Latvian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.597	.592	.629	.608	.561
Mistral-7B-Instruct-v0.3	.582	.595	.593	.534	.608
Meta-Llama-3.1-8B	.580	.583	.632	.586	.520
Salamandra-7B-Instruct	.560	.603	.655	.471	.511
Aya-23-8B	.560	.576	.640	.509	.513
Pharia-1-LLM-7B-ctr-aligned	.557	.595	.652	.470	.510
Mistral-7B-v0.3	.553	.568	.584	.546	.512
Pharia-1-LLM-7B-ctr	.548	.595	.642	.448	.505
Salamandra-7B	.539	.593	.648	.431	.483
Occiglot-7B-eu5-Instruct	.512	.530	.563	.468	.485
Occiglot-7B-eu5	.498	.515	.549	.466	.465
Bloomz-7B1	.342	.264	.305	.303	.494
Bloom-7B1	.331	.277	.310	.253	.486
Ours (Instruct)	.560	.596	.642	.443	.558
Ours (Base)	.515	.561	.635	.372	.490

Table 27: Task accuracies for the Dutch language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.566	.570	.576	.572	.546
Meta-Llama-3.1-8B	.556	.569	.585	.552	.516
Mistral-7B-Instruct-v0.3	.556	.579	.561	.506	.577
Aya-23-8B	.548	.560	.620	.495	.518
Salamandra-7B-Instruct	.547	.592	.634	.460	.502
Mistral-7B-v0.3	.536	.568	.553	.519	.506
Salamandra-7B	.525	.595	.632	.409	.464
Occiglot-7B-eu5-Instruct	.474	.507	.500	.432	.457
Occiglot-7B-eu5	.458	.486	.489	.426	.430
Pharia-1-LLM-7B-ctr-aligned	.336	.282	.314	.313	.436
Bloomz-7B1	.334	.274	.299	.287	.478
Pharia-1-LLM-7B-ctr	.332	.275	.316	.297	.440
Bloom-7B1	.326	.272	.297	.260	.476
Ours (Instruct)	.546	.580	.622	.425	.558
Ours (Base)	.497	.546	.603	.350	.490

Table 28: Task accuracies for the Polish language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.608	.625	.651	.622	.535
Mistral-7B-Instruct-v0.3	.590	.625	.625	.551	.558
Meta-Llama-3.1-8B	.587	.624	.662	.596	.466
Aya-23-8B	.570	.608	.674	.517	.479
Pharia-1-LLM-7B-ctr-aligned	.563	.611	.661	.471	.508
Mistral-7B-v0.3	.562	.614	.624	.561	.449
Salamandra-7B-Instruct	.557	.631	.667	.473	.459
Pharia-1-LLM-7B-ctr	.542	.603	.649	.448	.468
Occiglot-7B-eu5-Instruct	.541	.593	.623	.490	.457
Salamandra-7B	.535	.619	.669	.430	.422
Occiglot-7B-eu5	.526	.578	.615	.491	.420
Bloomz-7B1	.455	.490	.555	.373	.403
Bloom-7B1	.433	.490	.553	.255	.434
Ours (Instruct)	.554	.614	.658	.450	.496
Ours (Base)	.516	.583	.652	.375	.452

Table 29: Task accuracies for the Portuguese language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.583	.589	.589	.596	.558
Meta-Llama-3.1-8B	.568	.577	.597	.580	.518
Mistral-7B-Instruct-v0.3	.563	.584	.550	.519	.597
Aya-23-8B	.559	.578	.638	.505	.516
Mistral-7B-v0.3	.546	.558	.550	.537	.540
Salamandra-7B-Instruct	.546	.598	.629	.465	.492
Salamandra-7B	.533	.595	.634	.418	.485
Occiglot-7B-eu5-Instruct	.490	.516	.499	.433	.510
Occiglot-7B-eu5	.474	.494	.488	.438	.477
Pharia-1-LLM-7B-ctr-aligned	.377	.322	.336	.352	.497
Pharia-1-LLM-7B-ctr	.372	.314	.333	.350	.490
Bloomz-7B1	.346	.287	.302	.299	.497
Bloom-7B1	.341	.287	.302	.258	.518
Ours (Instruct)	.553	.589	.615	.434	.572
Ours (Base)	.508	.558	.605	.355	.513

Table 30: Task accuracies for the Romanian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.535	.537	.521	.559	.525
Salamandra-7B-Instruct	.530	.581	.613	.462	.465
Meta-Llama-3.1-8B	.523	.522	.531	.542	.496
Salamandra-7B	.507	.586	.611	.411	.421
Mistral-7B-Instruct-v0.3	.476	.468	.468	.475	.491
Mistral-7B-v0.3	.458	.454	.464	.484	.432
Aya-23-8B	.456	.429	.489	.450	.454
Occiglot-7B-eu5-Instruct	.422	.418	.438	.412	.420
Occiglot-7B-eu5	.418	.406	.430	.412	.425
Bloomz-7B1	.330	.262	.291	.281	.488
Pharia-1-LLM-7B-ctr-aligned	.326	.264	.302	.313	.426
Pharia-1-LLM-7B-ctr	.324	.255	.301	.302	.437
Bloom-7B1	.316	.263	.294	.255	.453
Ours (Instruct)	.525	.552	.598	.423	.527
Ours (Base)	.481	.524	.590	.354	.454

Table 31: Task accuracies for the Slovak language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Salamandra-7B-Instruct	.528	.574	.599	.453	.486
Mistral-7B-Instruct-v0.3	.518	.537	.510	.495	.531
Meta-Llama-3.1-8B-Instruct	.513	.510	.489	.539	.513
Salamandra-7B	.512	.571	.600	.419	.456
Mistral-7B-v0.3	.503	.529	.506	.501	.476
Meta-Llama-3.1-8B	.502	.509	.501	.525	.474
Occiglot-7B-eu5-Instruct	.427	.432	.437	.394	.444
Occiglot-7B-eu5	.420	.418	.434	.393	.437
Aya-23-8B	.391	.334	.379	.395	.457
Bloomz-7B1	.319	.242	.292	.275	.467
Pharia-1-LLM-7B-ctr-aligned	.318	.254	.300	.299	.419
Pharia-1-LLM-7B-ctr	.317	.259	.300	.295	.413
Bloom-7B1	.313	.256	.294	.246	.454
Ours (Instruct)	.527	.555	.577	.418	.558
Ours (Base)	.477	.512	.575	.345	.474

Table 32: Task accuracies for the Slovenian language.

Model	Average	EU21-ARC	EU21-HeSw	EU21-MMLU	EU21-TQA
Meta-Llama-3.1-8B-Instruct	.601	.594	.636	.596	.579
Meta-Llama-3.1-8B	.592	.586	.644	.579	.558
Mistral-7B-Instruct-v0.3	.585	.601	.603	.527	.608
Salamandra-7B-Instruct	.566	.620	.664	.459	.519
Mistral-7B-v0.3	.555	.576	.595	.542	.507
Salamandra-7B	.543	.608	.660	.428	.477
Occiglot-7B-eu5-Instruct	.510	.512	.548	.452	.526
Occiglot-7B-eu5	.490	.496	.531	.445	.486
Aya-23-8B	.459	.418	.482	.444	.494
Pharia-1-LLM-7B-ctr-aligned	.378	.324	.352	.360	.476
Pharia-1-LLM-7B-ctr	.374	.320	.347	.340	.490
Bloomz-7B1	.339	.263	.294	.301	.497
Bloom-7B1	.332	.275	.298	.257	.498
Ours (Instruct)	.566	.606	.654	.436	.566
Ours (Base)	.509	.567	.640	.343	.483

Table 33: Task accuracies for the Swedish language.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.174	.196	.173	.219
Mistral-7B-Instruct-v0.3	.139	.160	.140	.183
Salamandra-7B-Instruct	.206	.216	.196	.230
Aya-23-8B	.177	.197	.176	.218
Occiglot-7B-eu5-Instruct	.202	.212	.193	.226
Bloomz-7B1	.100	.115	.106	.134
Ours (Instruct)	.216	.223	.201	.230

Table 34: The aggregated $PTP_{small}^{Aggregated}$ evaluated on instruction-tuned models.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.162	.184	.188	.235
Mistral-7B-Instruct-v0.3	.186	.188	.214	.236
Aya-23-8B	.212	.216	.235	.262
Bloomz-7B1	.122	.132	.149	.173
Occiglot-7B-eu5-Instruct	.175	.184	.193	.226
Salamandra-7B-Instruct	.205	.207	.221	.251
Ours (Instruct)	.191	.191	.204	.227

Table 35: $PTP_{small}^{English}$ evaluated on instruction-tuned models.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.035	.081	.029	.097
Mistral-7B-Instruct-v0.3	.022	.066	.022	.086
Aya-23-8B	.039	.085	.030	.098
Bloomz-7B1	.017	.049	.020	.066
Occiglot-7B-eu5-Instruct	.054	.097	.045	.116
Salamandra-7B-Instruct	.045	.091	.037	.102
Ours (Instruct)	.045	.088	.038	.102

Table 36: PTP_{small}^{German} evaluated on instruction-tuned models.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.264	.258	.250	.268
Mistral-7B-Instruct-v0.3	.156	.171	.138	.179
Aya-23-8B	.227	.232	.221	.245
Bloomz-7B1	.127	.129	.119	.136
Occiglot-7B-eu5-Instruct	.301	.281	.275	.272
Salamandra-7B-Instruct	.314	.294	.286	.288
Ours (Instruct)	.329	.308	.294	.289

Table 37: PTP_{small}^{French} evaluated on instruction-tuned models.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.153	.188	.170	.230
Mistral-7B-Instruct-v0.3	.111	.145	.135	.189
Aya-23-8B	.135	.176	.151	.215
Bloomz-7B1	.063	.090	.077	.119
Occiglot-7B-eu5-Instruct	.184	.208	.192	.241
Salamandra-7B-Instruct	.170	.194	.175	.227
Ours (Instruct)	.194	.212	.198	.240

Table 38: $PTP_{small}^{Italian}$ evaluated on instruction-tuned models.

Model	$EP_{Profanity}$	$A_{Profanity}$	$EP_{Toxicity}$	$A_{Toxicity}$
Meta-Llama-3.1-8B-Instruct	.259	.268	.227	.266
Mistral-7B-Instruct-v0.3	.220	.228	.193	.225
Aya-23-8B	.275	.277	.243	.269
Bloomz-7B1	.172	.177	.163	.177
Occiglot-7B-eu5-Instruct	.293	.287	.262	.277
Salamandra-7B-Instruct	.296	.295	.261	.282
Ours (Instruct)	.321	.315	.272	.290

Table 39: $PTP_{small}^{Spanish}$ evaluated on instruction-tuned models.

Subset	Language	Count	Ø Turns	License
anon8231489123/ShareGPT_Vicuna_unfiltered	EN	37.6K	4.6	Apache-2.0
MBZUAI/Bactrian-X	EN	26.9K	1.0	MIT
Open-Orca/OpenOrca	EN	26.9K	1.0	MIT
WizardLM/WizardLM_evol_instruct_70k	EN	26.9K	1.0	MIT
WizardLM/WizardLM_evol_instruct_V2_196k	EN	26.8K	1.0	MIT
sahil2801/CodeAlpaca-20k	EN	12.1K	1.0	CC-BY-4.0
lmsys/lmsys-chat-1m	EN	11.2K	1.5	commercial
HuggingFaceH4/ultrachat_200k	EN	7.0K	3.2	MIT
MBZUAI/Bactrian-X	DE	63.7K	1.0	MIT
FreedomIntelligence/evol-instruct-deutsch	DE	55.9K	1.0	Not provided
FreedomIntelligence/alpaca-gpt4-deutsch	DE	47.5K	1.0	Apache-2.0
FreedomIntelligence/sharegpt-deutsch	DE	5.8K	2.9	Apache-2.0
LeoLM/German_Songs	DE	943	1.5	Not provided
bjoernp/ultrachat_de	DE	909	1.0	MIT
LeoLM/German_Poems	DE	378	1.0	Not provided

Table 40: Instruction-tuning training dataset composition.

Subset	Language	Count	Ø Turns
CohereForAI/aya_collection	Multilingual	39.9K	1.0
MBZUAI/Bactrian-X	Multilingual	32.3K	1.0
CohereForAI/aya_dataset	Multilingual	13.1K	1.0
anon8231489123/ShareGPT_Vicuna_unfiltered	EN	37.7K	4.6
sahil2801/CodeAlpaca-20k	EN	12.1K	1.0
lmsys/lmsys-chat-1m	EN	11.2K	1.5
WizardLM/WizardLM_evol_instruct_V2_196k	EN	8.1K	1.0
Open-Orca/OpenOrca	EN	8.1K	1.0
WizardLM/WizardLM_evol_instruct_70k	EN	8.1K	1.0
HuggingFaceH4/ultrachat_200k	EN	6.9K	3.1
anon8231489123/ShareGPT_Vicuna_unfiltered	transl. DE	37.6K	4.6
lmsys/lmsys-chat-1m	transl. DE	11.2K	1.5
WizardLM/WizardLM_evol_instruct_70k	transl. DE	8.1K	1.0
Open-Orca/OpenOrca	transl. DE	8.1K	1.0
WizardLM/WizardLM_evol_instruct_V2_196k	transl. DE	8.1K	1.0
HuggingFaceH4/ultrachat_200k	transl. DE	6.9K	3.1
FreedomIntelligence/sharegpt-deutsch	DE	5.7K	2.9
bjoernp/ultrachat_de	DE	905	1.0

Table 41: Instruction-tuning training dataset composition for 24 EU languages.

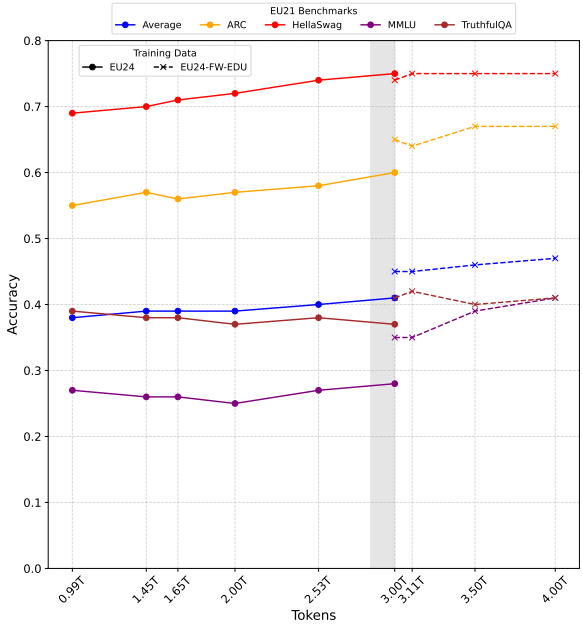


Figure 8: Downstream performance of the base model from 0.99T to 4T tokens for English. The grey area highlights the ablation comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens

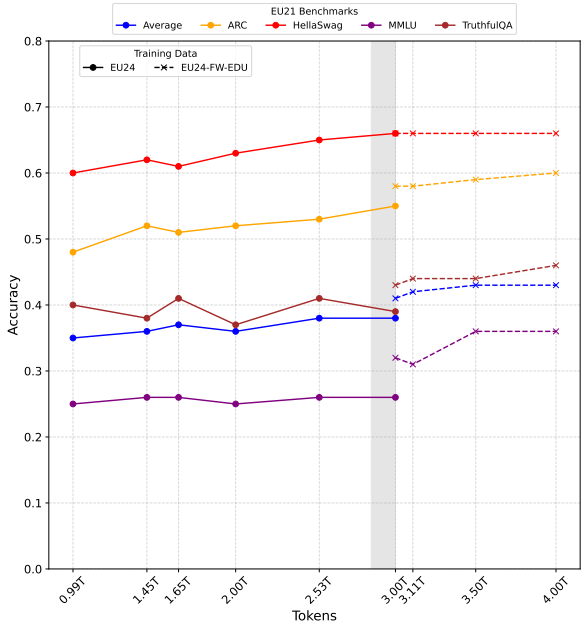


Figure 9: Downstream performance of the base model from 0.99T to 4T tokens for French. The grey area highlights the ablation comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens

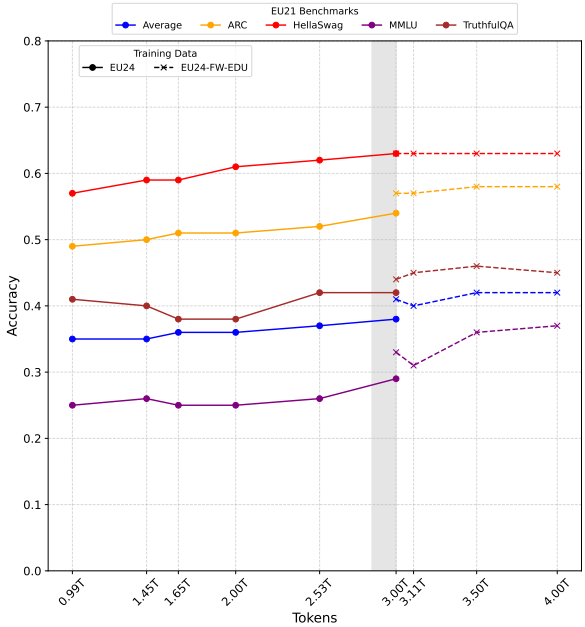


Figure 10: Downstream performance of the base model from 0.99T to 4T tokens for German. The grey area highlights the ablation comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens

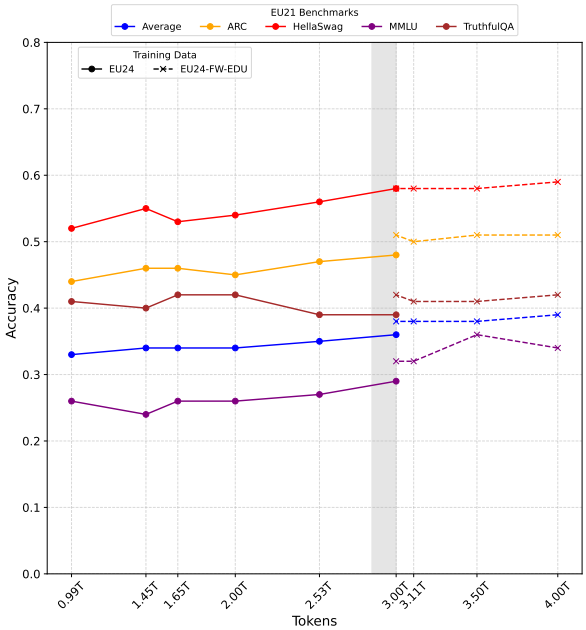


Figure 11: Downstream performance of the base model from 0.99T to 4T tokens for Finish The grey area highlights the ablation comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens

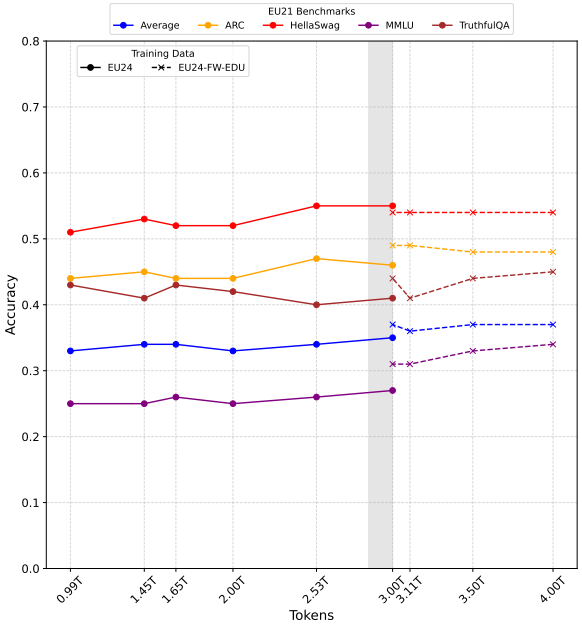


Figure 12: Downstream performance of the base model from 0.99T to 4T tokens for Estonian The grey area highlights the ablation study comparing the performance of EU24 data to the FineWeb-EDU dataset between 2.85T and 3T tokens

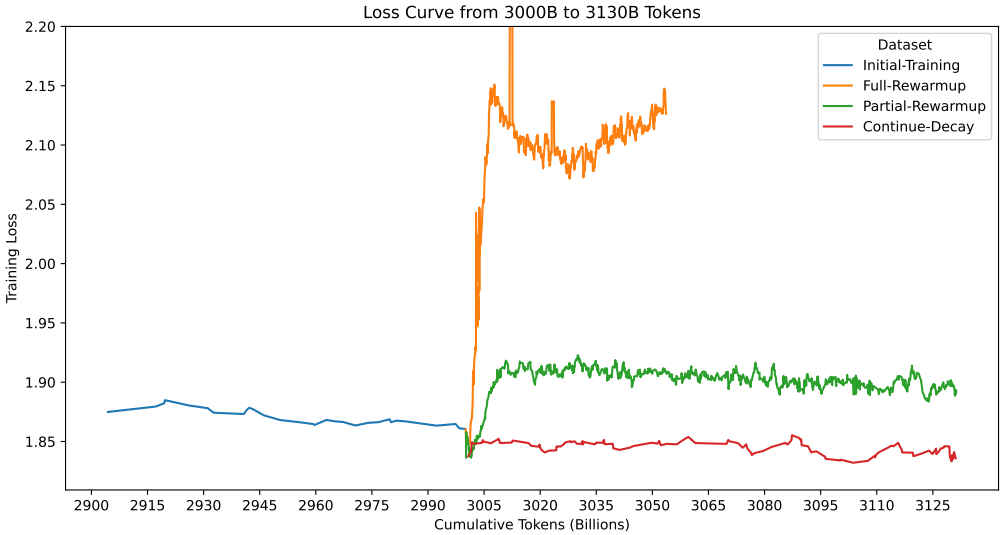


Figure 13: Loss curves up to 130B tokens for the investigated learning rate schedules.

(a) Cross-lingual avg. MT-Bench-X

(b) MT-Bench-EN

(c) MT-Bench-DE

(d) MT-Bench-FR

(e) MT-Bench-IT

(f) MT-Bench-ES

Figure 14: In-depth MT-Bench-X quality assessment by GPT-4.

LAN	Curated (M)	Web (M)	Curated in %	Total (M)	%
bg	21 233	23 810	47.14	45 042	1.13
cs	6 229	46 919	11.72	53 148	1.33
da	6 220	18 257	25.41	24 477	0.61
de	36 850	312 006	10.56	348 856	8.72
el	21 078	40 494	34.23	61 572	1.54
en	215 204	1 452 819	12.90	1 668 024	41.70
es	23 788	295 850	7.44	319 638	7.99
et	8 785	6 253	58.42	15 038	0.38
fi	3 234	35 996	8.24	39 231	0.98
fr	30 606	333 399	8.41	364 005	9.10
ga	474	75	86.26	549	0.01
hr	15 296	1	99.99	15 298	0.38
hu	3 376	37 551	8.25	40 927	1.02
it	19 834	169 069	10.50	188 903	4.72
lt	1 990	9 141	17.88	11 131	0.28
lv	1 980	5 683	25.84	7 663	0.19
mt	3 515	0.5	99.99	3 516	0.09
nl	6 788	125 106	5.15	131 893	3.30
pl	9 397	67 361	12.24	76 758	1.92
pt	6 670	136 284	4.67	142 954	3.57
ro	4 524	25 988	14.83	30 512	0.76
sk	40 114	11 162	78.23	51 276	1.28
sl	12 599	1 228	91.12	13 827	0.35
sv	4 079	40 070	9.24	44 148	1.10
code	301 612	-	100.00	301 612	7.54
Total	805 477	3 194 523	51.57	4 000 000	100.00

Table 42: Comparison of token counts (in millions) between curated and web data across different languages. The total word count reflects the sum of both curated and web-sourced data after deduplication and filtering. Source code is treated separately from natural language data.