SambaLingo: Teaching Large Language Models New Languages

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan Leon Zhang
Yun Du, Hengyu Zhao, Changran Hu, Urmish Thakker
SambaNova Systems
`zoltan.csaki@sambanovasystems.com`

Abstract

Despite the widespread availability of LLMs, there remains a substantial gap in their capabilities and availability across diverse languages. One approach to address these issues has been to take an existing pre-trained LLM and continue to train it on new languages. While prior works have experimented with language adaptation, many questions around best practices and methodology have not been covered. In this paper, we present a comprehensive investigation into the best practices for adapting LLMs to new languages. Our study explores the key components in this process, including vocabulary extension and initialization of new tokens, direct preference optimization and the data scarcity problem for human alignment in low-resource languages. We scale these experiments across 9 languages and 2 parameter scales (7B and 70B). We compare our models against Llama 2, Aya-101, XGLM, BLOOM and existing language experts, outperforming all prior published baselines. Additionally, all evaluation code[1] and checkpoints[2] are made public to facilitate future research.

1 Introduction

New state of the art large language models are being released at a breakneck speed, yet their training data, tokenizer, and evaluations remain primarily centered around a few popular languages such as English, Chinese, French and Arabic. In principle, the way to create large language models for specific languages is to pre-train models from scratch (Sengupta et al., 2023; Zhang et al., 2020). However, it is difficult to obtain a large amount of compute resources and a vast quantity of data in diverse languages. Researchers have tackled this problem by training monolithic multi-lingual models that cover a wide range of languages (Workshop et al., 2023; Lin et al., 2022; Shliazhko et al., 2023; Xue et al., 2021). These models can still struggle to achieve uniformly good results across all languages due to various factors such as the curse of multilinguality (Chang et al., 2023; Conneau et al., 2020) and the scarcity of pre-training data in many languages (Chung et al., 2023).

Recently, adapting English centric models to new languages has gained prominence (Blevins et al., 2024; Yong et al., 2023; Ebrahimi and Kann, 2021; Pires et al., 2023; Pipatanakul et al., 2023; Lin et al., 2024). The resulting models can outperform large multilingual models and even language specific models pre-trained from scratch. Adaptation requires various design choices around the tokenizer, data, alignment and evaluation strategies. This paper aims to provide a comprehensive study to help inform these decisions, outlining a clear protocol to adapt a pre-trained model to a new language. We show that our methodology works by training models across 9 languages and 2 parameter scales (7B and 70B) and comparing them against publicly available models. Figure 1 and 2 show that our methodology can lead to better models than existing state of the art models in these languages.

The key studies and contributions include:

- Best practices for adapting existing LLMs to new languages scaled across 9 typologically and linguistically diverse languages including Arabic, Bulgarian, Hungarian, Japanese, Russian, Serbian, Slovenian, Thai, and Turkish
  - Expanding the vocabulary for the target language improves the tokenizer fertility (12), but does not have a siginficant impact on downstream accuracy (5.1.1)
  - Various embedding initialization methods have minimal impact on accuracy,

[1]Fork of lm-evaluation-harness Gao et al., 2023 with new multilingual benchmarks: lm-evaluation-harness

[2]All SambaLingo Checkpoints: SambaLingo Checkpoints

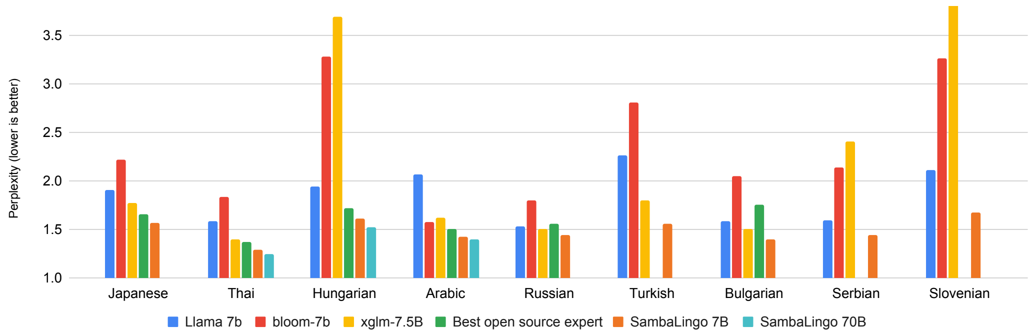


Figure 1: Evaluation perplexity on hold out dataset, we also evaluate perplexity over wikipedia and Mc4 in appendix F. Open source expert baselines: Japanese - Swallow-7b-hf (TokyoTech, 2023), Thai: typhoon-7b (Pipatanakul et al., 2023), Arabic: jais-13b (Sengupta et al., 2023), Hungarian: PULI-GPTrio (Yang et al., 2023), Russian: saiga-7b (Gusev, 2023), Bulgarian: mGPT-bulgarian(Shliazhko et al., 2023). We could not find Serbian, Slovenian and Turkish languages models with low enough perplexity that would fit the graph so we chose to omit them here to ensure readability.

  but sub word averaging accelerates training loss convergence (5.1.2)
  - The quality of the base checkpoint on English benchmarks can improve downstream language adaptation results (5.3)

- A recipe for human preference alignment in any language using open source data
  - Aligning the adapted model requires minimal data from the target language, reducing the need of gathering expensive alignment data (5.2.1)
  - The choice of translated versus human written alignment data does not have a large impact on win rates (5.2.2)

- Open sourcing code and checkpoints to promote future research
  - State of the art models adapted from Llama 2 in 9 languages and 2 parameter scales (7B, 70B)[2]
  - Integration of FLORES-200, SIB-200, EXAMS and multilingual perplexity benchmarks with lm-eval-harness[1] (Gao et al., 2023)

2 Related Work

While prior work has explored adapting pre-trained LLMs to new languages, they do not extensively study the methodology to do so. None of these works explore the design choices around aligning models in new languages, for example the mixture of data in the base models language and the new language or the impact of translated data on qualitative evaluation. Pires et al. (2023) and Cui et al. (2023b) adapt Llama models to Portuguese and Chinese respectively, but they do not explore the impact of vocabulary extension and/or initialization. Blevins et al. (2024) explores training language experts to break the curse of multilinguality starting from a pre-trained model, but they do not explore the impact of vocabulary extension, initialization and quality of the base model. Extension of vocabulary was discussed in Zhao et al. (2024b); Tikhomirov and Chernyshev (2023), however they do not explore token embedding initialization strategies or impact of quality of base model. Lin et al. (2024) studies simultaneous language adaptation to 500 languages. Nevertheless, they also do not answer questions around alignment or token initialization strategy. Ye et al. (2023) studies language adaptation of a wide variety of English-centric and multilingual models, however they only focus on finetuning XNLI tasks.

There has been a significant body of work around open-source multi-lingual models (Workshop et al., 2023; Lin et al., 2022; Shliazhko et al., 2023). Our work differs from the aforementioned studies as we solely focus on adapting pre-trained LLMs to new languages and not on pre-training from scratch. Notably, these multilingual open-source models tend to be pretrained on significantly fewer tokens than the base models we adapt from. As the models in this work tend to outperform these multilingual models, this presents a promising path forward for obtaining the state of the art in new languages.

3 Adaptation Methodology

We present our methodology to adapt large languages models to a new language, with state of the art results in 9 target languages: Arabic, Thai, Turkish, Japanese, Hungarian, Russian, Bulgarian, Serbian and Slovenian. We select these languages because they provide a mix of high resource and lower resources languages with diverse character sets and linguistic patterns. We additionally limit the scope of the languages studied in this paper to languages with easily available text datasets from CulturaX (Nguyen et al., 2023). See Section 4 for evaluation results on the final checkpoints produced by this methodology, and Section 5 for ablations justifying our methods.

We use the term *initial language* to describe the original language that the base model was trained on (in this case, English) and the term *target language* as the new language this model is being adapted to.

3.1 Selecting a Base Model

Our methodology starts with an existing base checkpoint instead of pre-training from scratch. Previous work has shown that starting from an existing checkpoint leads to faster training convergence, better downstream evaluation accuracy and lower compute/data requirements (Pires et al., 2023; Lin et al., 2024; Csaki et al., 2023). Section 5.3 demonstrates that it is important to select a starting checkpoint with the best results for the initial language, as that will improve the downstream results for the target language. Based on these observations, we chose Llama2 7B as our base model to adapt to target languages, the best open source model available at the time of the experiments.

We additionally scale this methodology to Llama 2 70B. Given compute restrictions, we only do this for 3 languages - Arabic, Thai and Hungarian. See Section 4.2 for in-depth comparisons of our 7B and 70B models.

3.2 Extending Model Vocabulary

Llama 2 (et al, 2023) was trained predominantly on English text, and has poor tokenizer efficiency for other languages (see Section 5.1). To address this inefficiency, we chose to extend the vocabulary of the Llama 2 tokenizer by adding non overlapping tokens from the target language and initializing them using sub-word embeddings from the original tokenizer. See Section 5.1 for experiments that justify our approach.

3.3 Continual Pre-training

We train each language independently on data that consists of a 1:3 mixture of English and target language web data biased towards the target language. Pretraining data for all languages, including English, is sourced from CulturaX (Nguyen et al., 2023). These decisions are grounded in results from previous works: Zhao et al. (2024b); Csaki et al. (2023) show that mixing in data from the base model domain helps downstream accuracy and training stability, Gupta et al. (2023) find that including a higher proportion of data from the target distribution helps improve the convergence in the target distribution, Almazrouei et al. (2023) showed the importance of cleaned web data. Additionally, hyperparameters used for training can be found in Appendix A.

3.4 Aligning To Human Preferences In Other Languages

To train a chat-aligned version of the model, we follow the two-stage approach from Tunstall et al. (2023) - supervised finetuning (SFT) followed by direct preference optimization (DPO). More details about hyperparameters for each of these phases used can be found in Appendix A.

- For SFT, we use `ultrachat-200k` (Tunstall et al., 2023), in a 1:1 ratio with a Google translated version of `ultrachat-200k`.

- For human preference alignment, we use the `ultrafeedback` (Cui et al., 2023a) and `cai-conversation-harmless` dataset (Huang et al., 2024). We mix these datasets with a 10:1 ratio of English to machine translated data. Section 5.2.1 shows that this ratio of data performs almost as well as other ratios and section 5.2.2 shows that machine-translated data can perform as well as human written data.

4 Evaluation

4.1 Quantitative Evaluation

We use a wide variety of benchmarks to quantitatively evaluate the performance of our models and compare them to prior work. See Table 1 for the full list of quantitative benchmarks. In summary, we evaluate language modeling with perplexity on a holdout set of CulturaX (Nguyen et al.,

Datasets	Task Category	Num Few-Shot	Number Of Languages	Metric
mc4 , Wikipedia	Perplexity	-	323	Perplexity
FLORES-200	Translation	8	200	CHRF
SIB-200	Text Classification	3	200	Accuracy
BELEBELE	Question Answering	3	122	Accuracy
Exams	Knowledge	3	11	Accuracy
XNLI XStoryCloze XCOPA XWinograd PAWS-X	Natural Language Understanding	0	25+	Accuracy

Table 1: Multi-lingual evaluation suite

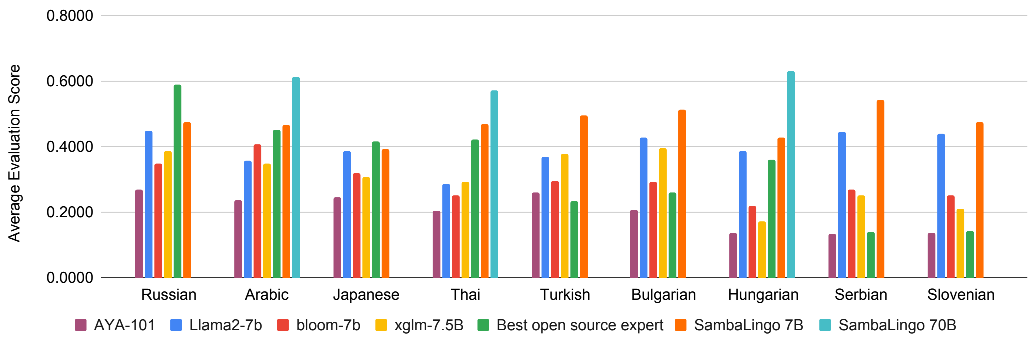


Figure 2: Quantitative evaluation results. The "best open source experts" are the same as ones specified in Figure 1. See Appendix F for the full breakdown.

2023), translation with CHRF (Popović, 2015) on FLORES-200 (Goyal et al., 2021; Zhu et al., 2023), text classification accuracy on SIB-200 (Adelani et al., 2024; Lin et al., 2024), open-book question answering on BELEBELE (Bandarkar et al., 2023), closed-book question answering on EXAMS (Hardalov et al., 2020), and a variety of natural language understanding benchmarks (XNLI (Conneau et al., 2018), XStoryCloze (Lin et al., 2022), XCOPA (Ponti et al., 2020), XWinograd (Emelin and Sennrich, 2021), and PAWS-X (Yang et al., 2019)).

All quantitative evaluations are performed on our adapted models after continuous pretraining, but before the alignment stage. We evaluate each checkpoint only on the language that it was trained on. Note that not all of our target languages are covered across all benchmarks. However, each language we examine has evaluations in at least 4 of these benchmarks. We ensured that perplexity measurements were done on a held out set in the target language, and verify that evaluating perplexity on different domains of text such as Wikipedia and MC4 (Raffel et al., 2019) have very similar results in appendix F.

4.1.1 Quantitative Results

We compare our continuously pretrained models against the best open source models available in each target language and state of the art multilingual models. Figure 1 shows that our SambaLingo models have a lower perplexity across all existing baselines on a holdout set from our training data. Perplexity on other domains also follows the same trend as shown in appendix F. Figure 2 shows the average evaluation score across the evaluation benchmarks introduced in Section 4.1, where we see our models outperform all other models in 7/9 languages.

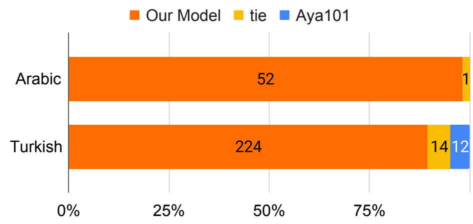
4.2 Scaling to 70B

Scaling to 70B consistently leads to better results as seen in table 2. The 70B models in the table have trained on fewer tokens than the 7B models.

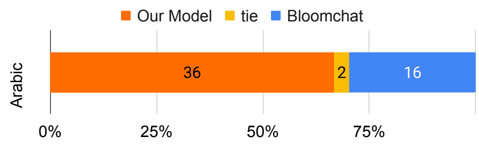
Additionally, we evaluate compute-matched checkpoints of our 7B and 70B Llama 2 models

Language	Checkpoint	ppl (↓)	FLORES EN→X (↑)	FLORES X→EN (↑)	Belebele (↑)	SIB-200 (↑)	XNLI (↑)	XStoryCloze (↑)
Arabic	70B	**1.44**	**54.25**	**65.60**	**0.78**	**0.69**	0.33	**0.68**
	7B	**1.44**	53.67	61.66	0.29	0.26	**0.34**	0.65
Hungarian	70B	**1.57**	**58.81**	**64.03**	**0.82**	**0.64**	-	-
	7B	1.63	52.70	58.31	0.33	0.25	-	-

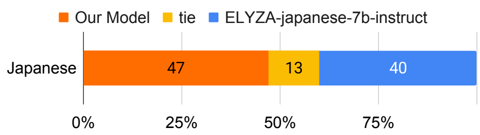
Table 2: This table compares compute matched 7B and 70B checkpoints. We look at intermediate checkpoint results and compare 7B models trained for 40B tokens with 70B models trained for 4B tokens.



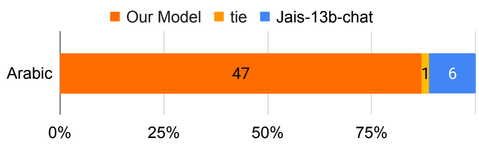
(a) SambaLingo vs Aya101



(b) SambaLingo vs BloomChat-v1



(c) SambaLingo vs ELEYZA-7b-instruct



(d) SambaLingo vs Jais-13b-chat

Figure 3: GPT4 evaluation result

in table 2. The compute-matched 70B checkpoints are trained for 10x fewer steps (4B tokens vs 40B tokens) and perform as well as or better than 7B variants trained over 40B tokens in every benchmark across Arabic and Hungarian.

4.3 Evaluating Human Aligned Checkpoints

4.3.1 GPT-4 as a Judge

To test our human aligned models' ability to generate high quality responses to real user prompts, we use GPT-4 (OpenAI and et al, 2024) as a judge. This method was first introduced by Zheng et al. (2023) to evaluate English models, and then used by Üstün et al. (2024) as a method to evaluate multilingual models. The resulting model generations are shuffled and fit to the prompting style suggested by (Zheng et al., 2023) before being fed to GPT-4. See Appendix D for the manually collected prompts and section 4.3.2 for the evaluation results.

GPT-4 as a judge has been widely accepted by the community as a way to evaluate chat models (Zheng et al., 2023; Verga et al., 2024), and we extend this to multilingual models. To ensure that GPT-4 is understanding the multilingual text we have native speakers read through a few examples of GPT-4 explaining its decision making process. The native speakers unanimously agree that GPT-4 clearly understands the content in other languages. In appendix D.2 we include example model generations along with GPT-4's corresponding preferences and explanations. Further work is needed to do a large scale study to see how GPT-4 preferences align with human preferences in other languages.

4.3.2 Qualitative Results

Measuring win-rate using GPT-4 as a judge only works in scenarios where a human aligned or instruction tuned model is available in a language. Given this constraint, we were only able to find relevant comparisons for Arabic, Japanese and Turkish, and do not have qualitative evaluations for our models in the other 6 languages. We do not compare to llama2-chat because we found that Llama2-chat and other open source English foundation chat models reply in English when prompted in the target language, instead of replying back in the target language. The results of our evaluation are shown in Figure 3. Our SambaLingo models consistently outperform other models in the same language. For details about the native speaker-curated prompts, see Appendix D. We additionally run evaluations with Claude Opus (Anthropic, 2024) as a judge to ensure that there is no bias by GPT-4 and find very similar results in appendix D.1

Added Tokens	Hungarian	Russian	Turkish	Bulgarian	Arabic	Japanese	Thai
0	2.70	2.28	3.28	2.36	4.23	2.07	4.84
1000	2.52	2.25	2.56	2.19	2.11	1.75	2.10
4000	2.14	2.05	2.20	1.92	1.67	1.23	1.50
25000	1.78	1.78	1.77	1.66	1.26	0.93	1.10

Table 3: Number of added tokens vs fertility (average number of tokens per "word")

Language	Tokenizer	ppl (↓)	FLORES EN→X (↑)	FLORES X→EN (↑)	Belebele (↑)	SIB-200 (↑)	XNLI (↑)	XStoryCloze (↑)
Arabic	Original	1.50	48.27	57.35	0.27	0.27	**0.34**	0.63
	Expanded	**1.46**	**52.66**	**61.05**	**0.32**	**0.35**	**0.34**	**0.64**
Hungarian	Original	**1.61**	**52.70**	**58.31**	**0.33**	0.26	-	-
	Expanded	1.63	51.82	57.12	0.30	**0.34**	-	-
Serbian	Original	**1.403**	56.15	64.89	0.32	**0.59**	-	-
	Expanded	1.435	**58.30**	**66.35**	**0.37**	0.52	-	-

Table 4: Accuracy after training with expanded vocabulary vs original tokenizer

5 Ablations

In this section, we present ablations of our design decisions in Section 3. Section 5.1 presents experiments motivating the modifications we make to the base model's tokenizer and how we initialize its new embeddings. Section 5.2 ablates the amount of target language data and use of machine translated data in the DPO phase of our methodology. Finally, section 5.3 looks at the impact of the quality of the base model.

5.1 Vocabulary Expansion

The Llama2 tokenizer is centered towards English. While this tokenizer can encode characters in any language, it will be very inefficient for non-English text. In fact, the BPE tokenizer may tokenize non-Latin characters as multiple independent bytes. One way to mitigate this problem is to extend the vocabulary of the base model by adding new tokens that represent the target language to it, and start adaptation training with this expanded vocabulary. This method also helps improve the inference efficiency in the target language. We explore different sizes for the expanded vocabulary and their impacts on fertility (Ács, 2019) in Table 3 and Figure 12. We chose to expand the vocabulary by 25,000 tokens for all languages as it yields the lowest fertility for all languages and highest throughput on the hardware platform.

5.1.1 Vocabulary Expansion vs Original Tokenizer

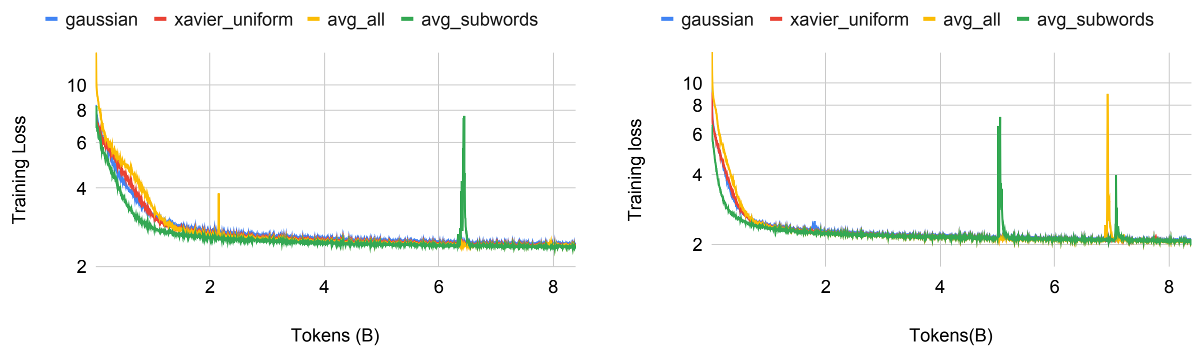
To measure the impact of vocabulary expansion on accuracy, we train two models—one using an expanded vocabulary and the other using the original vocabulary—across two three languages: Hungarian, Arabic and Serbian. We find that expanding the vocabulary does not have significant impact on the downstream accuracy. Nonetheless, given the benefit that the expanded vocabulary has for inference and sequence length utilization in the target language, we chose to expand the vocabulary of the base model.

5.1.2 Initializing new token embeddings

We experiment with 4 different token initialization strategies for the new tokens added to the vocabulary across 3 languages - Hungarian Arabic and Thai. For each experiment, we train the model for 10 billion tokens and compare the loss values. Let V be the set of tokens in the original vocabulary, and $E(t)$ the embedding vector of a token $t \in V$. The four token initialization methods we consider are as follows:

- `gaussian`: $\mathcal{N}(0, 0.02)$
- `xavier_uniform`
- `avg_all` (Hewitt, 2021): For each new token t', initialize $E(t') = \text{mean}(\{E(t) \forall t \in V\})$
- `avg_subwords` (Liu et al., 2024; Koto et al., 2021): For each new token t', let $L_{t'} = [t_1, ..., t_k]$ be the list of k tokens that t' would have been tokenized as under the original tokenizer. Initialize the new embedding with $E(t') = \text{mean}([E(t_1), ..., E(t_k)])$.

Figure 4 shows that after continuous pretraining for 10B tokens, all methods converge to similar loss values, with `avg_subwords` showing faster convergence. Table 5 shows the impact on downstream



(a) Arabic

(b) Hungarian

Figure 4: Training loss for different token initialization methods

Language	Initialization Method	ppl (↓)	FLORES EN→X (↑)	FLORES X→EN (↑)	Belebele (↑)	SIB-200 (↑)	XNLI (↑)	XStoryCloze (↑)
Arabic	`gaussian`	1.50	48.48	57.31	0.34	0.25	0.34	0.61
	`xavier_uniform`	1.49	50.46	58.90	0.36	0.26	0.33	0.62
	`avg_all`	**1.48**	50.54	58.29	0.34	0.25	**0.35**	0.63
	`avg_subwords`	**1.48**	**50.87**	**59.62**	**0.38**	**0.27**	0.34	**0.64**
Hungarian	`gaussian`	**1.65**	**51.42**	**56.92**	0.32	**0.50**	-	-
	`xavier_uniform`	**1.65**	49.52	55.81	**0.34**	0.42	-	-
	`avg_all`	1.76	51.39	56.86	**0.34**	0.45	-	-
	`avg_subwords`	**1.65**	50.79	56.77	0.33	0.30	-	-
Thai	`gaussian`	1.31	51.50	52.95	0.33	0.53	0.44	-
	`xavier_uniform`	1.31	52.88	55.34	0.32	0.30	0.38	-
	`avg_all`	1.31	52.89	55.36	0.35	**0.60**	**0.46**	-
	`avg_subwords`	**1.30**	**53.34**	**55.36**	**0.37**	0.35	**0.46**	-

Table 5: Multilingual evaluations across token embedding initialization methods

benchmarks. For Thai and Arabic, `avg_subwords` achieves marginally better scores while for Hungarian the results are mixed. These results show that the choice of initialization has minimal impact on the accuracy of end model when trained for 10 billion tokens. However `avg_subwords` gives faster training loss convergence, so we chose to initialize the new embeddings using `avg_subwords`.

5.2 Direct Preference Optimization

5.2.1 DPO Data Mixture

There is a lack of supervised finetuning and human alignment data across different languages. Collecting such data can be difficult and expensive. Given that the models obtained from our methodology are bilingual, we explore the question of how much of the human alignment data can be English and how much of it has to be from the target language. We run DPO on data mixtures of the English/Target language data ratio across 100:1, 10:1, 10:3 and 1:1, and observe the resulting win-rate in pairwise comparisons with the model trained on a 1:1 data ratio. For each experiment we keep the amount of English data the same and downscale the target language. We run these experiments on two languages: Hungarian and Arabic, with results in Table 6. We show that a 10:1 data ratio can perform almost as well as 1:1 data ratio for Hungarian. For Arabic, even a 10:3 data ratio still falls behind the performance of 1:1. One hypothesis is that Hungarian is more linguistically similar to English than Hungarian so there is more language transfer during fine tuning, but further research is needed to understand how the language impacts optimal alignment data mixture ratio.

5.2.2 Impact of Translated Human Preference Data

Results in Table 6 are based on translated data from the target language. Üstün et al. (2024) emphasized the importance of human written prompt completion pairs and claim that translated data does not perform as well. However, their work does not start with a high quality pretrained base model, nor do they use DPO. In order to understand whether machine translated data is a viable option for human alignment, we explore the impact of alignment using both approaches. We use Google translated `ultrafeedback-200k` data for one run and human-written data from Open Assistant Conversations (OASST1) (Köpf et al., 2023) for the other. We run this study on Russian, as it is has the most hu-

Target Language: English Ratio	100:1	10:1	10:3	1:1
Arabic	30.39%	35.00%	34.62%	50.00%
Hungarian	39.29%	45.18%	45.78%	50.00%

Table 6: DPO data mixture result (win-rate compared with 1:1 data mixture)

Base Model	**ppl**(↓)	**FLORES EN→X**(↑)	**FLORES X→en**(↑)	**Belebele**(↑)	**SIB-200**(↑)
GPT-13B	1.80	37.94	48.99	0.28	**0.25**
Llama-2-7b	**1.61**	**53.72**	**58.65**	**0.34**	**0.25**

Table 7: Performance of GPT-13B and Llama 2 7B on Hungarian benchmarks after adaptation

man written data from OASST1 (Köpf et al., 2023). The model trained using translated data attains a 50.47% win rate compared to the model trained with OASST1. This comparison does not control for the diversity and quality of the question answer pairs in the dataset because chat datasets with parallel human translated data in multiple languages. so this comparison is not meant to illustrate that translated data is as good or better than native data, but rather to show that human written data is not a silver bullet required to obtain good quality aligned models in other languages.

5.3 Importance Of Base Model Quality

To explore the relationship between the quality of the base model employed for language adaptation and its subsequent impact on accuracy in the target language, we ablate using two different base models - Llama 2 7B and GPT-13B (Srinivasan et al., 2023). The GPT-13B model is trained on much fewer tokens compared to llama2. We measure the GPT-13B model on some commonly accepted English benchmarks instead of our multilingual evaluation suite because these benchmarks are used more frequently to compare English checkpoints. GPT-13B lags behind Llama 2 7B in every English evaluation tasks we measured in Table 9.

We adapt both of these models to Hungarian. Table 7 illustrates that using a higher quality base model (Llama 2 7B) leads to better downstream performance in the target language. These results show that many of the benefits of training come from the base model quality not just the continuous training we do. This additionally indicates that as newer higher quality models become available, there is value in applying our proposed adaptation methodology on new base models.

6 Limitations

Our work has several limitations, including the need for extensive data from the target language, which is often unavailable for many languages. We study 9 diverse languages, but further research is required to address multilingual data scarcity and generalize our recipe. Due to compute and time constraints, our ablation studies focus on around 3 languages each, assuming similar results for other languages, although linguistic diversity and data availability may affect this. Additionally, we evaluate our chat-based model using GPT-4 as a judge, and while this has been shown to strongly correlate with human preferences in English, we are uncertain how well this works in other languages. We acknowledge that publicly releasing LLMs is risky because they can inadvertently generate harmful or biased content, compromise privacy, and be exploited for malicious purposes such as spreading misinformation. Moreover, while our models are adapted to other languages and cultures, the English base model, data biases, and use of translation may prevent them from fully capturing the nuances of cultures and languages from around the world.

7 Conclusion

We present a methodology to adapt pretrained LLMs to new languages. The methodology encompasses both continuous pretraining and alignment to human preferences in the target language. We present experimental results to justify our design choices and scale our methodology to 9 typologically diverse languages and 2 parameter scales. We make our evaluation scripts and final checkpoints publically available to facilitate future research, and we hope this work outlines a clearer path towards attaining state of the art language models in every language.

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *Preprint*, arXiv:2401.10440.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. *Preprint*, arXiv:2311.09205.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *Preprint*, arXiv:2304.09151.

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. *Preprint*, arXiv:1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *Preprint*, arXiv:2311.05741.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023a. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *Preprint*, arXiv:2106.02124.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Ilya Gusev. 2023. Saiga 7b.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models.

Edward J Hu, Yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shengyi Huang, Lewis Tunstall, Edward Beeching, Leandro von Werra, Omar Sanseviero, Kashif Rasul, and Thomas Wolf. 2024. Constitutional ai recipe. *Hugging Face Blog*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization. *Preprint*, arXiv:2109.04607.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, Bonita Bhaskaran, Bryan Catanzaro, Arjun Chaudhuri, Sharon Clay, Bill Dally, Laura Dang, Parikshit Deshpande, Siddhanth Dhodhi, Sameer Halepete, Eric Hill, Jiashang Hu, Sumit Jain, Ankit Jindal, Brucek Khailany, George Kokai, Kishor Kunal, Xiaowei Li, Charley Lind, Hao Liu, Stuart Oberman, Sujeet Omar, Sreedhar Pratty, Jonathan Raiman, Ambar Sarkar, Zhengjiang Shao, Hanfei Sun, Pratik P Suthar, Varun Tej, Walker Turner, Kaizhe Xu, and Haoxing Ren. 2024. Chipnemo: Domain-adapted llms for chip design. *Preprint*, arXiv:2311.00176.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *Preprint*, arXiv:2305.16264.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. *Sabiá: Portuguese Large Language Models*, page 226–240. Springer Nature Switzerland.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Together Computer SambaNova Systems. 2023. BLOOMChat: a New Open Multilingual Chat LLM.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. 2023. Elyza-japanese-llama-2-7b.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. mgpt: Few-shot learners go multilingual. *Preprint*, arXiv:2204.07580.

Venkat Srinivasan, Darshan Gandhi, Urmish Thakker, and Raghu Prabhakar. 2023. Training large language models efficiently with sparsity and dataflow. *Preprint*, arXiv:2304.05511.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

SambaNova Systems. 2023a. x-self-instruct-seed-32.

SambaNova Systems. 2023b. xoa22.

Mikhail Tikhomirov and Daniil Chernyshev. 2023. Impact of tokenization on llama russian adaptation. *Preprint*, arXiv:2312.02598.

TokyoTech. 2023. Swallow 7b.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

BigScience Workshop, :, and Teven Le Scao et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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. *Preprint*, arXiv:2010.11934.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zijian Győző Yang, László János Laki, Tamás Váradi, and Gábor Prószéky. 2023. Mono- and multilingual gpt-3 models for hungarian. In *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, pages 94–104, Plzeň, Czech Republic. Springer Nature Switzerland.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *Preprint*, arXiv:2306.06688.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. Bloom+1: Adding language support to bloom for zero-shot prompting. *Preprint*, arXiv:2212.09535.

Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, Fanchao Qi, Xiaozhi Wang, Yanan Zheng, Guoyang Zeng, Huanqi Cao, Shengqi Chen, Daixuan Li, Zhenbo Sun, Zhiyuan Liu, Minlie Huang, Wentao Han, Jie Tang, Juanzi Li, Xiaoyan Zhu, and Maosong Sun. 2020. Cpm: A large-scale generative chinese pre-trained language model. *Preprint*, arXiv:2012.00413.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024a. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

Judit Ács. 2019. Exploring bert's vocabulary.

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. *Preprint*, arXiv:2402.07827.

A Hyperparameters

- **Continuous Pre-training**: We pack the pre-training mixture into sequences of length 4096 and pretrain with *document attention* as described in Section 3.2 of Iyer et al. (2022) to ensure we only attend to tokens in the context of the corresponding text document. We train with a global batch size of 1024, sequence length of 4096, maximum learning rate of 1e-4 with cosine decay, warm-up ratio of 0.01 and a weight decay of 0.1. Each expert is trained for a maximum of 4 epochs, following (Muennighoff et al., 2023). Notably, we train all model parameters, foregoing use of PEFT methods such as LoRA (Hu et al., 2022), which are known to be inferior to full parameter training (Zhao et al., 2024a)(Sun et al., 2023).
- **Supervised Finetuning**: We use a global batch size of 512 and a maximum sequence length of 2048 tokens. We used a linear decay learning rate of 2e-5 with 10% warm up
- **Direct Preference Optimization**: We train with a global batch size 32 for 3 epochs, a linear decay learning rate of 5e-7, 10% warmup and $\beta = 0.1$ as the regularization factor for DPO

B Language Experts vs Monolith Multilingual Model

"The Curse Of Multilinguality" (Chang et al., 2023; Conneau et al., 2020) is the idea that LLMs have a fixed capacity with which to learn various languages. This theory claims that as one expands the number of languages a model is trained on, the various languages compete for the capacity of the model, therefore degrading the models performance across all languages. Blevins et al. (2024) attempt to address this phenomenon by adapting multiple small-scale language experts from XGLM-1.7B (Lin et al., 2022), one for each language, and show that each expert outperforms training a single monolithic model trained simultaneously on one language. We build on these results by scaling this study to 7B parameters and use more comprehensive evaluation metrics than just perplexity. We compare our 9 Llama 2 7B language experts against a monolith Llama 2 7B model continuously pretrained on all 9 languages. We ensure that each language is represented equally in the monolith's training data and the vocabulary is expanded to represent all 9 languages evenly.

For comparison's sake, we select intermediate model checkpoints such that each individual language expert has used the same amount of compute as the monolith multilingual model. This means that the experts required 9x more compute to train then the monolith. Table 8 averages the evaluation results across all 9 languages and finds that the monolith model and language experts have very similar performance. This implies that if one wants to adapt to many languages at once, it may be more compute-efficient to continuously train a multi-linugal model rather then independent experts. Further work is warranted to determine how this result scales with an increasing number of target languages.

Benchmark (Num Shots)	Llama2-7b Avg	Multilingual Monolith Avg	Language Expert Avg
↓ Holdout PPL	1.75	1.55	**1.50**
↑ FLORES X>en(8)	40.42%	50.69%	**51.71%**
↑ Belebele (3)	**36.24%**	33.36%	32.09%
↑ SIB-200(3)	26.67%	**38.04%**	33.43%
↑ XNLI (0)	39.00%	**43.44%**	43.04%
↑ XStoryCloze (0)	56.35%	65.75%	**68.03%**
↑ XWinograd (0)	69.48%	**72.39%**	71.97%
↑ PAWS-X (0)	51.00%	**54.40%**	53.50%
↑ MGSM (3)	**5.40%**	4.00%	4.20%

Table 8: Monolith multilingual continuous training vs language experts, averaged over all 9 languages.

C Base Model English Evaluation

	HellaSwag(↑)	**OpenBookQA**(↑)	**ARC-E**(↑)	**ARC-C**(↑)	**PiQA**(↑)	**Winogrande**(↑)
GPT-13B	0.60	0.36	0.53	0.30	0.76	0.60
Llama-2-7B	**0.76**	**0.57**	**0.73**	**0.48**	**0.80**	**0.70**

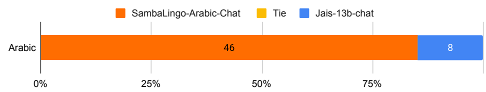
Table 9: Performance of GPT-13B and Llama-2-7B on English NLU benchmarks

D Qualitative Results

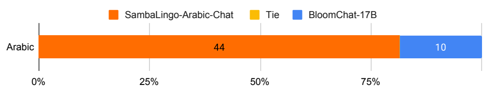
For Arabic, we compare our 7B arabic expert with aya-101 (Üstün et al., 2024), Jais-13b-chat (Sengupta et al., 2023), and Bloomchat-v1 (SambaNova Systems, 2023) and use prompts from x-self-instruct-seed-32 (Systems, 2023a) and xOA22 (Systems, 2023b). Our Arabic chat model reaches 87.96% win rate compared to Jais-13B-chat, 99.06% win rate compared to Aya101, and 68.52% compared to Bloomchat-v1. For Japanese, we compare our Japanese chat model with ELYZA-japanese-Llama-2-7b-instruct (Sasaki et al., 2023) on 100 randomly sampled prompts aya dataset (Üstün et al., 2024), reaching a win rate of 53.5% For Turkish, we compare our Turkish chat model against aya-101 (Üstün et al., 2024) using prompts from the test set of aya dataset (Üstün et al., 2024), leading to win-rate of 92.4%.

D.1 Evaluating Chat Models With Claude

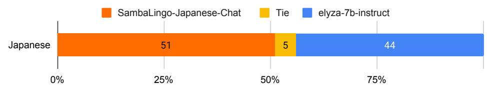
We run evaluations using the same prompt as GPT-4 as a judge, but use Claude Opus (Anthropic, 2024) as a judge. Figure 5 shows the evaluations to be in line with our previous results with GPT-4 as a judge 3. This shows that there is no strong bias in using GPT-4 as a judge.



(a) SambaLingo vs Jais-13b-chat



(b) SambaLingo vs BloomChat-v1



(c) SambaLingo vs ELEYZA-7b-instruct

Figure 5: Claude Opus evalution results

D.2 GPT-4 As A Judge

Below are some examples of how GPT-4 judged two candidate responses in Japanese, Arabic and Turkish. See figures 6, 7, 8, 9, 10, 11

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
１８６９年に開通した世界最長の運河の名前は何でしょう?	１８６９年に開通した世界最長の運河の名前は、「グランド・キャニオン・アンド・サンタフェ・アンド・トレイル・パス運河」です。 この運河はアメリカ合衆国のアリゾナ州とニューメキシコ州を結ぶ全長約435kmの運河です。 ちなみに、現在世界最長の運河は、アフリカ大陸とアジア大陸を結ぶ「スエズ運河」です。	世界最長の運河は、1869 年に開通したスエズ運河です。	Assistant A's response is incorrect. The canal opened in 1869 is the Suez Canal, not the "Grand Canyon and Santa Fe and Trail Pass Canal" as stated by Assistant A. Furthermore, the Suez Canal is not only the longest canal opened in 1869, but it is also currently the longest canal in the world. Assistant B correctly identifies the Suez Canal as the longest canal opened in 1869. Therefore, Assistant B's response is more accurate and relevant to the user's question.

Figure 6: Japanese Example 1

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
俳句は全部で１７文字からなりますが、短歌は全部で何文字からなるでしょう?	短歌は通常、俳句よりも長く、全編で少なくとも 31 文字あります。	短歌は、全部で５文字からなる「五七五」が基本です。	Assistant A's response is correct. A Tanka (短歌) in Japanese poetry is composed of 31 syllables (or "on" in Japanese), not 5 as Assistant B incorrectly states. Assistant B seems to be confusing Tanka with Haiku, which is composed of 17 syllables and is often structured as 5-7-5. Therefore, Assistant A provides a more accurate and relevant answer to the user's question.

Figure 7: Japanese Example 2

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
من 1 إلى 10 ما هو تأثير Novalgin على تخفيف الآلام إعتماداً على الإحصائيات المتاحة	لا توجد إحصائيات دقيقة لتقييم مدى تم شركة Google Inc.، المعروفة ببساطة باسم \"Google\"، هي شركة أمريكية متعددة الجنسيات متخصصة في مجال الإعلان المرتبط بخدمات الإنترنت والبحث. تأسست الشركة في 4 سبتمبر 1998 من قبل Larry Page و Sergey Brin ، وهما طالبا الدكتوراه في جامعة ستانفورد . يقع مقرها الرئيسي في Menlo Park ، كاليفورنيا ، الولايات المتحدة الأمريكية . دأت Google كشركة خاصة ، وتم طرح أسهمها للاكتتاب العام في 19 أغسطس 2004. اعتبارًا من 3 يناير 2023 ، تعد Alphabet Inc. الشركة الأم لشركة Google. في غضون بضع سنوات من تأسيسه ، أصبح Google أحد أكثر الأسماء شهرة وشعبية بين مستخدمي الإنترنت حول العالم. بحلول عام 2021 ، احتفظ Google بأكثر من 95٪ من حصة السوق العالمية لمحركات البحث ، مما جعله المهيمن بلا منازع في هذا المجال.	إنشاء Google في عام 1998 على يد سيرجي برين ولاري بيج عندما كانا طالبين في جامعة ستانفورد. لقد بدأوا كمشروع بحثي في جامعة ستانفورد، وسرعان ما أدركوا إمكانات محرك البحث الذي قاموا بإنشائه. في عام 1998، قاموا بتسجيل Google كشركة، وفي عام 1999، تم إطلاق Google كمحرك بحث عام. كان أحد العوامل الرئيسية التي ساهمت في هيمنة Google على محركات البحث هو تركيزها على تزويد المستخدمين بنتائج بحث عالية الجودة وذات صلة. لقد استثمروا بكثافة في تطوير خوارزميات البحث الخاصة بهم، والتي تم تصميمها لتحديد مواقع الويب الأكثر صلة واستجابة لاستعلامات المستخدم. بالإضافة إلى ذلك، قدمت Google ميزات مثل التصفية الجغرافية، واقتراحات البحث التلقائية، ونتائج البحث الفوري، والتي جعلت من السهل على المستخدمين العثور على المعلومات التي يحتاجون إليها بسرعة. هناك عامل آخر ساهم في نجاح Google وهو تركيزها على تجربة المستخدم. لقد أدركوا أن المستخدمين يريدون نتائج بحث سريعة ودقيقة وسهلة الاستخدام، وقد استثمروا في إنشاء واجهة سهلة الاستخدام وبديهية. لقد قاموا أيضًا بدمج العديد من الميزات التي جعلت من السهل على المستخدمين التنقل في نتائج البحث الخاصة بهم، مثل شريط البحث، ونتائج البحث المنسدلة، ونتائج البحث المنسدلة. بشكل عام، كان الجمع بين خوارزميات البحث عالية الجودة، وتجربة المستخدم الرائعة، والتركيز على الابتكار هو الذي ساعد Google على أن تصبح محرك البحث الأكثر شعبية في العالم.	Both assistants provided relevant and accurate information about the effectiveness of Novalgin in pain relief. However, Assistant A provided a more comprehensive response by mentioning a specific study published in a pain research journal in 2016, and also discussed the factors that might affect the effectiveness of Novalgin. Assistant B also mentioned some studies, but the information provided was less detailed and less specific compared to Assistant A. Therefore, Assistant A's response is more helpful and detailed. Final Verdict: [[A]]

Figure 8: Arabic Example 1

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
من 1 إلى 10 ما هو تأثير Novalgin على تخفيف الآلام إعتماداً على الإحصائيات المتاحة	لا توجد إحصائيات دقيقة لتقييم مدى فعالية النوفالجين في تسكين الألم. ومع ذلك، فإن بعض الدراسات تشير إلى أن النوفالجين قد يكون فعالًا مثل مضادات الالتهاب غير الستيرويدية (NSAIDs) الأخرى في علاج آلام الأسنان والتهاب المفاصل. وقد وجدت دراسة صغيرة عام 2013 أن النوفالجين كان أكثر فعالية من دواء وهمي في تقليل ألم عصب الضرس بعد خلعه. ووجدت دراسة أخرى عام 2014 أن النوفالجين كان له نفس الفعالية مثل الإيبوبروفين في علاج التهاب مفصل الفك الصدغي. ومع ذلك، هناك حاجة لمزيد من الأبحاث لتحديد مدى فعالية النوفالجين في تسكين الألم بشكل دقيق.	لا يمكنني الوصول إلى أحدث الإحصائيات، ولكن وفقًا لدراسة نشرت في مجلة أبحاث الألم في عام 2016، فإن نوفالاجين فعال في تخفيف الألم المعتدل إلى الشديد لدى المرضى الذين يعانون من آلام ما بعد الجراحة. وجدت الدراسة أن نوفالاجين كان أكثر فعالية من العلاج الوهمي في تقليل شدة الألم ومدته. ومع ذلك، قد تختلف فعالية نوفالاجين في تخفيف الألم اعتمادًا على عوامل مثل شدة الألم، ومدة الألم، والحالة الصحية العامة للمريض. من الأفضل دائمًا استشارة أخصائي الرعاية الصحية لتحديد أفضل خيارات إدارة الألم بناءً على احتياجات الفرد الخاصة.	Both assistants provided relevant and accurate information about the effectiveness of Novalgin in pain relief. However, Assistant A provided a more comprehensive response by mentioning a specific study published in a pain research journal in 2016, and also discussed the factors that might affect the effectiveness of Novalgin. Assistant B also mentioned some studies, but the information provided was less detailed and less specific compared to Assistant A. Therefore, Assistant A's response is more helpful and detailed. Final Verdict: [[A]]

Figure 9: Arabic Example 2

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
Cilt bakımında buhar kaç dakika verilir?	ika arasında yapılır.	Cilt bakımı rutinine buhar uygulaması genellikle 5-10 dakika kadar kısa bir süre için yapılır. Buhar, gözeneklerin açılmasına ve cildin nemlendirilmesine yardımcı olur, bu da onu cilt bakımı rutininin önemli bir parçası haline getirir. Ancak cilt tipine ve kişisel tercihe bağlı olarak buhar süresi ayarlanabilir. Buhar uygulaması sırasında ciltte herhangi bir rahatsızlık veya tahriş hissedilirse, buhar uygulaması durdurulmalı ve bir dermatoloğa veya cilt bakımı uzmanına danışılmalıdır.	Assistant A's response is incomplete and does not provide any useful information to the user's question. On the other hand, Assistant B provides a detailed and informative response, explaining that steam is typically applied for 5-10 minutes during a skincare routine, the benefits of steam, and advice on what to do if discomfort or irritation is experienced. Therefore, Assistant B's response is more helpful, relevant, and accurate. My final verdict is: [[B]]

Figure 10: Turkish Example 1

Prompt	Model 1 Completion	Model 2 Completion	GPT4 Judge
Sütçü İmam Üniversitesi ne zaman kuruldu?	li ilçesinde, Sivas Valisi Süleyman Demirel'in talimatıyla, Sivas'ın tanınmış tarihi ve kültürel değerlerine sahip, Sivas'ın ilk ve tek devlet üniversitesi olarak kurulmuştur.	Kahramanmaraş Sütçü İmam University (KSÜ) was founded in 1986 and is located in the city of Kahramanmaraş in southeastern Turkey.	Assistant A provides a direct and accurate answer to the user's question, stating that Sütçü İmam University was established in 1986 and is located in Kahramanmaraş in southeastern Turkey. On the other hand, Assistant B's response is irrelevant and confusing, as it talks about a university in Sivas and mentions Süleyman Demirel, who was not related to the establishment of Sütçü İmam University. Therefore, Assistant A's response is more helpful, relevant, and accurate. Final Verdict: [[A]]

Figure 11: Turkish Example 2

E Expanded Vocabulary Tokenizer Fertility

E.1 Expanded Vocabulary Tokenizer Fertility

In figure 12 We measure the fertility of the tokenizer as we expand the vocabulary, and see that we can improve the fertility from about 4.8 to 1.1 on Thai. This is about a 4.35x improvement, implies that inference speeds can improve up to 4.35x compared to the Llama2 tokenizer.

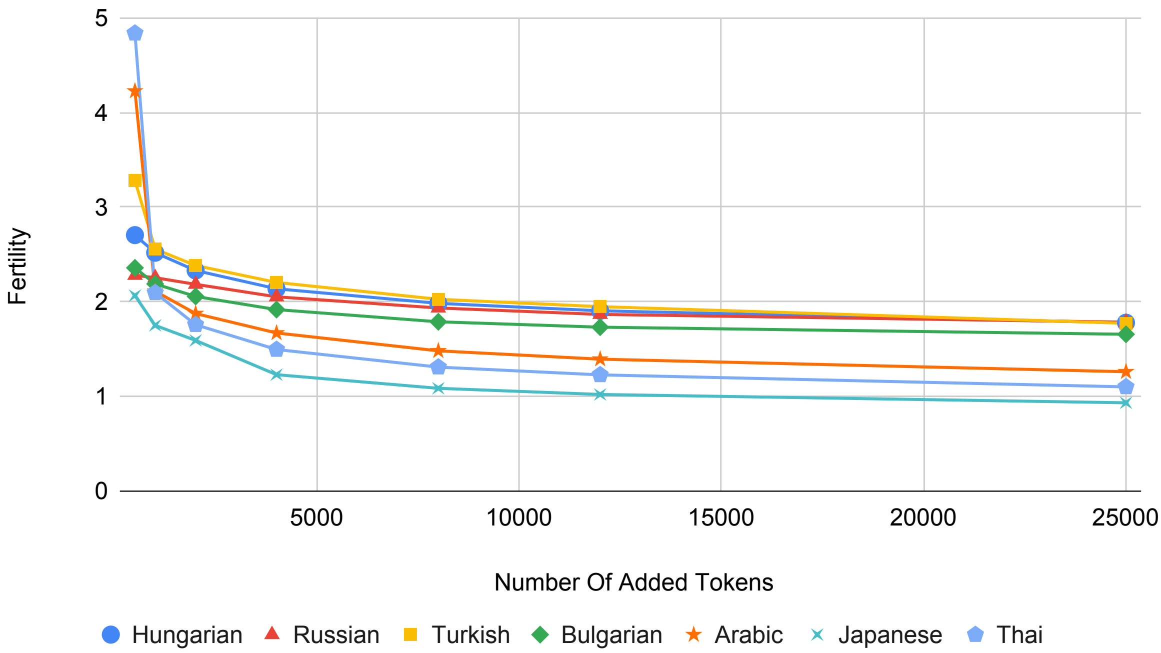


Figure 12: Tokenizer Fertility: the average number of tokens per "word" (Ács, 2019)

F Main Results Details

See tables 13 and 14 for all evaluation results

Langauge	Checkpoint	Holdout PPL	MC4 PPL	Wikipedia PPL	FLORES EN->X 8 shot CHRF	FLORES X->en 8 shot CHRF	FLORES EN->X 8 shot BLEU	FLORES X->en 8 shot BLEU	Belebele 3 shot	SIB-200 3 shot	Exams 3 shot	XNLI 0 shot	XStory Cloze 0 shot	XWinograd 0 shot	PAWS-X 0 shot	XCOPA 0 shot	MGSM 3 shot	Average
Russian	Our-Russian-7b	1.44	1.35	1.30	54.90	61.13	27.67	35.07	0.35	0.43		0.36	0.72	0.69			0.09	0.47
Russian	Our Monolith	1.48	1.37	1.35	53.45	59.72	25.57	31.93	0.34	0.44		0.50	0.69	0.67			0.05	0.48
Russian	Saiga-7b	1.56	1.73	2.89	52.95	61.70	24.28	34.09	0.75	0.63		0.50	0.69	0.67			0.33	0.59
Russian	Llama2-7b	1.53	1.49	1.36	49.59	59.55	21.83	33.70	0.42	0.26		0.41	0.63	0.69			0.08	0.45
Russian	bloom-7b1	1.51	1.40	1.38	28.22	44.75	5.60	18.07	0.27	0.25		0.43	0.53	0.57			0.02	0.35
Russian	xglm-7.5B	1.80	1.79	1.81	32.62	50.64	4.33	17.48	0.25	0.26		0.46	0.63	0.63			0.02	0.39
Russian	AYA-101				0.45	1.04	0.00	0.00	0.23	0.28		0.40	0.60	0.63			0.00	0.27
Arabic	Our-Arabic-70B	1.39			57.67	66.99	26.05	41.21	0.74	0.73		0.35	0.69					0.63
Arabic	Our-Arabic-Base	1.42	1.40	1.38	54.11	63.09	22.71	38.66	0.34	0.29		0.34	0.66					0.47
Arabic	Our Monolith	1.52	1.49	1.49	49.39	58.39	69.82	54.70	0.31	0.36		0.33	0.62					0.45
Arabic	Jais-13b	1.50	1.44	1.39	49.39	58.39	18.92	33.35	0.31	0.36		0.33	0.62					0.45
Arabic	Llama2-7b	1.81	1.93	1.89	28.06	44.96	3.40	21.12	0.31	0.26		0.35	0.50					0.36
Arabic	bloomz-7	1.57	1.57	1.52	44.11	54.88	14.60	29.37	0.28	0.25		0.34	0.59					0.41
Arabic	xglm-7.5	1.60	1.60	1.59	18.87	49.39	1.17	22.80	0.25	0.26		0.34	0.56					0.35
Arabic	AYA-101				0.37	0.88	0.00	0.00	0.23	0.28		0.33	0.56					0.24
Japanese	Our-Japanese-7B	1.56	1.54	1.59	36.64	52.10	2.10	23.43	0.28	0.26				0.77	0.54		0.02	0.39
Japanese	Our Monolith	1.66	1.56	1.65	34.97	49.88	0.77	20.88	0.33	0.40				0.78	0.54		0.03	0.42
Japanese	ELYZA-japanese-Llama	1.75	1.76	1.72	16.88	43.94	0.01	15.24	0.34	0.43				0.77	0.49		0.02	0.38
Japanese	Swallow	1.65	1.77	1.75	39.58	50.87	1.12	22.16	0.39	0.27				0.81	0.44		0.10	0.42
Japanese	Llama2-7b	1.90	2.03	1.67	30.44	48.92	1.30	20.93	0.39	0.29				0.70	0.51		0.02	0.39
Japanese	bloom-7b1	2.26	2.44	2.20	14.25	39.44	0.23	12.17	0.26	0.25				0.59	0.55		0.04	0.32
Japanese	xglm-7.5B	1.80	1.72	1.80	19.68	30.04	0.01	4.58	0.27	0.25				0.65	0.48		0.00	0.31
Japanese	AYA-101				0.32	0.94	0.00	0.00	0.23	0.29				0.63	0.56		0.00	0.25
Thai	Our-Thai-70B	1.24			55.41	62.45	11.92	36.13	0.74	0.77		0.44				0.64	0.24	0.57
Thai	Our-Thai-7B	1.29	1.23	1.22	54.21	56.07	11.64	29.94	0.38	0.65		0.45				0.61	0.08	0.47
Thai	Our Monolith	1.35	1.28	1.26	51.79	54.17	11.35	27.02	0.31	0.40		0.48				0.60	0.06	0.42
Thai	typhoon-7b	1.37	1.41	1.40	50.16	26.28	5.95	6.10	0.60	0.30		0.42				0.61	0.27	0.42
Thai	Llama2-7b	1.59	1.55	1.55	19.85	29.78	2.24	8.80	0.31	0.25		0.36				0.56	0.04	0.29
Thai	bloom-7b1	1.83	1.78	1.83	13.56	18.92	0.13	1.33	0.27	0.25		0.34				0.55	0.02	0.25
Thai	xglm-7.5B	1.40	1.30	1.27	30.24	24.10	0.50	5.13	0.24	0.25		0.42				0.59	0.01	0.29
Thai	AYA-101				0.68	0.81	0.00	0.00	0.23	0.26		0.35				0.58	0.00	0.21
Turkish	Our-Turkish-7B	1.56	1.59	1.55	54.22	58.27	20.47	31.95	0.37	0.33		0.45				0.70		0.50
Turkish	Our Monolith	1.63	1.69	1.66	51.61	55.98	77.91	60.55	0.30	0.41		0.49				0.66		0.49
Turkish	TURNA				0.00	0.02	0.00	0.00	0.23	0.25		0.37				0.56		0.24
Turkish	Llama2-7b	2.27	2.44	2.16	27.61	43.03	161.70	76.56	0.32	0.26		0.37				0.55		0.37
Turkish	bloom-7b1	2.95	3.24	3.00	12.75	24.78	351.11	121.72	0.29	0.25		0.35				0.51		0.30
Turkish	xglm-7.5B	1.83	1.91	1.80	28.63	42.99	485.24	113.21	0.25	0.25		0.47				0.58		0.38
Turkish	AYA-101				1.02	1.13	98.59	98.58	0.23	0.31		0.40				0.60		0.26

Figure 13: Main results, evaluation benchmarks described in 4.1.This data is averaged to create 2.

Langauge	Checkpoint	Holdout PPL	MC4 PPL	Wikipedia PPL	FLORES EN->X 8 shot CHRF	FLORES X->en 8 shot CHRF	FLORES EN->X 8 shot BLEU	FLORES X->en 8 shot BLEU	Belebele 3 shot	SIB-200 3 shot	Exams 3 shot	XNLI 0 shot	XStory Cloze 0 shot	XWinograd 0 shot	PAWS-X 0 shot	XCOPA 0 shot	MGSM 3 shot	Average
Bulgarian	Our-Bulgarian-Base	1.42	1.35	1.29	62.94	65.06	36.25	39.31	0.36	0.43	0.52	0.49						0.51
Bulgarian	Our Monolith	1.46	1.44	1.34	61.92	64.50	34.45	38.28	0.35	0.45	0.49	0.47						0.50
Bulgarian	mGPT-1.3B-bulgarian	1.75	1.65	1.53	18.53	24.78	2.89	4.19	0.23	0.25	0.30	0.34						0.26
Bulgarian	Llama2-7b	1.59	1.61	1.39	51.55	62.43	23.12	36.55	0.37	0.27	0.38	0.41						0.43
Bulgarian	bloom-7b1	2.06	2.14	2.10	23.08	35.55	2.20	10.91	0.28	0.25	0.26	0.39						0.29
Bulgarian	xglm-7.5B	1.50	1.48	1.36	45.05	58.48	10.38	28.13	0.23	0.25	0.41	0.45						0.40
Bulgarian	AYA-101				0.60	0.86	0.00	0.00	0.23	0.30	0.33	0.37						0.21
Hungarian	Our-Hungarian-70B	1.52			57.18	63.40	25.89	37.07	0.76	0.56								0.63
Hungarian	Our-Hungarian-7B	1.61	1.63	1.56	53.72	58.65	22.84	31.64	0.34	0.25								0.43
Hungarian	Our Monolith	1.66	1.78	1.67	52.81	57.06	21.00	29.52	0.30	0.35								0.44
Hungarian	PULI-GPTrio	1.72	1.74	1.67	46.33	48.23	15.81	21.05	0.24	0.25								0.36
Hungarian	Llama2-7b	1.95	2.24	1.78	42.47	53.89	13.47	27.26	0.33	0.25								0.39
Hungarian	bloom-7b1	3.02	3.75	3.68	12.55	23.73	0.62	4.15	0.27	0.25								0.22
Hungarian	xglm-7.5B	3.36	4.36	4.17	6.40	11.87	0.15	0.36	0.23	0.27								0.17
Hungarian	AYA-101				0.51	0.87	0.00	0.00	0.23	0.30								0.14
Serbian	Our-Serbian-7B	1.44	1.35	1.26	56.16	64.89	29.03	40.51	0.32	0.59	0.59							0.54
Serbian	Our Monolith	1.45	1.46	1.35	58.53	65.43	31.05	40.19	0.35	0.41	0.37							0.47
Serbian	sr-gpt2				0.15	7.49	0.00	0.03	0.23	0.25								0.14
Serbian	Llama2-7b	1.60	1.67	1.42	49.19	63.98	20.65	39.52	0.39	0.25								0.44
Serbian	bloom-7b1	2.13	2.33	2.21	22.08	32.80	1.68	9.64	0.28	0.25								0.27
Serbian	xglm-7.5B	2.39	2.60	2.42	19.38	31.24	0.47	6.84	0.24	0.27								0.25
Serbian	AYA-101				0.64	1.03	0.00	0.00	0.23	0.29								0.13
Slovenian	Our-Slovenian-7B	1.68	1.69	1.70	54.37	58.60	26.39	32.54	0.35	0.43								0.48
Slovenian	Our Monolith	1.72	1.62	1.80	55.79	60.12	27.59	33.43	0.30	0.51								0.49
Slovenian	sl-gpt2				9.84	8.62	0.06	0.01	0.24	0.15								0.14
Slovenian	Llama2-7b	2.11	2.08	1.86	44.45	57.57	15.77	31.73	0.38	0.36								0.44
Slovenian	bloom-7b1	3.40	3.44	3.66	17.26	30.21	1.25	7.02	0.28	0.25								0.25
Slovenian	xglm-7.5B	4.42	4.66	4.58	10.94	23.84	0.32	2.70	0.24	0.25								0.21
Slovenian	AYA-101				0.93	1.22	0.00	0.00	0.23	0.30								0.14

Figure 14: Main results, evaluation benchmarks described in 4.1.This data is averaged to create 2.