LLMs for Extremely Low-Resource Finno-Ugric Languages

Anonymous ACL submission

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

The development of large language models (LLMs; OpenAI et al., 2024; Touvron et al., 2023, etc) has primarily focused on widely spoken languages, leaving low-resource languages with minimal support. Potential causes for this are not only extremely limited amounts of training data but also the lack of evaluation benchmarks and low numbers of speakers. Therefore, merely developing training methods for low-resource settings is insufficient for low-resource languages to benefit. Initiative from the community is also needed to draw attention to the lack of NLP tools for their languages and to support the creation of the tools, datasets and benchmarks (Orife et al., 2020).

In this work, we focus on LLM development for low-resource Finno-Ugric languages (SMUGRI[1]). Aside from the progress in machine translation (Yankovskaya et al., 2023; Rikters et al., 2022; Tars et al., 2022, 2021), most of these languages

	class	script	code	speakers
Livonian	0	Latin	liv	40
Võro	1	Latin	vro	100K
Komi	1	Cyrillic	kpv	100K
Finno-Ugric support languages				
Estonian	3	Latin	et	1.1M
Finnish	4	Latin	fi	5M

Table 1: Language statistics of Finno-Ugric languages covered in this work. The first column (*class*) is a language classification according to Joshi et al. (2020) indicating the amount of resources available for that language and ranging from 0 to 5.

have not yet benefited from the rapid advancement of NLP technologies, although the advantages of pretraining models have led to methods that achieve high-quality results even in limited-resource settings. We cover the full pipeline of LLM creation for three low-resource Finno-Ugric languages: Võro, Livonian, and Komi (see Figure 1). We report our experience with every step, including collecting and processing training data, designing training methodologies and training models, creating benchmarks to evaluate the resulting models and running manual evaluation. Thus our contributions are:

1. a study and experimental results of pretraining and instruction-tuning strategies applicable in low-resource settings, resulting in open-source, multilingual base and instruction-tuned models;
2. extension of the automatic evaluation benchmarks Belebele (Bandarkar et al., 2023) and SIB-200 (Adelani et al., 2024) to Komi, Livonian, and Võro;
3. creation and release of a novel multi-turn conversational benchmark, titled SMUGRI-MT-BENCH; using it to conduct a human evaluation.

[1] *Finno-Ugric* translates to Estonian as *soome-ugri*, to Finnish as *suomalais-ugrilaiset*, to Võro as *soomõ-ugri*, and to Livonian as *sūomõ-ugrõ*, hence we refer to it as SMUGRI.

2 Background and Related Work

2.1 Low-resource Finno-Ugric Languages

While all the languages covered in this paper belong to the same Finno-Ugric language group, they vary in terms of scripts and resources available (see Table 1). Regarding resources, we refer not only to the size of the available corpora but also to the ease or difficulty of finding language speakers who can help create benchmarks and evaluate model outputs. For instance, there is only around 40 speakers of Livonian (Ernštreits, 2019).

Võro and Livonian belong to the smaller Balto-Finnic language group, spoken around the Baltic Sea. We will utilise the two higher-resourced languages in the group, Finnish and Estonian, as sources of cross-lingual transfer (or "support languages") during pretraining to alleviate data scarcity. Due to the geographical location of the speakers, Livonian has been heavily influenced by Latvian, and Komi by Russian. Additionally, Komi speakers often know Russian, and Livonian speakers often know Latvian. Therefore, we will also use Latvian and Russian as supporting languages during pretraining.

2.2 Multilingual LLMs

Multilingual LLMs are a widely explored for expanding language coverage of LLMs. Traditional methods involve training models from scratch (Luukkonen et al., 2024, 2023; Wei et al., 2023; Kudugunta et al., 2023). However, the approach of adapting pre-trained English-centric models to other languages by continued pre-training has also shown promising results on 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). The closest work to ours is from Kuulmets et al. (2024), who adapted Llama-2 7B to Estonian.

The development of multilingual LLMs involves techniques that often improve the model's quality. For example, one common practice is incorporating parallel data into the pre-training phase (Luukkonen et al., 2024; Owen et al., 2024; Wei et al., 2023). Another technique is curriculum learning applied by Wei et al. (2023).

2.3 Instruction Tuning

Previous works have also explored a variety of techniques for using cross-lingual instruction-tuning (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 then translated and code-switched. Mixing 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) also find that using a diverse set of instructions in English can increase performance in Estonian tasks.

2.4 Evaluation

Common approaches for evaluating the multilingual capabilities of generative LLMs include using existing cross-lingual benchmarks (Ahuja et al., 2023a,b) or translating English benchmarks into target languages, either through machine translation (Lai et al., 2023) or manually (Shi et al., 2022). However, extending the evaluation of conversational capabilities to other languages is more complex as the gold standard involves using human annotators (Touvron et al., 2023). Human annotators are required for both the recently popularized method of ranking models using the Elo rating system (Zheng et al., 2024) and the more 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 related work explores LLM-judges as potential replacements for human annotators (Zheng et al., 2024; Kim et al., 2023, 2024). While it has been shown that strong LLMs can substitute human annotators for English, it is unclear to what extent such capabilities extend to non-English languages. Hada et al. (2024) study this question across eight very high and high resource non-English languages, finding a bias in GPT-4-based evaluators towards higher scores, which highlights the need for calibration. To the best of our knowledge, the behaviour of LLM-judges on low-resource languages, including Finno-Ugric languages, has not been systematically studied.

3 Experiments

3.1 Training datasets

We utilize CulturaX (Nguyen et al., 2023) to continue pre-training the base model on high-resource languages. The Komi documents are sourced from

FU-LAB's Komi corpus[2]. The Livonian dataset consists of sentence-level data from Rikters et al. (2022). Our Võro dataset is compiled from various pre-existing corpora as well as data we have crawled. A more detailed overview can be found in Appendix G.

The parallel data used during pretraining is collected for translation directions that involve Võro, Komi, and Livonian and is sourced from Yankovskaya et al. (2023); Rikters et al. (2022); Tars et al. (2022, 2021) (see Table 16).

Lang	Characters	Docs	Sampled Characters			
			Stage 1	Stage 2	Total	Ratio
LIV	2.6M	no	-	10.3M	10.3M	4.00
VRO	14.0M	yes	-	56.1M	56.1M	4.00
KPV	578.9M	yes	-	1.4B	1.4B	2.48
LV	27.8B	yes	3.0B	300.0M	3.3B	0.12
ET	32.6B	yes	8.2B	300.0M	8.5B	0.26
FI	114.0B	yes	7.6B	300.0M	7.9B	0.07
RU	>1T	yes	2.7B	300.0M	3.0B	<0.01
EN	>1T	yes	2.7B	300.0M	3.0B	<0.01

Table 2: Training dataset composition. Docs - *yes* if the dataset is document level, *no* if sentence-level.

3.2 Continued Pre-training

We take the approach of adapting English-centric Llama-2 7B (Touvron et al., 2023) to our chosen target languages through full fine-tuning. Due to computational budget limitations, we opt for a two-stage training approach where we first continue pre-training on high-resource Finno-Ugric languages along with other related supporting languages (see §2.1) and only during the second phase teach the model low-resource target languages. The training hyperparameters are listed in Appendix D.

Stage 1: learning supporting languages. As a first step, we continue pre-training of Llama-2 7B (Touvron et al., 2023) on high-resource Finno-Ugric languages and supporting languages. We set the training budget at 10B tokens and sample documents from Estonian, Finnish, English, Latvian, and Russian Culturax with 32%, 32%, 12%, 12%, 12% probability of choosing the document from the respective language.

Stage 2: learning low-resource Finno-Ugric languages. The second stage of continued pre-training aims to enhance understanding and generative capabilities for low-resource languages. We employ a character-based budget to achieve 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. We chose the N value according to held-out validation set perplexity (see Table 12 in Appendix E).

We also investigate the role of parallel translation examples by formatting them into various templates and using them during this stage of pre-training. Specifically, we add translation examples up to 1% of the Stage 2 character budget (30M) and use Unimax N=1 to balance the budget between language pairs (see Table 16). We refer to it as **Stage 2 + parallel**.

3.3 Instruction Tuning

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) in these languages. Kuulmets et al. (2024) showed improved cross-lingual knowledge transfer from using an additional set of high-quality English instructions. We sample 5,000 such instructions from FLAN-V2 (Longpre et al., 2023) TULU mixture (Wang et al., 2023) and 20,000 examples from Alpaca-GPT-4 (Peng et al., 2023). Additional 20,000 Estonian instructions are sampled from Alpaca-est (Kuulmets et al., 2024). We refer to the data mixture of the aforementioned sources as `Inst`.

We create instruction datasets for the target languages by translating 1000 examples per language from Alpaca-style instruction datasets into low-resource Finno-Ugric languages. An external system Neurotõlge[3] (Yankovskaya et al., 2023) is used as a translator. While Võro and Livonian are translated from Alpaca-est (Kuulmets et al., 2024), Komi is obtained by first translating Alpaca-GPT-4 (Peng et al., 2023) with `GPT-3.5-turbo` into Russian and then with Neurotõlge into Komi. We refer to this dataset as `TrAlpaca`.

To investigate a scenario where a translation model is not available, we additionally explore handling low-resource translation directions with our LLM tuned for the translation task (discussed in §I). We found that LLM-based models unpredictably leave sentences untranslated. Therefore,

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

[3] https://neurotolge.ee/

we removed examples where the BLEU score is greater than 70 between the original and translated text. This also removes some valid examples, since sometimes a text can be the same in both languages. We refer to this dataset as `LLMTrAlpaca`.

Finally, we explore augmenting the general instructions with translation task instructions to/from Võro, Livonian, and Komi – 250 examples per direction from sources listed in §3.1 (see Table 16). We refer to this dataset as `TrInst`.

Instruction tuning examples are formatted into a multi-turn conversational format (see Figure 6).

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

Table 3: Test benchmarks created or extended for Komi, Võro, and Livonian. SIB-SMUGRI additionally includes 30 validation examples.

3.4 Training on Parallel Data

One potential bottleneck of our approach is the low quality of machine translation when translating instructions to the low-resource SMUGRI languages. However, adapting general-purpose LLMs to the machine translation task yields competitive results with dedicated MT systems (Xu et al., 2023; Kuulmets et al., 2024). Therefore, we fine-tune our base model on available translation data (see §3.1) by sampling up to 100,000 sentence pairs from each language pair (see Table 16). We call this configuration `TrTuning`.

In preliminary experiments, we noticed that the model sometimes struggles with multi-line or multi-sentence inputs, which is crucial for translating instructions as they include entire texts from Alpaca-style examples for accurate translation. To address this, we trained a model where 50% of the training data consists of 2–6 concatenated sentences, while the rest are single sentences. We refer to this configuration as`TrTuningConcat`.

4 Benchmarks

4.1 Automatic Evaluation

Our automatic evaluation heavily relies on FLORES-200 benchmark (NLLB-Team et al., 2022; Goyal et al., 2022), an extensive multilingual dataset designed for evaluating machine translation. Notably, the first 250 sentences have already been translated into several Finno-Ugric languages by Yankovskaya et al. (2023). Building on FLORES-200 we extend benchmarks like topic classification benchmark SIB-200 (Adelani et al., 2024) and multiple-choice QA dataset Belebele (Bandarkar et al., 2023) to Livonian, Võro and Komi. We align Finno-Ugric translations by Yankovskaya et al. (2023) with sentences and topics in SIB-200 and paragraphs, questions and answers in Belebele. To ensure the high quality of the benchmark, we manually translate questions and answer choices into the target languages since FLORES-200 does not contain them. Table 3 shows the details of all evaluation benchmarks.

We also report byte-level perplexity of base models on held-out validation data, sampled from the same corpora as the training examples (see Table 13). Further evaluation details will be described in Appendix F.

4.2 Multi-turn Conversational Benchmark

4.2.1 Requirements and Limitations

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 the real-life use cases of low-resource Finno-Ugric LLM. Consequently, our evaluation benchmark should consist of user prompts similar to real-life queries. Another benefit of real-life data is that it helps quickly reveal the model's usefulness in practical scenarios, which standard NLP benchmarks typically do not cover. It also helps to identify potential weaknesses of the model in real-life situations.

However, usefulness is a vague term as it depends on the specific use case of the user and is, therefore, difficult to measure. During the training of LLM-based assistants, usefulness is indirectly optimized with RLHF (Ouyang et al., 2022) that rewards model outputs with high helpfulness and safety scores as determined by the reward model (Touvron et al., 2023). During evaluation, the models are ranked using a pairwise comparison, where human annotators are asked to select a better response (more helpful, safe, and honest) from two model responses (Touvron et al., 2023).

One danger of pairwise comparison is the potential for many ties between the two models. This could indicate that the models have very similar output quality or that the evaluation prompts are too trivial to differentiate between them. 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.

The chosen low-resource Finno-Ugric languages impose a set of limitations on benchmarking LLMs (see §2.1). Firstly, creating high-quality benchmarks for these languages is tricky. They cannot be obtained through machine translation from other languages, as the machine translation systems for these languages are 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.

Secondly, a key requirements 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. Fnding human annotators who are both speakers of the target language and knowledgeable enough to judge answers to expert-level questions is a significant challenge.

Taking into account the expectations and limitations set and discussed above, we list the requirements for the benchmark of low-resource Finno-Ugric languages:

- translating it to a new language should be feasible both content-wise and time-wise for non-professional translators;
- answering questions should not require expert knowledge, as expert annotators can not be used;
- questions should cover real-life usage scenarios to reflect real-life usefulness;
- questions should be challenging enough for LLMs to differentiate the models accurately.

4.2.2 Initial Dataset Collection

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 §4.2.1. Finally, we remove 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.

4.2.3 Finalising the Dataset

	general	reasoning	maths	writing	**total**
questions	20	20	20	20	80
follow-ups	14	8	11	9	42

Table 4: Statistics of human evaluation dataset.

We organize conversations into four categories: math, reasoning, writing, and general. As we wanted the final dataset to consist of 80 questions (similar to Zheng et al. (2024)) — 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 (see Appendix A for more details). 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 statistics of the dataset are shown in Figure 4. The final dataset was translated to the target languages by non-professional translators who could speak the language at the native level. The translators were asked to preserve any informality of the text in the translations, e.g. missing uppercase and punctuation.

5 Results

5.1 Pre-training

Stage 1 pre-training on supporting high-resource languages demonstrates visible improvements in

Model	SIB-SMUGRI 5-shot, acc			BELEBELE-SMUGRI 3-shot, acc			FLORES-SMUGRI 5-shot, BLEU			byte-PPL		
	VRO	LIV	KPV	VRO	LIV	KPV	ET-VRO	ET-LIV	RU-KPV	VRO	LIV	KPV
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	10.0	4.0	1.7	3.3548	12.1081	3.1959
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	10.5	4.4	2.5	6.1528	14.8055	3.1198
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	10.3	3.6	2.4	3.4895	11.4210	3.1341
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	22.1	3.5	12.3	2.1885	3.8351	1.4055
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	23.7	4.5	14.5	2.1837	3.7615	1.4050

Table 5: Pre-training results for low-resource Finno-Ugric languages. Standard errors are reported for the scores (*score* ± *stderr*). *Stage 2 + parallel* incorporates additional parallel translation data into training. For comparison, we report GPT-models and Llammas-base (Kuulmets et al., 2024).

SIB-200 and perplexity (Võro, Livonian, Komi) compared to the Llama-2-7B model (see Stage 1 in Table 5). This indicates that there are benefits from similar languages even when low-resource SMUGRI languages are not directly seen during training.

Stage 2 pre-training focusing on low-resource Finno-Ugric languages further improves both perplexity and FLORES-200 scores, suggesting the model has learned generative capabilities for SMUGRI languages. The performance gains on the SIB-200 benchmark are modest for Komi and Võro, and there is a decrease for Livonian. Belebele scores remain unchanged from those of Llama-2-7B, except Võro, which shows improvement.

Incorporating parallel translation data (1% of the training budget) into the stage 2 pre-training yields minimal improvements in benchmark performance and byte-perplexity (Stage 2 + parallel in Table 5). Either the impact of including this data is minimal, or our benchmarks are too limited to show it sufficiently. Given a slightly positive impact of the parallel data, we will use `Stage 2 + parallel` as a foundation for subsequent instruction-tuning.

It is possible that the available benchmarks are not ideal at discriminating between models at this stage. This could be the case for multiple reasons. It is possible that the model can choose the correct answer from clues in the text that do not require understanding the language well. Furthermore, judging by the low scores, Belebele questions might be sometimes too difficult for the models to answer. Finally, our benchmarks are very small and the standard errors are too high to make confident choices about fine-grained model differences. Therefore these benchmarks are only suitable to make more general claims about the models' capabilites.

5.2 Instruction-Tuned Models

Looking at the scores of commercial systems in Table 6, it is visible that they have at least some level of understanding of Võro, Livonian, and Komi. Judging by benchmark scores, they seem to understand Võro and Livonian the best. A possible explanation is that the languages are very similar to Estonian - an average Estonian speaker will understand most of a Võro text and some of a Livonian text but not much Komi since it is more distant and in a different script. The scores of these languages' benchmarks on GPT-4-Turbo and GPT-3.5-Turbo are primarily in this order as well. For example, since GPT-4-turbo achieves 92% accuracy on Belebele Estonian, it is not surprising that Võro also achieves a high score.

Our models show comparable performance to GPT-3.5-Turbo on Võro and Livonian, and slightly better performance on Komi. However, GPT-4-Turbo significantly outperforms our models on Võro and matches our performance on Livonian and Komi.

On the SIB benchmark, a similar pattern emerges: our models surpass GPT-4-Turbo on Livonian and Komi but fall short on Võro. Meanwhile, GPT-3.5-Turbo consistently scores lower across all low-resource languages.

When examining our trained models, the different instruction-tuning strategies yield similar results. Due to the small size of our benchmarks and the resulting high standard errors, we cannot draw definitive conclusions about the best strategy.

LLM-translated instructions. Automatic metrics show that instructions translated with our translation-tuned LLM provide similar results to translations obtained with an external system (Neurotõlge). Unfortunately, there is not enough confidence or clarity in the results to indicate a clear preference in one method or another. These results demonstrate that even when external translation systems are unavailable the translation-tuned LLM can be used.

Does augmenting the data with translation

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
Ours:						
`Inst`	42.5 ± 4.4	30.7 ± 4.1	**44.1 ± 4.4**	86.4 ± 3.1	79.2 ± 3.6	**88.8 ± 2.8**
`Inst+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
`Inst+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
`Inst+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
`Inst+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*).

instructions improve the results? Incorporating a small amount of translation instructions (250 for each Võro, Komi, and Livonian direction) does not yield a clear and consistent improvement across discriminative benchmarks (see Table 6). On the other hand we see a substantial increase in the translation benchmark in Table 7.

Translation abilities. Judging language generation abilities by the FLORES translation benchmark, results in Table 7 demonstrate that GPT-models can translate from Estonian to Võro quite well. This might indicate that they had Võro in their training data. The BLEU scores of Livonian and Komi are very low, suggesting almost nonexistent translation abilities. Our LLMs that have not seen translation examples as part of the instruction-tuning can not translate to the low-resource SMUGRI languages. However, they are successful in translating in the opposite direction, even outperforming GPT-models for Komi. A closer look reveals that they copy the high-resource language sentences to the output. When the `TrAlpaca` and `LLMTrAlpaca` were added, we also observed that the models often copied the source text in these languages to the output when asked to translate, resulting in lower scores. This can be addressed by including a small amount of translation data during instruction-tuning or possibly few-shot prompting.

5.3 Translation-tuning

We compare our LLM-based translation models to Neurotõlge, which supports low-resource Finno-Ugric languages. Our translation-tuned models outperform Neurotõlge in the VRO-ET and ET-VRO translation directions (see Table 7). For ET-LIV and RU-KPV, our models achieve performance on par with Neurotõlge. However, when translating from low-resource to high-resource languages (with the exception of Võro), our models fall short.

In addition to regular fine-tuning with sentence pairs, we concatenate sentence examples into larger sequences to enhance the model's ability to translate longer texts (`TrTunedConcat`). This approach is particularly useful for translating instructions. Notably, the concatenation of examples does not compromise translation quality and increases training effectiveness in a similar way to packing.

5.4 Human evaluation

We pick 3 instruction-tuned models for human evaluation: `TrAlpaca`, `LLMTrAlpaca+TrInst` and `TrAlpaca+TrInst`. As a baseline we use GPT-3.5-turbo, which can be freely accessed via a chat-interface[4]. For each target language, we create a survey where participants were asked to rate the helpfulness of the answer from a randomly chosen model in 5-point Likert scale. Additionally, we ask participants to rate how natural the answer sounds in the target language as Kuulmets et al. (2024) reports that model outputs tend to sound unnatural in the target language. The surveys were shared within the communities of target language speakers through social media and by directly reaching out to the language speakers (see Appendix C for the screenshot of the survey). We did not collect any personal data from the respondents.

In addition to Võro, Liivi and Komi we gather and present human evaluation data also for Estonian as it is closely related to Võro and Liivi (see §2.1) but at the same time is well-supported by GPT-3.5-Turbo (Kuulmets et al., 2024). This gives us a meaningful anchor point to compare our human evaluation results against.

The results reveal that our models underperform in terms of helpfulness compared to GPT-3.5-Turbo in Estonian, which is not surprising (Kuulmets et al., 2024). For Võro, the disparity persists, with our models still trailing behind. In the case of Võro

[4]https://chatgpt.com/

Model	VRO-ET	ET-VRO	LIV-ET	ET-LIV	KPV-RU	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**
Instruction-tuned:						
`Inst`	41.9	10.7	11.1	4.6	21.4	3.0
`Inst+LLMTrAlpaca`	23	10.8	9.2	4.6	13.5	2.9
`Inst+TrAlpaca`	16.8	10.6	9.7	4.7	17	3.2
`Inst+LLMTrAlpaca+TrInst`	47.7	21.2	20.6	6.2	20.9	16.4
`Inst+TrAlpaca+TrInst`	45.3	19.1	19.9	5.5	21.4	15.2
Trainslation-tuned:						
`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 7: BLEU scores on FLORES-SMUGRI (0-shot). Translations are generated with beam size 4 for our models.

	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 8: Human evaluation data collection statistics.

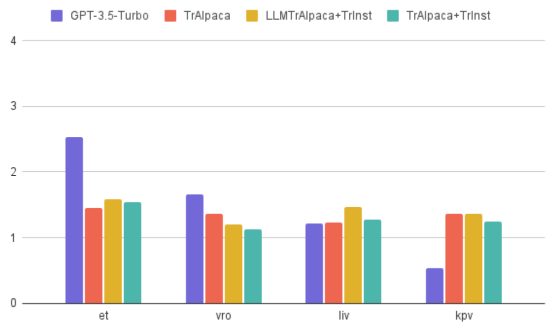


Figure 1: Human evaluation scores for helpfulness.

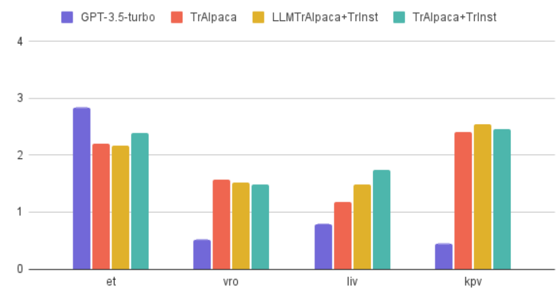


Figure 2: Human evaluation scores for naturalness.

and Livonian, the helpfulness scores of our models and GPT-3.5-turbo are comparable, whereas, for Komi, our system exceeds the commercial baseline. Although it is likely that variations in annotator expectations for different languages affect individual language results, it is noteworthy that our models consistently achieve similar helpfulness scores across various languages.

In terms of the naturalness of responses, GPT-3.5-Turbo performs slightly better for Estonian; however, our models exhibit greater naturalness in all other languages, with the difference being particularly pronounced for Komi.

Category-wise comparisons (see Appendix B) indicate that the scores of GPT-3.5-turbo are inflated by *maths* and *reasoning* examples, where our models lag in helpfulness. However, our models perform comparably in the *general* and *writing* categories. Notably, in Komi, our models outperform GPT-3.5-Turbo in *general* and *writing* tasks while achieving similar scores in *maths* and *reasoning* tasks.

When comparing models trained by us, no clear ranking emerges, reinforcing the observations from automatic benchmarks that incorporating translation instructions does not yield definitive benefits and that there is no significant difference between using LLM-translated instructions and those translated by an external system.

6 Conclusion

We adopted a comprehensive approach from data collection to instruction-tuning and human evaluation for three low-resource Finno-Ugric languages: Võro, Livonian, and Komi. Our contributions include an exploration of pre-training and instruction-tuning strategies, resulting in open-source multilingual base and instruction-tuned models for these languages. We extend the automatic evaluation benchmarks Belebele and SIB-200 to Komi, Livonian, and Võro and release a novel multi-turn conversational benchmark, SMUGRI-MT-BENCH. Human evaluation using SMUGRI-MT-BENCH shows our models surpass `GPT-3.5-Turbo` in naturalness and achieve higher helpfulness for Komi, with similar 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.

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.

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.

Valts Ernštreits. 2019. Electronical resources for Livonian. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 184–191, Tartu, Estonia. 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*.

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.

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.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–

machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

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.

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*.

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*.

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.

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, 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 Detecting Difficult Question with LLMs

Zheng et al. (2023) uses GPT-3.5-Turbo to classify whether the prompt is a good prompt for benchmarking. They find the technique effective for filtering out trivial or ambiguous user prompts. We use the same prompt as Zheng et al. (2023) to assess the difficulty of a prompt. To measure whether the scores are effective, we plot the difficulty scores against answer grades, obtained with GPT-4. The plots in Figure 3 reveal somewhat surprisingly that answers to easier questions tend to get slightly lower grades from GPT-4, indicating that GPT-4 might underestimate the difficulty of a question. This is especially evident in weaker LMs such as `Llammas`. We hypothesize that our differing results from Zheng et al. (2023) may be due to our initial dataset being handpicked, which likely included more challenging questions.

B Usefulness Scores by Categories

The usefulness scores by categories from human evaluation are shown in 4

C Collecting Data for Human Evaluation

The screenshot of the survey is shown in Figure 5. For Võro, Liivi, and Estonian, the instructions were given in Estonian, while for Komi, they were given in Russian.

D Training Details

The hyperparameters of pre-training stages 1 and 2 are listed in Table 9. The instruction-tuning and translation-tuning parameters are in Table 10. 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 11.

Parameter	Stage 1	Stage 2
updates	19073	-
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 9: 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 10: 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 11: GPU-hours elapsed for training the models.

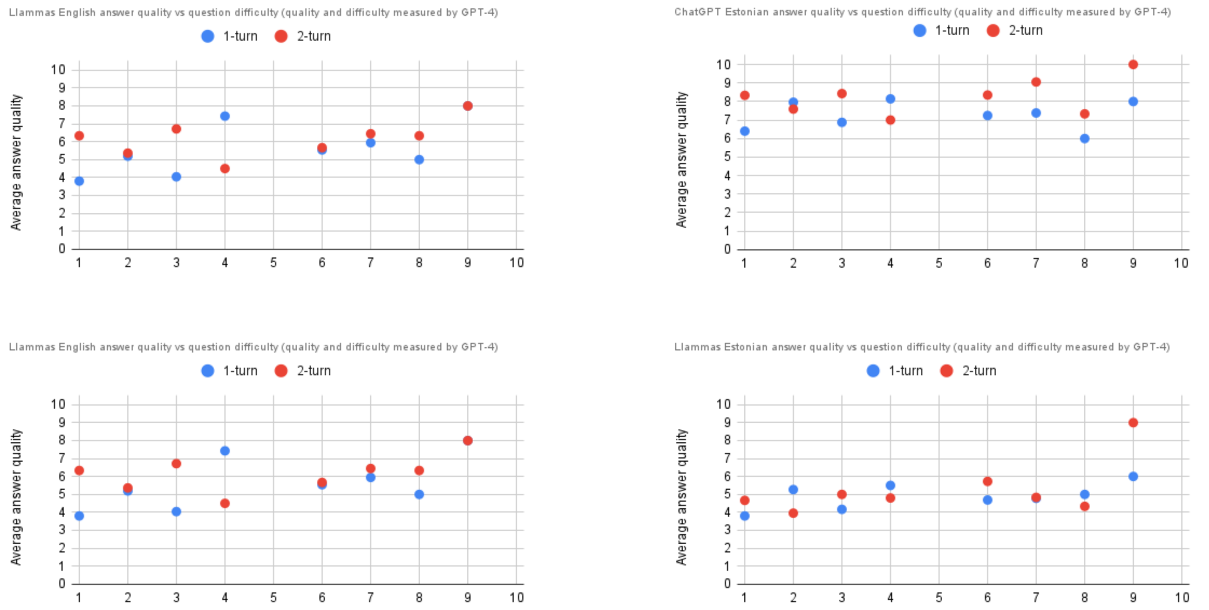


Figure 3: Plotting the difficulty of a question (assessed by GPT-4) against the quality of an answer (assessed by GPT-4).

E 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 12).

Unimax N	byte-PPL		
	VRO	LIV	KPV
N=1	2.3072	4.1986	1.4508
N=4	**2.1885**	**3.8351**	**1.4055**
N=8	2.5983	4.725	1.4159

Table 12: The effect of Unimax N (max data repeat epochs) on held-out validation set byte perplexity.

F Evaluation details

The base models are evaluated with `lm-evaluation-harness` (Gao et al., 2023), and bootstrap standard errors are reported.

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.

`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[5] (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, otherwise they are documents.

G Võro Data Collection

We collect Võro data from Võro language Wikipedia dump (Foundation), Corpus of Fiction in Võro and Seto languages[6], Additionally, we crawled Võro language newspaper articles from *Uma Leht*[7]. 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[8] which is also in Seto dialect. The collected Võro dataset composition is shown in Table 14.

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

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

[7]https://umaleht.ee/

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

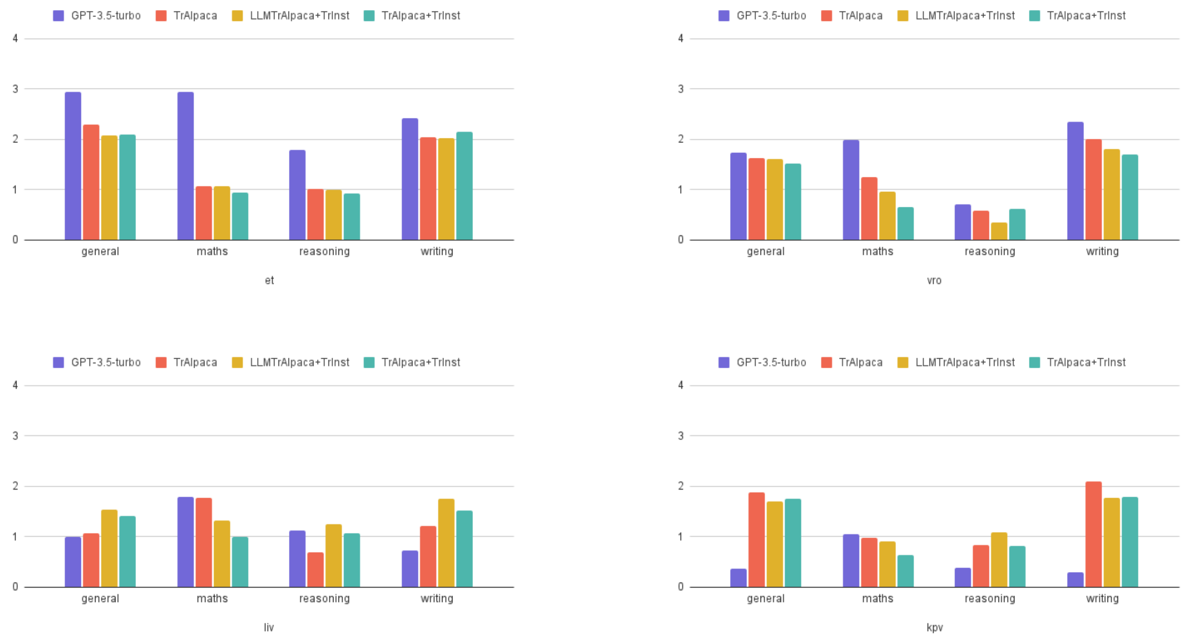
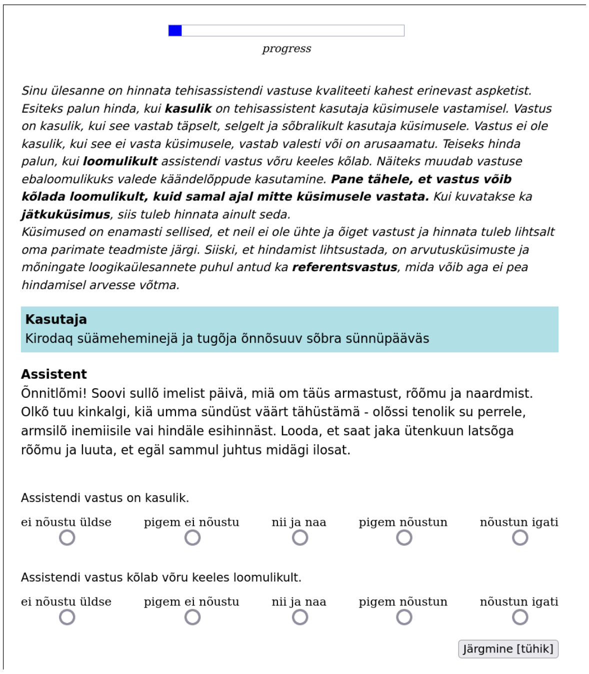


Figure 4: Usefulness by different categories.



progress

Sinu ülesanne on hinnata tehisassistendi vastuse kvaliteeti kahest erinevast aspketist. Esiteks palun hinda, kui ***kasulik*** *on tehisassistent kasutaja küsimusele vastamisel. Vastus on kasulik, kui see vastab täpselt, selgelt ja sõbralikult kasutaja küsimusele. Vastus ei ole kasulik, kui see ei vasta küsimusele, vastab valesti või on arusaamatu. Teiseks hinda palun, kui* ***loomulikult*** *assistendi vastus võru keeles kõlab. Näiteks muudab vastuse ebaloomulikuks valede käändelõppude kasutamine.* ***Pane tähele, et vastus võib kõlada loomulikult, kuid samal ajal mitte küsimusele vastata.*** *Kui kuvatakse ka* ***jätkuküsimus****, siis tuleb hinnata ainult seda.*
Küsimused on enamasti sellised, et neil ei ole ühte ja õiget vastust ja hinnata tuleb lihtsalt oma parimate teadmiste järgi. Siiski, et hindamist lihtsustada, on arvutusküsimuste ja mõningate loogikaülesannete puhul antud ka ***referentsvastus****, mida võib aga ei pea hindamisel arvesse võtma.*

Kasutaja
Kirodaq süämeheminejä ja tugõja õnnõsuuv sõbra sünnüpääväs

Assistent
Õnnitlõmi! Soovi sullõ imelist päivä, miä om täüs armastust, rõõmu ja naardmist. Olkõ tuu kinkalgi, kiä umma sündüst väärt tähüstämä - olõssi tenolik su perrele, armsilõ inemiisile vai hindäle esihinnäst. Looda, et saat jaka ütenkuun latsõga rõõmu ja luuta, et egäl sammul juhtus midägi ilosat.

Assistendi vastus on kasulik.

ei nõustu üldse | pigem ei nõustu | nii ja naa | pigem nõustun | nõustun igati

Assistendi vastus kõlab võru keeles loomulikult.

ei nõustu üldse | pigem ei nõustu | nii ja naa | pigem nõustun | nõustun igati

Järgmine [tühik]

Figure 5: Screenshot of the survey that was used to collect human annotations.

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

Table 14: Võro data composition by source.

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

Figure 6: Chat format following Wang et al. (2023) and Kuulmets et al. (2024). The model responds after `<|assistant|>`.

H Instruction-tuning details

The composition of our instruction-tuning dataset is listed in Table 15. Instructions are formated into a char-format shown in Figure 6. Translation data format is shown in Figure 7.

I Parallel data

Composition of the parallel data is shown in Table 16.

Dataset	LIV	VRO	KPV	ET	FI	EN	RU
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			
Tr-Alpaca (ours)	1000	1000	1000				
TOTAL	1,000	1000	1000	20000	747	32458	1104

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

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
Pre-training	28504	14212	11606	492	3835	7272	81487	7286	5018	159712

Table 16: Number of sentences of parallel translation data used in various configurations during training. In all cases, the language pair data is split in two so that, for example, in ET-LIV, 50% of the reported sentences are for ET→LIV and the other 50% for LIV→ET

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

Figure 7: Translation-tuning data format based on Figure 6.