LLMs for Extremely Low-Resource Finno-Ugric Languages

Taido Purason*, **Hele-Andra Kuulmets***, **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{taido.purason, hele-andra.kuulmets, mark.fisel}@ut.ee

Abstract

The advancement of large language models (LLMs) has predominantly focused on high-resource languages, leaving low-resource languages, such as those in the Finno-Ugric family, significantly underrepresented. This paper addresses this gap by focusing on Võro, Livonian, and Komi. We cover almost the entire cycle of LLM creation, from data collection to instruction tuning and evaluation. Our contributions include developing multilingual base and instruction-tuned models; creating evaluation benchmarks, including the SMUGRI-MT-BENCH multi-turn conversational benchmark; and conducting human evaluation. We intend for this work to promote linguistic diversity, ensuring that lesser-resourced languages can benefit from advancements in NLP.

1 Introduction

Large language models (LLMs) have recently demonstrated unprecedented flexibility in responding to unstructured text queries (OpenAI et al., 2024; Touvron et al., 2023, etc). However, their development requires high amounts of training material: while the (continued) pre-training stage only needs raw text, instruction tuning relies on sets of instructions, which are much more expensive to obtain. The challenge is exacerbated for endangered languages, where both the availability of data as well as the number of speakers are severely limited.

We present a case study in developing LLMs for extremely low-resource (XLR) languages, focusing on three Finno-Ugric (SMUGRI[1]) languages: Livonian, Võro and Komi. According to Joshi et al. (2020), Livonian is in the lowest category out of 6 (0 / *The Left-Behinds*) while Võro and Komi are in the second-lowest (1 / *The Scrabing-Bys*). Developing LLMs and other tools for these languages is thus a significant challenge, as well as a vital step for ensuring their digital survival and support.

	Script	Code	Class	Speakers	Status
Livonian	Latin	liv	**0**/5	~30[†]	CE
Võro	Latin	vro	**1**/5	~100K	DE
Komi	Cyrillic	kpv	**1**/5	~160K	DE

Table 1: Language statistics of the targeted languages. The *class* column indicates the amount of data available (on a scale from 0 to 5) as defined by Joshi et al. (2020). *Status* according to Moseley (2010): DE - *definitely endangered*; CE - *critically endangered*. †- people able to communicate in Livonian[2].

Our contributions cover the full cycle of LLM development, including continued pre-training and instruction tuning, as well as benchmark creation, and both automatic and manual evaluation. During pre-training and instruction-tuning, we make use of cross-lingual transfer from related higher-resourced languages and parallel translation data. Additionally, we rely on the small amounts of available parallel data for the included languages in order to develop intermediate translation functionality, which is then applied to machine-translate instructions into the target languages.

We also describe our significant manual effort. First, this includes extending two existing benchmarks to these target languages (multiple-choice question-answering and topic classification), one of which required additional manual translation. Second, it involves creating a new parallel multi-turn benchmark for these XLR target languages. The extended benchmarks enable automatic evaluation of language models on the target languages, while the multi-turn benchmark allows us to assess the real-life usefulness of instruction-tuned models.

We use the multi-turn benchmark to conduct

*Equal contribution

[1] *Finno-Ugric* translates to Estonian as ***s**oome-**ugri***, to Finnish as ***s**uo**m**alais-**ugri**laiset*, to Võro as ***s**oo**m**õ-**ugri***, and to Livonian as ***s**ūo**m**õ-**ugr**õ*, which is why we refer to it as SMUGRI.

[2] www.livones.net/en/valoda/the-livonian-language

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6692–6712

extensive human evaluation, comparing our models to GPT-3.5-turbo. Translation, benchmark creation, and human evaluation were all carried out by native speakers of Komi and Võro. As there are no native speakers of Livonian, this work was handled by fluent speakers.

Evaluation on multiple-choice QA benchmarks indicates that our instruction-tuned models either outperform or are on par with strong proprietary baselines (GPT-3.5-turbo and GPT-4-turbo), for Livonian and Komi. Extensive human evaluation on the multi-turn benchmark further supports these findings. However, both automatic and human evaluations reveal that our models slightly underperform on Võro compared to proprietary models, likely due to Võro's close similarity to Estonian, a language in which proprietary models excel. Nonetheless, human evaluation shows that our models significantly outperform proprietary ones in terms of naturalness for Võro, Komi, and Livonian.

We publish the training implementation, evaluation benchmarks, and models[3].

2 Background and Related Work

2.1 Finno-Ugric Languages

Finno-Ugric languages belong to the Uralic language family and are spoken primarily in regions surrounding the Baltic Sea and the Ural Mountains. Most of these morphologically complex languages are considered low-resource or extremely low-resource (XLR), with Finnish, Hungarian, and Estonian being the most well-resourced. In this work, we focus on three XLR Finno-Ugric languages: Võro, Komi, and Livonian. These languages differ in both script and resource availability (see Table 1), which includes not only textual data but also the number of speakers. For example, Livonian has only 30 speakers, yet its community is highly active in preserving and revitalizing the language, exemplified by the establishment of the UL Livonian Institute in 2018.

The communities of Finno-Ugric language speakers have actively contributed to the development of modern NLP tools for their languages, including core NLP technologies, such as foundation models (Tanvir et al., 2020; Kuulmets et al., 2024; Luukkonen et al., 2023, 2024), which underpin many advanced NLP applications. Additionally, practical tools like machine translation systems (Yankovskaya et al., 2023; Tars et al., 2022, 2021; Rikters et al., 2022) and speech synthesis technologies (Rätsep and Fishel, 2023) have been developed to further support the use of these languages.

Supporting languages. To address the extreme data scarcity during continued pre-training, we include additional languages into the training data. First, we include Estonian and Finnish, which belong to the same Balto-Finnic subgroup as Võro and Livonian. Second, Latvian, due to its significant influence on Livonian and the fact that many Livonian speakers also speak Latvian. Third, Russian, because of its strong influence on Komi and the proficiency of many Komi speakers in Russian.

2.2 Multilingual LLMs

Multilingual LLMs are widely explored to expand the language coverage of LLMs. Traditional approaches involve training models from scratch (Luukkonen et al., 2024, 2023; Wei et al., 2023; Kudugunta et al., 2023). However, adapting pre-trained English-centric models to other languages through continued pre-training has also yielded promising results across various languages (Csaki et al., 2024; Dou et al., 2024; Rijgersberg and Lucassen, 2023; Lin et al., 2024; Andersland, 2024; Basile et al., 2023; Owen et al., 2024; Cui et al., 2024; Cui and Yao, 2024; Zhao et al., 2024; Etxaniz et al., 2024b). In the context of Finno-Ugric languages, the most relevant works to ours include Kuulmets et al. (2024), who adapted LLaMA-2 7B for Estonian, and Luukkonen et al. (2023), who adapted BLOOM (Workshop et al., 2023) for Finnish. Additionally, Luukkonen et al. (2023) demonstrate that continued pre-training of BLOOM outperforms Finnish models trained from scratch, emphasizing the advantages of this approach.

The development of multilingual LLMs often employs techniques that enhance model quality. Common practices include incorporating parallel data into the pre-training phase (Luukkonen et al., 2024; Owen et al., 2024; Wei et al., 2023) and utilizing curriculum learning (Wei et al., 2023).

2.3 Instruction Tuning

Previous works have also explored a variety of cross-lingual techniques for teaching the models to follow instructions (Li et al., 2023; Zhu et al., 2023; Zhang et al., 2024; Chai et al., 2024; Ranaldi and Pucci, 2023; Chen et al., 2023). Zhang et al. (2024) creates model answers to instructions in a high-resource/high-quality language, which are

[3]https://github.com/TartuNLP/smugri-llm

then translated and code-switched. Adding translation data during instruction-tuning has also been widely explored (Cui et al., 2024; Kuulmets et al., 2024; Zhu et al., 2023; Zhang et al., 2024; Ranaldi and Pucci, 2023; Chen et al., 2023). Kuulmets et al. (2024) find that using a diverse set of instructions in English can increase performance in Estonian tasks.

2.4 Evaluation

Common approaches to evaluating the multilingual capabilities of LLMs include using existing cross-lingual benchmarks (Ahuja et al., 2023a,b) or translating English benchmarks into target languages through either machine translation (Lai et al., 2023) or manual translation (Shi et al., 2022). However, extending the evaluation of conversational capabilities to other languages is more complex, as a gold standard requires the involvement of human annotators (Touvron et al., 2023). Human annotators are essential for both the recently popularized method of ranking models using the Elo rating system (Zheng et al., 2024) and the traditional method of pairwise comparison of answers from different models to predefined prompts (Zheng et al., 2024; Touvron et al., 2023).

An alternative line of research explores using LLMs as potential replacements for human annotators (Zheng et al., 2024; Kim et al., 2023, 2024). While strong LLMs can effectively serve as substitutes for human annotators in English, their capabilities in non-English languages remain unclear. Hada et al. (2024) investigate this across eight high-resource non-English languages, finding a bias in GPT-4-based evaluators toward assigning higher scores. To our knowledge, the behavior of LLM judges on XLR languages, including Finno-Ugric languages, has not been systematically studied.

3 Experimental Setup

Lang	Characters	Sampled Characters			
		Stage 1	Stage 2	Total	Ratio
LIV	2.6M	-	10.3M	10.3M	4.00
VRO	14.0M	-	56.1M	56.1M	4.00
KPV	578.9M	-	1.4B	1.4B	2.48
LV	27.8B	3.0B	300.0M	3.3B	0.12
ET	32.6B	8.2B	300.0M	8.5B	0.26
FI	114.0B	7.6B	300.0M	7.9B	0.07
RU	>1T	2.7B	300.0M	3.0B	<0.01
EN	>1T	2.7B	300.0M	3.0B	<0.01

Table 2: Training dataset composition. All of the data for Livonian is sentence-level, for other languages, the data is document-level.

3.1 Continued Pre-training

We take the approach of adapting the English-centric Llama-2 7B model (Touvron et al., 2023) to the target languages through full fine-tuning. Given our computational budget limitations, we employ a two-stage training strategy. In the first phase, we continue pre-training Llama-2 7B on higher-resource languages Finnish, Estonian, Russian and Latvian. In the second phase, we focus on teaching the model the XLR target languages resulting in **Llama-SMUGRI**. The training hyperparameters are detailed in Appendix B.

Stage 1: Learning supporting languages. In the first step, we continue pre-training Llama-2 7B (Touvron et al., 2023) on higher-resource languages Estonian, Finnish, English, Latvian, and Russian. We allocate a training budget of 10 billion tokens and sample documents from CulturaX (Nguyen et al., 2023), with 32%, 32%, 12%, 12%, and 12% probability of choosing the document from the respective language.

Stage 2: Learning Võro, Komi and Livonian. The second stage of continued pre-training focuses on enhancing the understanding and generative capabilities for XLR languages. We employ a character-based budget to ensure a balanced representation of languages in the training dataset. This budget is set at 3 billion characters, with 50% allocated to sampling Võro, Komi, and Livonian using Unimax with N=4 (Chung et al., 2023), and the remaining 50% uniformly distributed among the supporting languages to maintain the quality achieved in Stage 1. The N was chosen based on perplexity from the held-out validation set (see Appendix C). The Komi documents are sourced from FU-LAB's Komi corpus[4]. The Livonian dataset consists of sentence-level data from Rikters et al. (2022), while Võro dataset is compiled from various pre-existing corpora as well as data we have scraped. A more detailed overview of Võro dataset can be found in Appendix E.

Stage 2 + parallel: making use of parallel translation data. To investigate the role of parallel translation examples in the pre-training data, we incorporate translation examples formatted into various templates, accounting for up to 1% of the Stage 2 character budget (159,712 sentence pairs). We use Unimax with N=1 to balance the budget between language pairs. For further details on the

[4] http://wiki.fu-lab.ru/index.php/Электронная_база_коми_текстов

parallel data, we refer the reader to Appendix H. This stage yields the final base model we refer to as **Llama-SMUGRI**.

3.2 Instruction-tuning

Supporting instructions. We utilize existing instruction-tuning datasets across multiple languages. For English, Russian, and Finnish, we use Aya (Singh et al., 2024), and the highest-rated conversation paths of OASST-2 (Köpf et al., 2023). Additionally, we sample 20,000 Estonian instructions from Alpaca-est (Kuulmets et al., 2024). Following Kuulmets et al. (2024), we include 5,000 instructions from the FLAN-V2 (Longpre et al., 2023) TULU mixture (Wang et al., 2023) and 20,000 examples from Alpaca-GPT-4 (Peng et al., 2023), to improve cross-lingual knowledge transfer from high-quality English instructions. We refer to this instruction mixture as `SupInst` (Supporting Instructions). Further details are listed in Appendix F.

XLR Language Instructions. Due to LLMs' insufficient capabilities in XLR languages, it is not feasible to create Alpaca-style instructions directly. Consequently, we create instruction datasets for Võro, Livonian and Komi by translating 1000 examples per language from Alpaca-style instruction datasets into these languages. An external system, Neurotõlge[5] (Yankovskaya et al., 2023), is used for translation. While Võro and Livonian are translated directly from Alpaca-est (Kuulmets et al., 2024), Komi is generated by first translating Alpaca-GPT-4 (Peng et al., 2023) into Russian using `GPT-3.5-turbo`, and then translating that result into Komi with Neurotõlge. We refer to this dataset as `TrAlpaca`.

To investigate a scenario where a translation model is unavailable, we explore handling translating Alpaca instructions to XLR languages by fine-tuning our base model (Llama-SMUGRI) for the translation task (discussed in §3.3) resulting in `LLMTrAlpaca` instructions. This is similar to self-translate-train (Ri et al., 2024) and self-translate (Etxaniz et al., 2024a) concepts with the difference that we add a fine-tuning step to obtain an LLM specialized for translation. Further instruction-translation details and human evaluation of translations are discussed in Appendix G.

Translation instructions. We augment the general instructions with translation task instructions for Võro, Livonian, and Komi, using 250 examples per direction. We refer to these translation task instructions as `TrInst` (see Appendix H for dataset overview).

[5]https://neurotolge.ee/

3.3 Translation-tuning

Adapting general-purpose LLMs for the machine translation task has been shown to yield competitive results compared to dedicated MT systems (Xu et al., 2023; Kuulmets et al., 2024). Therefore, we fine-tune our base model on available translation data by sampling up to 100,000 sentence pairs from each language pair (see Appendix H for further details) to compare the quality of our model to using an MT system. We call this configuration `TrTuning`.

4 Benchmarks

Benchmark		Size	Type
MT-bench-SMUGRI	**[new]**	80	multi-turn questions
Belebele-SMUGRI	**[extended]**	127	multi-choice QA
SIB-SMUGRI	**[extended]**	125	topic classification
FLORES-SMUGRI	Yankovskaya et al. (2023)	250	translation

Table 3: Test benchmarks for Komi, Võro, and Livonian.

4.1 Automatic Evaluation

Existing benchmarks. From the existing benchmarks we use FLORES-SMUGRI (Yankovskaya et al., 2023) machine translation benchmark. It includes the first 250 sentences of FLORES-200 (NLLB-Team et al., 2022; Goyal et al., 2022) translated into several Finno-Ugric languages.

New benchmarks. We extend the topic classification benchmark SIB-200 (Adelani et al., 2024) and the multiple-choice QA benchmark Belebele (Bandarkar et al., 2023) to include Livonian, Võro, and Komi. Both SIB-200 and Belebele build on top of FLORES-200 and, therefore, can be extended to Livonian, Võro, and Komi using translations by Yankovskaya et al. (2023). We align these translations with sentences in SIB-200 and with paragraphs in Belebele. We then manually translate the questions and answer choices in Belebele into the target languages, as FLORES-200 does not contain these components. Table 3 shows the details of all evaluation benchmarks.

4.2 A Novel Multi-turn Benchmark

4.2.1 Requirements

We formulate the following desiderata for a human-evaluation benchmark considering the XLR use-case.

1) Questions should cover real-life usage scenarios to reflect real-life usefulness. The easiest and most likely way for speakers of low-resource Finno-Ugric languages to benefit from LLMs is through interaction via a chat-like interface. Our novel Finno-Ugric benchmark is designed to cover such real-life use cases. Consequently, it should consist of user prompts similar to real-life queries. Another benefit of using real-life data is that it helps quickly reveal the model's usefulness and potential weaknesses in practical scenarios, which standard NLP benchmarks typically do not cover.

2) Questions should be challenging enough for LLMs to differentiate the models accurately. Zheng et al. (2023) show that challenging prompts from real-life conversations reveal larger performance gaps between different models compared to a manually designed benchmark of high-quality challenging questions.

3) Answering questions should not require expert knowledge. A key requirement for the benchmark is that it should comprise questions that are challenging for language models. However, such questions are often challenging for humans as well, requiring expert-level knowledge in various domains. For example, Zheng et al. (2024) uses graduate students as labelers, considering them more knowledgeable than average crowd workers. Requiring expert-level knowledge from evaluators shrinks the potential evaluator pool, making it nearly impossible to find them from the communities of XLR language speakers.

4) Translating the benchmark into a new language should be feasible in terms of time and content (i.e., should not require expert knowledge). Since no data on human interactions with chat LLMs exists for XLR languages, we collect the data in English and translate it. Given the limited availability of translators, we carefully select examples that are straightforward to translate.

4.2.2 Constructing the Dataset

We manually collect the initial dataset from LMSYS-Chat-1M (Zheng et al., 2023), which consists of real-world user interactions with LLMs. First, we extract all two-turn English conversations that have not been redacted or flagged by OpenAI moderation API. We only allow conversations with user prompts no longer than 50 tokens to ease the translation process. We then use `all-MiniLM-L12-v2` model from SentenceTransformers (Reimers and Gurevych, 2019) to compute the sentence embedding and apply fast clustering implemented in `sentence-transformers` which finds local groups of texts that are highly similar. We manually examine a few examples from each cluster and pick user prompts that fill the criteria specified in Chapter 4.2.1. Finally, we removed the observed clusters from the dataset and recluster the remaining examples with a smaller similarity threshold until we had collected 248 multi-turn conversations in total.

We organize conversations into four categories: math, reasoning, writing, and general. As we wanted similarly to Zheng et al. (2024), the final dataset to consist of 80 questions — 20 from each category (potentially with follow-ups) — the initial dataset had to be filtered. For that purpose, we asked GPT-4 to rate the difficulty of each question as was done by Zheng et al. (2023). However, we observed no correlation between the difficulty of the question and the quality of the answer given by ChatGPT when quality was assessed by GPT-4. Therefore, the final dataset was also picked manually by removing near duplicate questions and — after looking at the generated answers — also questions where judging the answer still seemed to require too specific knowledge. The final dataset consists of 80 real-life prompts among which 42 are multiturn. It was then translated into Võro, Komi, and Livonian by fluent speakers with a linguistic background or previous experience in translation. The translators were asked to preserve any informality of the text in the translations, e.g. missing uppercase and punctuation.

Instead of human translators, one could use a machine translation system or a proprietary LM for test data generation or translation. We explore both options and find that proprietary models struggle to generate text in our target languages, while even the best machine translation systems produce translations that are often judged inferior to human translations (see Appendix J for further details).

5 Results

5.1 Pre-training

Stage 1 continued pre-training on high-resource supporting languages shows notable improvements in SIB-SMUGRI for Võro and Livonian compared to Llama-2-7B (see Stage 1 in Table 4). There is also a clear reduction in perplexity (see Table 5) for Võro and Livonian, while no such improvement is observed for Komi. Similarly, the Llama

Model	SIB-SMUGRI 5-shot, acc			BELEBELE-SMUGRI 3-shot, acc			FLORES-SMUGRI 5-shot, BLEU					
	VRO	LIV	KPV	VRO	LIV	KPV	ET-VRO	ET-LIV	RU-KPV	VRO-ET	LIV-ET	KPV-RU
Llammas-base	78.4 (3.7)	69.6 (4.1)	64.0 (4.3)	30.7 (4.1)	28.4 (4.0)	32.3 (4.2)	11.5 (0.9)	4.3 (0.5)	1.7 (0.4)	28.7 (1.5)	8.0 (0.8)	2.2 (0.3)
Llama-2-7B	57.6 (4.4)	60.0 (4.4)	58.4 (4.4)	29.1 (4.1)	**29.9** (4.1)	**36.2** (4.3)	11.1 (1.0)	4.6 (0.6)	1.5 (0.3)	11.3 (0.9)	4.4 (0.6)	2.4 (0.3)
Llama-SMUGRI (ours)												
Stage 1	80.8 (3.5)	**75.2** (3.9)	65.6 (4.3)	32.3 (4.2)	26.8 (3.9)	26.0 (3.9)	11.5 (1.0)	4.2 (0.5)	2.6 (0.6)	29.6 (1.4)	7.2 (0.7)	4.1 (0.7)
Stage 2	78.4 (3.7)	65.6 (4.3)	74.4 (3.9)	31.5 (4.1)	26.0 (3.9)	28.4 (4.0)	26.5 (1.1)	3.4 (0.4)	15.7 (1.0)	45.3 (1.5)	10.6 (0.9)	18.6 (0.9)
Stage 2 + parallel	**84.0** (3.3)	66.4 (4.2)	**76.8** (3.8)	**35.4** (4.3)	27.6 (4.0)	29.1 (4.1)	**29.1** (1.2)	**4.3** (0.5)	**16.0** (1.0)	**48.7** (1.4)	**17.6** (1.0)	**22.1** (1.3)

Table 4: **Pre-training results for extremely low-resource Finno-Ugric languages.** Standard errors are reported for the scores in parentheses: *score* (*stderr*). For comparison, we report Llama-2 7B and Llammas-base (Kuulmets et al., 2024). **Stage 1** consists of pre-training with high-resource related languages while **Stage 2** additionally includes Võro, Livonian and Komi. **Stage 2 + parallel** incorporates additional parallel translation data into training.

Model	byte-PPL		
	VRO	LIV	KPV
Llammas-base	3.3548	12.1081	3.1959
Llama-2-7B	6.1528	14.8055	3.1198
Llama-SMUGRI (ours)			
Stage 1	3.4895	11.4210	3.1341
Stage 2	2.1885	3.8351	1.4055
Stage 2 + parallel	**2.1837**	**3.7615**	**1.4050**

Table 5: **Pre-training perplexity for extremely low-resource Finno-Ugric languages.** For comparison, we report Llama-2 7B and Llammas-base (Kuulmets et al., 2024).

model fine-tuned for Estonian (Llammas-base) exhibits lower perplexity than Llama for Livonian and Võro, which are closely related to Estonian. The lack of improvement for Komi may result from its more distant relationship (see Appendix I) to the other Finno-Ugric languages in the dataset, as well as its use of a different script. These results suggest that related languages generally improve benchmark scores for XLR languages that were not included in the training. For Belebele-SMUGRI, there is no improvement compared to Llama-2-7B, while FLORES-SMUGRI shows improvement only when translating from the XLR languages into the higher-resourced languages.

Stage 2 pre-training, which targets XLR Finno-Ugric languages, further enhances both perplexity and, with the exception of ET-LIV translation, FLORES-SMUGRI scores, indicating that the model has acquired generative capabilities for these languages. The performance improvements on the SIB-SMUGRI benchmark are modest for Komi, while Livonian and Võro show a slight decrease from the previous stage. This stage of training has also failed to improve the Belebele scores.

Stage 2 + parallel data results in minimal improvements in benchmarks and perplexity, with the exception of translation tasks from the XLR languages showing larger gains. This indicates that the inclusion of parallel data has a limited impact or that our benchmarks are insufficiently sensitive to capture these effects. Nevertheless, due to the slightly positive influence observed, we will use this setup as the basis for subsequent instruction tuning.

Benchmarks. The current benchmarks may not effectively differentiate between models at this stage, as their small size and high standard errors limit our ability to draw fine-grained conclusions about training strategies. Additionally, the low scores on the Belebele benchmark suggest it may be too challenging for the models. In contrast, the relatively high scores on the SIB-200 benchmark could result from its simplicity, allowing the models to classify texts based on single-word clues rather than a deeper understanding of the language. Designing automatic benchmarks with an appropriate difficulty level and relevant context for these languages is an important challenge for future research.

5.2 Instruction-Tuned Models

Examining the scores of commercial systems in Table 6, we observe that these models exhibit at least some understanding of Võro, Livonian, and Komi. Based on benchmark scores, they seem to understand Võro and Livonian better than Komi. This could be explained by the linguistic similarity between these languages and Estonian – an average Estonian speaker can understand most of a Võro text and some of a Livonian text, but not much Komi, as it is more distantly related and written in a different script. The performance of GPT-4-turbo and GPT-3.5-turbo aligns with this trend, with scores typically following this order. For instance, GPT-4-turbo achieves 92% accuracy on the Belebele Estonian benchmark, so it is unsurprising

Model	BELEBELE-SMUGRI 0-shot, acc			SIB-SMUGRI 5-shot, acc		
	VRO	LIV	KPV	VRO	LIV	KPV
GPT-3.5-turbo	45.7 (4.4)	37.8 (4.3)	34.6 (4.2)	81.6 (3.5)	73.6 (4.0)	68.8 (4.2)
GPT-4-turbo	**70.1** (4.1)	40.2 (4.3)	**44.1** (4.4)	**92.0** (2.5)	72.0 (4.0)	67.2 (4.2)
Llammas (Kuulmets et al., 2024)	36.2 (4.3)	32.3 (4.2)	27.6 (4.0)	80.8 (3.5)	78.4 (3.7)	63.2 (4.3)
Llama-SMUGRI-Instruct						
`SupInst`	42.5 (4.4)	30.7 (4.1)	**44.1** (4.4)	86.4 (3.1)	79.2 (3.6)	**88.8** (2.8)
`SupInst+LLMTrAlpaca`	39.4 (4.3)	35.4 (4.3)	42.5 (4.4)	85.6 (3.1)	**81.6** (3.5)	84.8 (3.2)
`SupInst+TrAlpaca`	35.4 (4.2)	32.3 (4.2)	40.2 (4.3)	85.6 (3.1)	79.2 (3.6)	85.6 (3.1)
`SupInst+LLMTrAlpaca+TrInst`	44.9 (4.4)	**40.9** (4.4)	**44.1** (4.4)	86.4 (3.1)	76.0 (3.8)	78.4 (3.7)
`SupInst+TrAlpaca+TrInst`	45.7 (4.4)	32.3 (4.2)	**44.1** (4.4)	86.4 (3.1)	78.4 (3.7)	78.4 (3.7)

Table 6: **Instruction-tuning evaluation results.** Standard errors are reported for the scores: *score* (*stderr*).

that it also performs well on Võro.

Our models demonstrate comparable performance to GPT-3.5-turbo on Võro and Livonian and slightly outperform it on Komi. However, GPT-4-turbo significantly surpasses our models on Võro and matches our performance on Livonian and Komi. A similar trend emerges in the SIB benchmark: our models outperform GPT-4-turbo on Livonian and Komi but underperform on Võro. Meanwhile, GPT-3.5-turbo consistently scores lower across all XLR languages.

We observe that the different instruction-tuning strategies produce similar results. Given the small size of our benchmarks and the associated high standard errors, we cannot make definitive conclusions about which strategy is superior.

LLM-translated instructions. Automatic metrics indicate that instructions translated using our translation-tuned LLM achieve results comparable to those produced by the external system Neurotõlge. However, the results do not provide enough clarity or confidence to definitively favor one method over the other. These findings suggest that even in the absence of external translation systems, the translation-tuned LLM can serve as a viable alternative.

Translation instructions. Incorporating a small set of translation instructions (250 for each Võro, Komi, and Livonian direction) does not lead to clear and consistent improvements in the discriminative benchmarks (see Table 6). Human evaluation of conversations (Section 5.4) produces similar findings. However, there is a notable improvement in the translation benchmark, even with this minimal data (see Table 7). We believe these translation examples mainly help the model respond in the correct language, while the underlying language and translation capabilities are already present in the base model.

Translation evaluation. When assessing language generation using the FLORES translation benchmark, the results in Table 7 show that GPT-family models can translate from Estonian to Võro quite effectively, suggesting that Võro might have been included in their training data. In contrast, the low BLEU scores for Livonian and Komi indicate very limited translation capabilities. Our LLMs, which were not exposed to translation examples during instruction-tuning, struggle to translate into Võro, Livonian, and Komi. However, they perform better in the reverse direction, even surpassing GPT models for Komi. A closer look reveals that they copy the high-resource language sentences to the output when translating to the low-resource languages. When the translated Alpaca instructions (`TrAlpaca` and `LLMTrAlpaca`) were added, we observed that when asked to translate from the low-resource languages, the models often copied the source text to the output as well, resulting in lower BLEU scores. This can be addressed by including a small amount of translation data during instruction-tuning (`TrInst`).

5.3 Translation-tuning

Model	ET-VRO		ET-LIV		RU-KPV	
	←	→	←	→	←	→
GPT-3.5-turbo	34.0	15.1	7.7	2.7	6.7	0.5
GPT-4-turbo	47.5	20.5	9.9	3.7	8.7	3.1
Neurotõlge	48.5	21.2	**29.7**	**10.2**	**31.5**	**17.7**
Llama-SMUGRI-Instruct						
`SupInst`	41.9	10.7	11.1	4.6	21.4	3.0
`SupInst+TrAlpaca`	16.8	10.6	9.7	4.7	17	3.2
`SupInst+LLMTrAlpaca`	23	10.8	9.2	4.6	13.5	2.9
`SupInst+TrAlpaca+TrInst`	45.3	19.1	19.9	5.5	21.4	15.2
`SupInst+LLMTrAlpaca+TrInst`	47.7	21.2	20.6	6.2	20.9	16.4
Llama-SMUGRI-translate	**50.5**	**29.2**	24.0	10.0	23.4	17.3

Table 7: BLEU scores on FLORES-SMUGRI (0-shot). Translations are generated with beam size 4 for our models.

We compare our LLM-based translation mod-

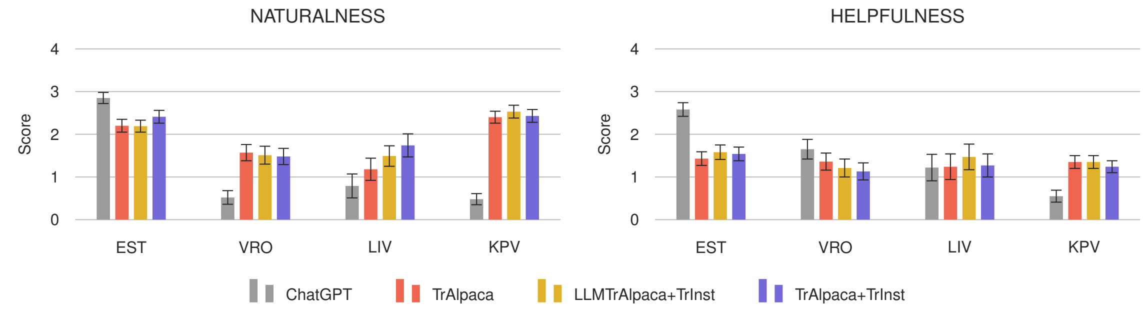

Figure 1: Human evaluation scores for naturalness and helpfulness across different models.

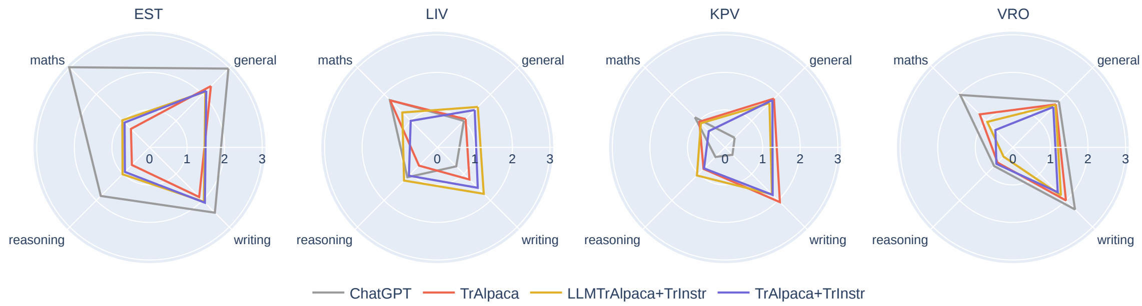

Figure 2: Helpfulness across languages and categories.

els with Neurotõlge, which supports low-resource Finno-Ugric languages. Our translation-tuned models outperform Neurotõlge in both the VRO-ET and ET-VRO translation directions (see Table 7). For the ET-LIV and RU-KPV directions, our models achieve performance that is comparable to Neurotõlge. However, in translations from low-resource to high-resource languages – except for Võro – our models do not perform as well.

5.4 Human Evaluation

We select 3 instruction-tuned models for human evaluation: `TrAlpaca`, `LLMTrAlpaca+TrInst` and `TrAlpaca+TrInst`. As a baseline, we use GPT-3.5-turbo, which was freely accessible via a chat interface[6] at the time of the evaluation. For each target language, we designed a survey where participants rate the helpfulness of the answer from a randomly selected model on a 5-point Likert scale. Additionally, we ask participants to rate how natural the response sounds in the target language, as Kuulmets et al. (2024) notes that model outputs often sound unnatural in these languages. The surveys were distributed within target language communities via social media and direct outreach to speakers (see Appendix A for the screenshot of the survey). We did not collect any personal data from respondents.

In addition to Võro, Livonian, and Komi, we also gather and present human evaluation data for Estonian, which is closely related to both Võro and Livonian. At the same time, Estonian is well-supported by GPT-3.5-turbo (Kuulmets et al., 2024), providing a meaningful anchor point for comparing our human evaluation results.

The results indicate that our models underperform in helpfulness compared to GPT-3.5-turbo in Estonian, which aligns with previous findings (Kuulmets et al., 2024). A similar disparity exists for Võro, where our models still lag behind. However, for both Võro and Livonian, the helpfulness scores of our models are comparable to those of GPT-3.5-turbo. In contrast, our system outperforms the commercial baseline for Komi. While variations in annotator expectations may influence results across different languages, it is noteworthy that our models consistently achieve similar helpfulness scores across various languages.

Comparisons by category (see Figure 2) reveal that the scores for GPT-3.5-turbo are inflated by examples in the *maths* and *reasoning* categories, where our models demonstrate less helpfulness. In contrast, our models perform comparably in the

[6] https://chatgpt.com/

general and *writing* categories. Notably, in Komi, our models surpass GPT-3.5-turbo in both *general* and *writing* tasks, while achieving similar scores in the *maths* and *reasoning* tasks.

In terms of response naturalness, GPT-3.5-turbo performs slightly better for Estonian; however, our models demonstrate greater naturalness in all other languages, especially in Komi, where the difference is particularly pronounced.

When comparing our trained models, no clear ranking emerges, reinforcing the findings from automatic benchmarks that incorporating translation instructions does not produce significant advantages. Additionally, there is little difference between using LLM-translated instructions and those translated by an external system.

6 Conclusion

We implemented a comprehensive approach encompassing data collection, instruction tuning, and human evaluation for three extremely low-resource Finno-Ugric languages: Võro, Livonian, and Komi. Our contributions include an exploration of pretraining and instruction-tuning strategies, leading to the development of open-source multilingual base and instruction-tuned models for these languages. We also extend the automatic evaluation benchmarks, Belebele and SIB-200, to include Komi, Livonian, and Võro, and we introduce a novel multi-turn conversational benchmark, SMUGRI-MT-BENCH. Human evaluation using SMUGRI-MT-BENCH demonstrates that our models surpass `GPT-3.5-turbo` in terms of naturalness and achieve higher helpfulness for Komi, while maintaining comparable levels for the other low-resource languages.

Limitations

There are several limitations that may affect the robustness and generalizability of our findings. Firstly, the automatic benchmarks used are small and exhibit high standard errors, making fine-grained comparisons difficult. This issue is compounded by our reliance on the FLORES-200 dataset, which limits the scope of our evaluation to the specific topics and set of sentences it covers. Furthermore, our automatic evaluation utilized only three tasks, which constrains the comprehensiveness of our assessment. From these three, only one (translation) measured generative performance, as no other suitable benchmarks exist for these languages. This narrow focus on translation might not fully capture the generative capabilities of the models across different tasks. However, human evaluation addresses these concerns to some extent, providing a more detailed and reliable assessment of the model's quality in a multi-turn chat assistant scenario.

The heavy reliance on the FLORES-200 dataset is caused by the difficulties related to creating new datasets. Creating high-quality benchmarks for XLR languages is tricky because the data can not be obtained by machine translating benchmarks from other languages, as the machine translation systems are potentially too weak. Additionally, hiring professional translators is difficult due to the scarcity or absence of individuals experienced in translating these languages, particularly when the languages are not officially recognized. Finally, since finding human annotators for XLR languages in itself is challenging, finding expert-level annotators becomes almost impossible, and thus, the set of prompts used for human evaluation must be constructed so that assessing the quality of the answer would not require any specific expert-level knowledge.

A limitation of our instruction-tuning process is that we only used machine-translated instructions for the XLR languages. As a result, some of these instructions were of low quality, potentially affecting the overall performance and reliability of the fine-tuned models.

Our emphasis on Finno-Ugric languages means that our findings might not apply to other language families, which could present different challenges or yield different results in a more diverse multilingual context. To address these limitations, future research should aim to develop larger and more diverse benchmarks and apply similar methodologies to a broader range of low-resource languages to validate and extend our findings.

Ethics Statement

Our models have not been extensively tested for the generation of harmful content. Furthermore, we were unable to check the training and instruction-tuning data for harmful content due to their sheer volume. Thus, we can not guarantee the models' harmlessness and advise them to be used with this in mind only for research purposes. Furthermore, our models still make many mistakes when generating the responses, and their output should not be considered an accurate representation of the low-resource languages without manual verification.

Acknowledgments

This work was supported by the Estonian Research Council grant PRG2006 (Language Technology for Low-Resource Finno-Ugric Languages and Dialects). All computations were performed on the LUMI Supercomputer through the University of Tartu's HPC center.

We would like to express our sincere gratitude to the translators Janek Vaab (Võro), Aleksei Ivanov (Komi), and Marili Tomingas (Livonian). We would also like to thank Võro, Komi, and Livonian communities for their help in annotating the data.

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.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. Megaverse: benchmarking large language models across languages, modalities, models and tasks. *arXiv preprint arXiv:2311.07463*.

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *Preprint*, arXiv:2403.06354.

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*.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *Preprint*, arXiv:2312.09993.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Preprint*, arXiv:2401.07037.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *Preprint*, arXiv:2310.20246.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages. *Preprint*, arXiv:2404.05829.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Yiming Cui and Xin Yao. 2024. Rethinking llm language adaptation: A case study on chinese mixtral. *Preprint*, arXiv:2403.01851.

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

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024a. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

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

Wikimedia Foundation. Wikimedia downloads.

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. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

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

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3519–3529. PMLR.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. *Preprint*, arXiv:2404.04042.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023. Align after pre-train: Improving multilingual generative models with cross-lingual alignment. *Preprint*, arXiv:2311.08089.

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.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality. *Preprint*, arXiv:2404.01856.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. In

Advances in Neural Information Processing Systems, volume 36, pages 50358–50376. Curran Associates, Inc.

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.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic 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. *Preprint*, arXiv:2207.04672.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Louis Owen, Vishesh Tripathi, Abhay Kumar, and Biddwan Ahmed. 2024. Komodo: A linguistic expedition into indonesia's regional languages. *Preprint*, arXiv:2403.09362.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Liisa Rätsep and Mark Fishel. 2023. Neural text-to-speech synthesis for Võro. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727, Tórshavn, Faroe Islands. University of Tartu Library.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ryokan Ri, Shun Kiyono, and Sho Takase. 2024. Self-translate-train: Enhancing cross-lingual transfer of large language models via inherent capability. *Preprint*, arXiv:2407.00454.

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

Matiss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernstreits, and Mark Fishel. 2022. Machine translation for livonian: Catering to 20 speakers. In *ACL (2)*, pages 508–514.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Hasan Tanvir, Claudia Kittask, and Kairit Sirts. 2020. Estbert: A pretrained language-specific bert for estonian. *Preprint*, arXiv:2011.04784.

Maali Tars, Taido Purason, and Andre Tättar. 2022. Teaching unseen low-resource languages to large translation models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 375–380, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maali Tars, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. *Preprint*, arXiv:2105.13065.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, 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, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. *Preprint*, arXiv:2402.12204.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,

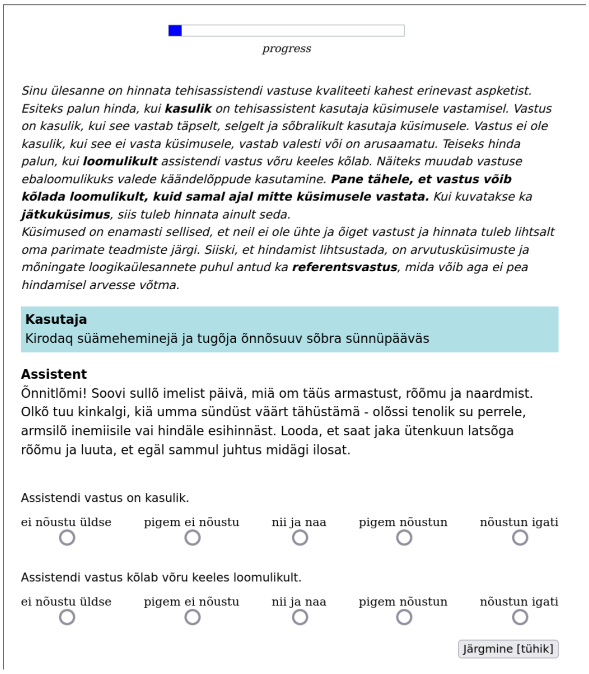

Figure 3: Screenshot of the survey that was used to collect human annotations.

Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *Preprint*, arXiv:2308.04948.

A Collecting Data for Human Evaluation

The screenshot of the survey is shown in Figure 3. For Võro, Livonian, and Estonian, the instructions were given in Estonian, while for Komi, they were given in Russian.

B Training Details

The hyperparameters of pre-training stages 1 and 2 are listed in Table 8. The instruction-tuning and translation-tuning parameters are in Table 9. The first epoch was used for evaluating instruction-tuned models.

All the models were trained using 4 AMD MI250x GPUs (acting as 8 units) on the LUMI supercomputer. We report GPU-hours elapsed for model training in Table 10.

Parameter	Stage 1	Stage 2 (translate)
updates	19073	2985 (3013)
LR	4.00e-5	2.00e-5
LR-schedule	cosine decay to 10%	
context length	2048	
batch size	256	
warmup ratio	0.01	
weight decay	0.05	
precision	bfloat16	
optimizer	AdamW	
packing	yes	

Table 8: Pre-training hyperparameters.

Parameter	Value
LR	2.00e-5
LR-schedule	cosine decay to 10%
context length	2048
batch size	256
epochs	2
warmup ratio	0.01
weight decay	0.05
precision	bfloat16
optimizer	AdamW
packing	no

Table 9: Instruction-tuning and translation-tuning hyperparameters.

Model	**GPU-hours**
Base:	
Stage 1	2008
Stage 2	308
Stage 2 + translate	316
Instruction:	
LLMTrAlpaca+TrInst	39
TrTuning	39

Table 10: GPU-hours elapsed for training the models.

Sampling	byte-PPL			Epochs			Proportion		
	VRO	LIV	KPV	VRO	LIV	KPV	VRO	LIV	KPV
Unimax									
N=1	2.3072	4.1986	1.4508	1	1	1	2.4%	0.4%	97.2%
N=4	**2.1885**	**3.8351**	**1.4055**	4	4	2.5	3.7%	0.7%	95.6%
N=8	2.5983	4.7250	1.4159	8	8	2.4	7.5%	3.3%	81.0%
Proportional	2.1983	3.8352	1.4065	2.5	2.5	2.5	2.4%	0.4%	97.2%

Table 11: The effect of Unimax N (max data repeat epochs) on held-out validation set byte perplexity.

C Choice of Unimax N

We chose the Unimax N according to the byte perplexity on our held-out validation set, with the best value for our setup being 4 (see Table 11).

Muennighoff et al. (2023) found that repeating data for 4 epochs is almost as good as new data with improvements continuing beyond 4 epochs for pre-training LLMs. We find that for continued pre-training with very small datasets and small data budgets, 4 epochs of repetition (Unimax N=4) provides an improvement in perplexity over 1 epoch of data for **Stage 2**. However, already at 8 epochs, the perplexity drops, suggesting overfitting (see Appendix C). Thus we keep the maximum repetitions at 4 and conclude that the number of repetitions of smaller datasets should be carefully chosen to avoid over- or underfitting.

	ET	VRO	LIV	KPV
surveys submitted	45	17	6	27
answers graded	1708	836	279	1306
grades per question	2.8	1.74	0.58	2.7

Table 12: Human evaluation data collection statistics.

D Evaluation details

The base models are evaluated with `lm-evaluation-harness` (Gao et al., 2023). For instruction-tuned models' SIB-SMUGRI outputs that do not conform to the expected format, we use `GPT-4-turbo` to verify that the prediction matches the ground truth. We calculate standard errors using bootstrap resampling implemented in `lm-evaluation-harness` (Gao et al., 2023). The evaluation prompts are listed in Figure 9. For Belebele, the instruction-tuned models' zero-shot evaluation method is based on Bandarkar et al. (2023).

`GPT-4-turbo` version used in evaluations was `gpt-4-turbo-2024-04-09` and `GPT-3.5-turbo` version used was `gpt-3.5-turbo-0125`.

We evaluate translations quality using BLEU (Papineni et al., 2002) calculated with sacreBLEU[7] (Post, 2018).

The held-out validation set (see Table 13) used to calculate perplexity is sampled from our pre-training data.

Language	Characters	Examples
LIV	86842	1246
VRO	131373	110
KPV	1308290	500

Table 13: Held-out validation set sizes. Examples for Livonian are sentences. For other languages they are documents.

E Võro Data Collection

We collect Võro data from Võro language Wikipedia dump (Foundation), Corpus of Fiction in Võro and Seto languages[8], Additionally, we scraped Võro language newspaper articles from *Uma Leht*[9]. Since the Seto dialect is similar to Võro, we do not filter it out of our Võro datasets that contain it, and additionally include "Setomaa" newspaper corpus[10] which is also in Seto dialect. The collected Võro dataset composition is shown in Table 14.

Name	Documents	Characters	Sentences
Võro			
Wikipedia (2024.02.20)	6385	3879212	88550
Fiction corpus	399	1987446	32121
Umaleht (scraped)	3392	6280588	93958
Seto dialect			
Fiction corpus	8	76361	869
Setomaa corpus	397	1791268	20693

Table 14: Võro data composition by source.

[7]signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.2`

[8]https://metashare.ut.ee/repository/browse/corpus-of-fiction-in-voro-and-seto-languages/2cf454fca0d411eebb4773db10791bcf485f3f9e7dee447b983f21b074ad3835

[9]https://umaleht.ee/

[10]https://metashare.ut.ee/repository/browse/setomaa-newspaper-corpus/3303e60ca0d411eebb4773db10791bcf2d01e0b55ce2419db34ef402460a1c99/

Dataset	**LIV**	**VRO**	**KPV**	**ET**	**FI**	**EN**	**RU**
Supporting language instructions:							
Aya (Singh et al., 2024)					742	3944	423
OASST-2 (Köpf et al., 2023)					5	3514	681
FLAN-V2 (Longpre et al., 2023)						5000	
Alpaca-GPT-4 (Peng et al., 2023)						20000	
Alpaca-est (Kuulmets et al., 2024)				20000			
TrAlpaca (ours)	1000	1000	1000				
TOTAL	1000	1000	1000	20000	747	32458	1104

Table 15: Instruction-tuning data with the number of sentences sampled

F Instruction-tuning details

The composition of our instruction-tuning dataset is listed in Table 15. Instructions are formatted into a chat-format shown in Figure 5. The translation-tuning data format is shown in Figure 6. The fine-tuning loss is calculated on target (assistant) tokens while the rest of the prompt is masked.

G Instruction translation details

When using our translation-tuned models for translating instructions (for `LLMTrAlpaca`), the models sometimes leave sentences untranslated in an unpredictable manner. Consequently, we removed examples where the BLEU (Papineni et al., 2002) score between the original and translated text exceeds 70. This process may also eliminate some valid examples, as identical texts can occur in some cases.

In preliminary experiments, we observed that the model sometimes struggles with multi-line or multi-sentence inputs, which are essential for accurately translating instructions that often consist of entire texts from Alpaca-style examples. To address this issue, we concatenate 50% of the training sentences into chunks of 2 to 6 sentences, training the model to handle longer inputs effectively. We refer to this configuration `TrTuningConcat` and the regular translation instructions as `TrTuning`.

We find that this concatenation does not harm model translation quality (see Table 17). Additionally, we observed more consistent outputs when translating whole instructions.

To get a glimpse into the quality of the translations, we conducted a small-scale human evaluation with native speakers or, in the case of Livonian a fluent speaker, due to the lack of native speakers. Given the original and the translation of 20 randomly chosen instructions, which stay the same across translation models, the evaluators were asked to rate the Fluency (*How fluent and natural does the translation sound in the target language?*) and Consistency (*Does the translation preserve the meaning and intent of the source text?*) on a 5-point Likert scale (see Table 18 for evaluation guidelines). We also asked them to report if the resulting *instruction-input-output* triplet does not form a correct instruction, e.g. the output does not satisfy the instruction.

We report the evaluation summary in Table 19 for Neurotõlge and Llama-SMUGRI-translate. In general, Llama-SMUGRI-translate produces better results for Võro and Livonian, achieving higher fluency and consistency scores than Neurotõlge on average. For an acceptable instruction translation, we would like the fluency and consistency to be at least 3, and the resulting instruction should still be correct. We see that for Neurotõlge, this is only achieved for 45% of evaluated instructions in Võro and Livonian, while for Komi 80% of instructions are acceptable. For Llama-SMUGRI-translate, we see this condition is satisfied for 70%, 80%, and 65% of translation for Komi, Võro, and Livonian, respectively. While this shows that machine translation can be a feasible option in some cases, it is far from ideal for instructions. We also report histograms of human ratings in Figure 4.

H Parallel data

Composition of the parallel data is shown in Table 16.

I Llama-SMUGRI Representations

The CKA scores (Kornblith et al., 2019) in Table 7 indicate that the more closely related Finno-Ugric languages – Estonian, Võro, and Livonian – written in the Latin script exhibit more similar representations in the intermediate layers. Meanwhile, Komi shows the highest CKA score with Russian, the only other language in our experiments that uses the Cyrillic script, while having a lower similarity score with languages for which the model has seen the most training data, such as English and

Dataset	VRO-ET	LIV-ET	LIV-LV	LIV-EN	KPV-ET	KPV-FI	KPV-RU	KPV-EN	KPV-LV	TOTAL
`TrInst`	500	500	500	493	500	500	500	500	500	**4493**
`TrTuning`	28505	14215	11608	493	3876	7273	100000	7288	5020	**178278**
Stage 2 + parallel	28504	14212	11606	492	3835	7272	81487	7286	5018	**159712**

Table 16: **Number of sentences of parallel data used in various training configurations.** In all cases, the language pair data is split equally; for instance, in ET-LIV, 50% of the reported sentences are for ET→LIV and the remaining 50% for LIV→ET. The data is sourced from Yankovskaya et al. (2023); Rikters et al. (2022); Tars et al. (2022, 2021).

Model	**ET-VRO**		**ET-LIV**		**RU-KPV**	
	←	→	←	→	←	→
Neurotõlge	48.5	21.2	**29.7**	**10.2**	**31.5**	**17.7**
Llama-SMUGRI-translate						
`TrTuning`	50.5	**29.2**	24.0	10.0	23.4	17.3
`TrTuningConcat`	**51.7**	28.7	22.9	9.7	23.5	17.4

Table 17: BLEU scores on FLORES-SMUGRI (0-shot). Translations are generated with beam size 4 for our models. TrTuningConcat uses concatenated sentences from multiple examples in a single translation instruction.

	Fluency
5	The translation is perfectly fluent, with no grammatical errors, unnatural phrasing, or awkward expressions.
4	The translation is mostly fluent, with minor grammatical or stylistic issues that do not affect readability.
3	The translation is somewhat fluent, but noticeable issues (e.g., awkward phrasing or grammatical errors) hinder smooth reading.
2	The fluency of the translation is poor, with significant issues that make it difficult to understand in parts.
1	The translation is completely unnatural or ungrammatical, making it incomprehensible.
	Consistency
5	The translation fully preserves the meaning of the source text, with no omissions, additions, or distortions.
4	The translation preserves the overall meaning, but there are minor inaccuracies or nuances lost.
3	The translation conveys the general meaning, but there are noticeable issues (e.g., omissions or slight distortions).
2	The translation distorts the meaning significantly or omits important details, making it partially inaccurate.
1	The translation fails to convey the meaning of the source text entirely.

Table 18: Instruction translation human evaluation guidelines.

Lang	**Fluency**	**Consistency**	**Incorrect**	**Both** ≥ 3
Neurotõlge				
KPV	3.55	3.5	5%	80%
VRO	2.75	3.2	0%	45%
LIV	2.8	2.85	15%	45%
Llama-SMUGRI-translate				
KPV	3.35	3.35	5%	70%
VRO	3.5	3.9	0%	80%
LIV	3.2	3.2	20%	65%

Table 19: Average fluency and consistency ratings of instruction translation (out of 5). Incorrect - the translation does not form a correct *instruction-input-response* triplet. Both ≥ 3 - the percentage of translated instructions where the Fluency and Consistency are at least 3, and the instruction is correct.

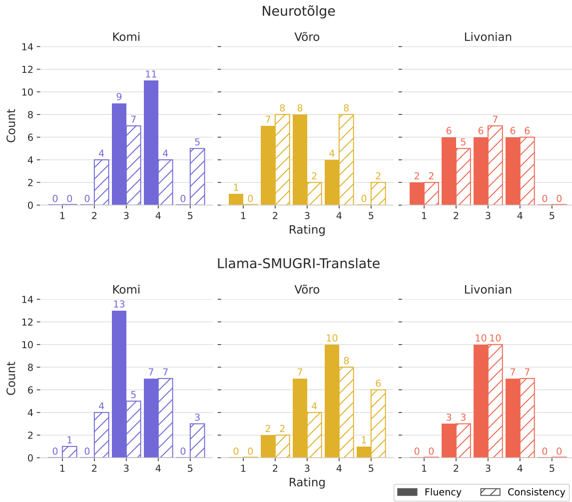

Figure 4: **Neurotõlge** and **Llama-SMUGRI-translate** instruction translation human evaluation scores.

```
<|user|>
Tere!
<|assistant|>
Tere! Kas saaksin teid kuidagi aidata?</s>
<|user|>
Kuidas alustada kirja kirjutamist?
<|assistant|>
```

Figure 5: Chat format following Wang et al. (2023) and Kuulmets et al. (2024). The model responds after `<|assistant|>`.

Estonian.

The t-SNE (van der Maaten and Hinton, 2008) plots in Figure 10 show that quite expectedly, the lexically similar languages Võro, Estonian, and Finnish are close or overlapping in the input embeddings, with the other embeddings being language-specific. In the middle layers, the embeddings become more language-agnostic. In the later layers, each language forms a separate cluster. t-SNE plots of Layer 16 embeddings at the end of different training stages suggest representations becoming more language-agnostic as the training progresses.

```
<|system|>
Translate the following {src_lang} text into
{tgt_lang}.
<|user|>
{src_text}
<|assistant|>
{tgt_text}</s>
```

Figure 6: Translation-tuning data format based on Figure 5.

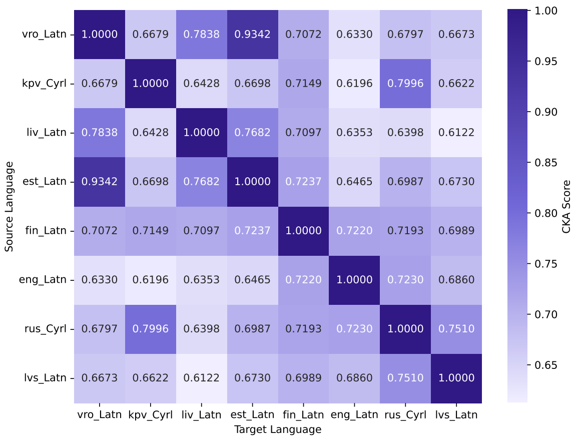

Figure 7: Llama-SMUGRI (Stage 2 + parallel) CKA scores (Kornblith et al., 2019) of mean-pooled layer 16 embeddings.

J Do We Still Need Human Translations?

We evaluate the applicability of the best proprietary LLMs for creating an MT-Bench-like evaluation dataset for XLR Finno-Ugric languages by asking the models to translate English or Estonian prompts from MT-bench-SMUGRI to the target languages. Table 20 shows that the best OpenAI models have not yet learned to translate to XLR Finno-Ugric languages. From that it can be concluded that they could also not generate synthetic data for our target languages with sufficiently good quality.

	EST→VRO	EST→LIV	ENG→KPV
gpt-4o-mini-2024-07-18	9.3	5	4.6
gpt-4o-2024-08-06	4.5	5.6	4.2
gpt-4-turbo-2024-04-09	18.9	5.9	3.6
Neurotõlge	24.7	21.4	**31.7**
Llama-SMUGRI-translate	**26.4**	**25.3**	19.1

Table 20: BLEU scores of translating MT-bench-SMUGRI.

We then compare translations from the best translation models with human translations using pairwise comparison where we ask human annotators to choose a better translation from the two alternatives (ties allowed). We gather 3 sets of annotatios for Livonian, 2 sets for Võro and 1 for Komi. Figure 8 shows that Komi and Livonian speakers mostly prefer human translations over machine translated data, however, Võro speaker prefer surprisingly often machine translated data suggesting a good quality of Võro machine translation. The average agreement between the pairs of Livonian annotations were 67.5% while between Võro annotations 42.5%.

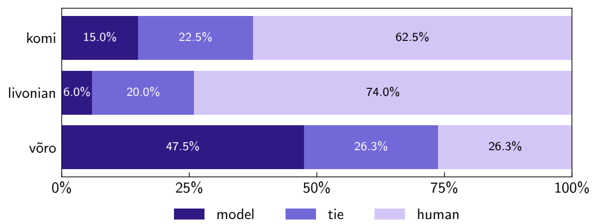

Figure 8: Preferred translations according to pairwise comparison.

PRE-TRAINED MODELS

FLORES-SMUGRI

```
{src_lang}: {src}\n{tgt_lang}:
```

BELEBELE-SMUGRI (prompt from Bandarkar et al., 2023)

```
P: {passage}\nQ: {question}\nA: {answer_1}\nB: {answer_2}\nC: {answer_3}\nD: {answer_4}\nAnswer:
```

SIB-SMUGRI (prompt from Lin et al., 2024; Csaki et al., 2024)

```
Topic Classification: science/technology, travel, politics, sports, health, entertainment, geography.\n\nThe label of [{sentence}] is
```

INSTRUCTION-TUNED MODELS

FLORES-SMUGRI

```
Translate the following {src_lang} text into {tgt_lang}.\n{src}
```

BELEBELE-SMUGRI (prompt from Bandarkar et al., 2023)

```
Given the following passage, query, and answer choices, output the letter corresponding to the correct answer.\n###\nPassage:\n{passage}\n###\nQuery:\n{query}\n###\nChoices:\n(A) {answer_1}\n(B) {answer_2}\n(C) {answer_3}\n(D) {answer_4}\n###\nAnswer:
```

SIB-SMUGRI (prompt from Adelani et al., 2024)

```
Is this a piece of news regarding science/technology, travel, politics, sports, health, entertainment, or geography?\n{sentence}
```

GPT-4-turbo as a Fallback Evaluator (for SIB-200)

```
Your task is to verify if the given model output classifies a text correctly. Answers in other languages should be allowed if they meaning matches closely with the expected class (e.g. \See on teadusuudis\" is correct when expected output is \"science/technology\"). If the model output does not choose a specific class, then the output is incorrect.\n\n### Expected class: {expected_answer}\n\n### Model output: {output_text}\n\n### Respond with Yes or No:
```

Figure 9: Prompts used for evaluation. Pre-trained models were evaluated with Language Model Evaluation Harness (Gao et al., 2023).

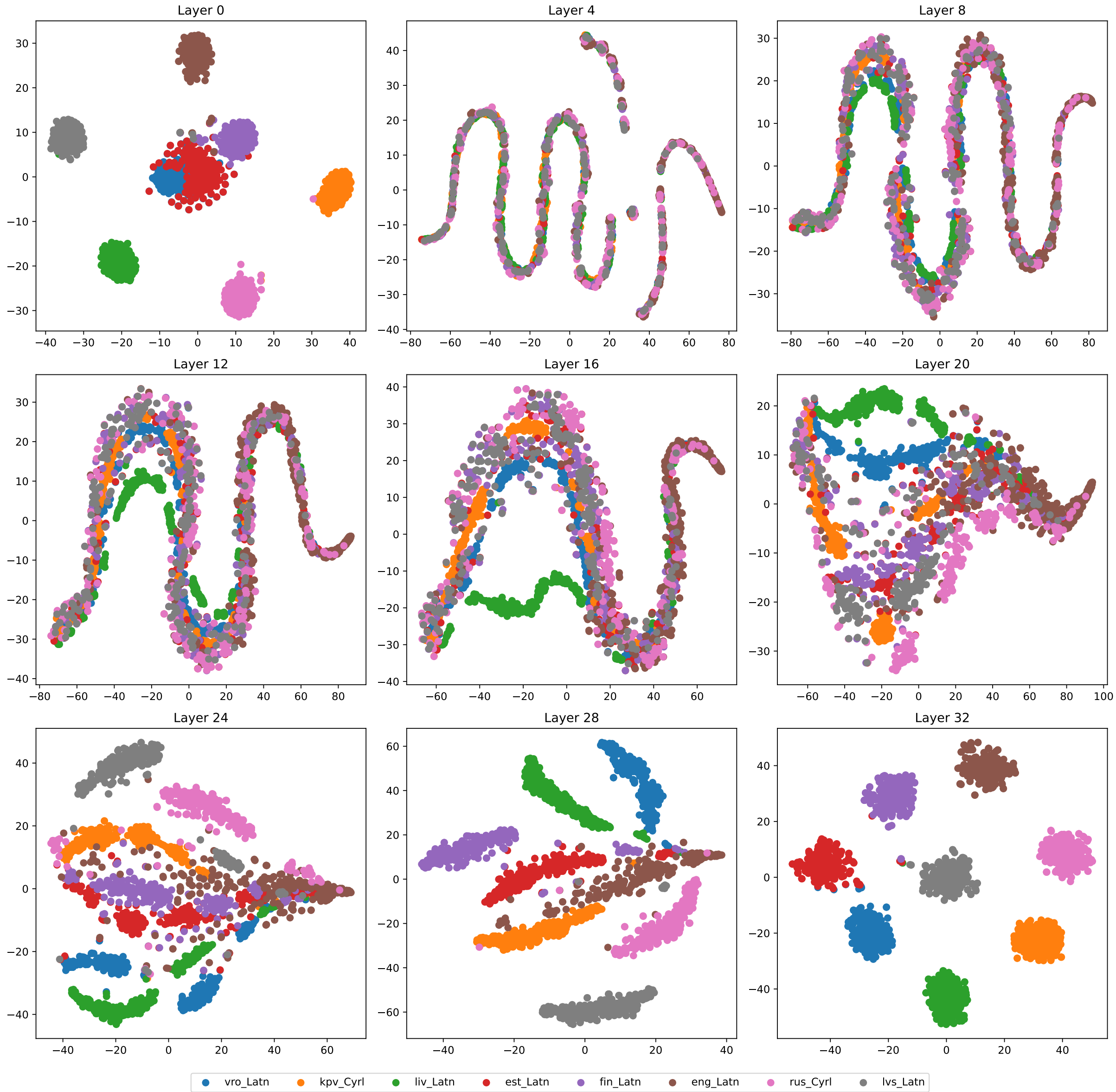

Figure 10: Llama-SMUGRI (Stage 2 + parallel) t-SNE (van der Maaten and Hinton, 2008) plots of mean-pooled embeddings. Layer 0 is the output of the embedding layer.

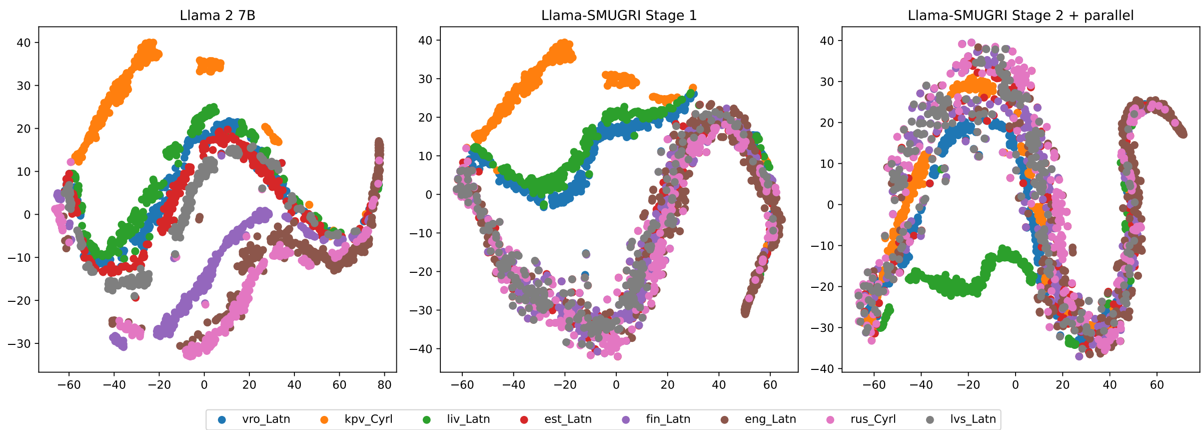

Figure 11: t-SNE (van der Maaten and Hinton, 2008) plots of mean-pooled 16th layer embeddings in different stages of model development.