Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer

Anonymous ACL submission

Abstract

This paper explores cost-efficient methods to adapt pretrained Large Language Models (LLMs) to new lower-resource languages, with a specific focus on Estonian. Leveraging the Llama 2 model, we investigate the impact of combining cross-lingual instruction-tuning with additional monolingual pretraining. Our results demonstrate that even a relatively small amount of additional monolingual pretraining followed by cross-lingual instruction-tuning significantly enhances results on Estonian. Furthermore, we showcase cross-lingual knowledge transfer from high-quality English instructions to Estonian, resulting in improvements in commonsense reasoning and multi-turn conversation capabilities. Our best model, named LLAMMAS, represents the first open-source instruction-following LLM for Estonian. Additionally, we publish Alpaca-est, the first general task instruction dataset for Estonia. These contributions mark the initial progress in the direction of developing open-source LLMs for Estonian.

1 Introduction

Instruction-tuning is a method for aligning large language models (LLMs) with human preferences (Ouyang et al., 2022; Mishra et al., 2022; Wei et al., 2021). However, the majority of instruction-tuning datasets and advancements focus on English. Moreover, to benefit from instruction tuning, a strong foundation model is needed but due to the extensive training training data required, such models are available only for a few languages.

To overcome the lack of a strong foundation model in the target language, one could try to elicit non-English abilities from English-centric LLMs through cross-lingual instruction-tuning. In this setup, instructions are given in both English and the target language, often including a translation task to directly stimulate the alignment (Ranaldi et al., 2023; Ranaldi and Pucci, 2023; Zhu et al., 2023). While empirical evidence indicates benefits from incorporating translation-following demonstrations into the training dataset, the best training strategy and its effectiveness with monolingual finetuning remain unclear.

In this paper, we investigate these aspects in the context of creating an instruction-following model for Estonian. We focus on a low-resource scenario where only a relatively small amount of monolingual data is available. By utilizing a novel general task instruction dataset, Alpaca-est, we examine the impact of combining monolingual pretraining with cross-lingual instruction-tuning using both general and translation task instructions. Our experiments with Llama 2 (Touvron et al., 2023b) demonstrate the benefits of translation task instructions when no monolingual data is available for additional pretraining. However, monolingual pretraining greatly diminishes the importance of the translation task.

Furthermore, we showcase that supplementing our instruction-tuning dataset consisting of Alpaca (Taori et al., 2023) and Alpaca-est with high-quality English instructions and English conversations further enhances results on Estonian through cross-lingual knowledge transfer. This is reflected in improved commonsense reasoning and the ability to engage in multi-turn conversations despite no Estonian conversations used during training. As a result, we present LLAMMAS - the first open-source instruction-following conversational LLM for Estonian that achieves competitive zero-shot performance on multiple tasks.

2 Related Work

2.1 Instruction Tuning

Instruction-tuning is a method for guiding pretrained LLMs to follow natural language instructions (Ouyang et al., 2022; Mishra et al., 2022; Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022; Wang et al., 2022b). For that purpose, both

human-written and synthetic instructions generated with LLMs have been shown to work remarkably well (Wang et al., 2022b, 2023b). One of the prerequisites for instruction-tuning is the availability of a strong pretrained language model which due to high training costs is the major limiting factor for many to contribute to the development of LLMs. Fortunately, over the last year, a few foundation models (Workshop et al., 2022; Touvron et al., 2023a,b; Jiang et al., 2023) have been publicly released which somewhat mitigates the issue. However, the models are mostly trained on English and perform poorly on other languages.

A common method of acquiring instruction data is using strong proprietary models such as GPT-4 for generating instructions (Taori et al., 2023; Chiang et al., 2023; Wang et al., 2022a). However, Gudibande et al. (2023) have shown that models trained on these generated datasets learn to imitate the style of strong LLMs but not necessarily the factuality.

2.2 Cross-lingual Instruction Tuning

Cross-lingual instruction tuning is a training method where the model is simultaneously instruction-tuned on instructions in multiple languages. Its goal is to strengthen cross-lingual semantic alignment in LLMs to make them understand and generate texts in a selected target language. In practice, it is one of the most cost-efficient ways to create instruction-following models for languages where data-heavy pretraining is not possible.

The approach has been explored, for example, by Zhu et al. (2023) and Ranaldi et al. (2023) who both use original and translated versions of Alpaca (Taori et al., 2023) dataset. Moreover, they both report additional benefits from supplementing the general task instruction datasets with translation task instructions. However, their approaches differ in the size of translation datasets. Zhu et al. (2023) use datasets that sometimes contain around 10 times more translation task instructions than general task instructions. Ranaldi et al. (2023) employ a translation task instruction dataset that contains only 20K instructions. Additionally, while Zhu et al. (2023) report benefits from using English to target language translations, Ranaldi et al. (2023) demonstrated that using both translation directions together is better than translating to only one direction.

Zhang et al. (2023a) propose to combine the task of strengthening cross-lingual semantic alignment and instruction-tuning via a multi-turn translation task. Zhang et al. (2023b) utilize the capabilities of LLMs to comprehend and execute instructions in a high-resource language by using that high-resource language as a pivot language during response generation for the target language.

2.3 Monolingual Pretraining

Another way to improve the ability of English-centric pretrained LLMs to understand and generate content in a target language is via continued pretraining on data in the target language. For example, Cui et al. (2023) continue pre-training LLaMA family models on a large-scale monolingual Chinese corpus before the instruction-tuning. Xu et al. (2023) show that continued pre-training with even a relatively small monolingual dataset can significantly improve the results of the translation instruction task. Moreover, they show that after continued pre-training only a small amount of high-quality parallel data is required to reach competent translation. Their analysis even discourages training with larger amounts as it leads to the dilution of the model's intrinsic knowledge.

3 Training Data

3.1 General Task Instructions

3.1.1 Alpacas

We combine the original Stanford Alpaca dataset (Taori et al., 2023) with an Estonian version of it which we create by ourselves. We refer to the combination of these two datasets as **Alpacas**.

Stanford Alpaca (Taori et al., 2023) A general task instruction dataset generated with Self-Instruct framework (Wang et al., 2023b). In our experiments we use the cleaned version[1] that consists of filtered Alpaca (Taori et al., 2023) instructions and GPT-4-LLM (Peng et al., 2023).

Alpaca-est Due to a lack of general task instruction data in Estonian, we generate an Estonian version of Alpaca. Following Taori et al. (2023), we first randomly sample from a set of Estonian seed instructions and use an LLM to generate new instructions based on the examples. Using `gpt-3.5-turbo-0613`[2], we generate a total of 52,006 instructions for Estonian. The seed instruction set consists of 90 translated examples from

[1] https://github.com/gururise/AlpacaDataCleaned
[2] https://platform.openai.com/docs/models

the original Alpaca seed set and 17 new instructions written by the authors. We make Alpaca-est publicly available[3].

3.1.2 High-Quality General Task Instructions

We supplement Alpacas with high-quality English instructions that are not obtained with synthetic data generation using OpenAI models. In our dataset creation, we take inspiration from (Wang et al., 2023a; Ivison et al., 2023). We use Open Assistant 1 (Köpf et al., 2023) top-scoring English-only path from each conversation tree. We also take 10,000 examples of both CoT and FLAN-2 (Chung et al., 2022) mixtures used in (Ivison et al., 2023). We refer to this high-quality mixture of data in short as **HQI**.

3.2 Translation Task Instructions

We create translation task instructions from relatively low-quality translation bitexts: CCMatrix (Schwenk et al., 2021b), WikiMatrix (Schwenk et al., 2021a), OpenSubtitles (Lison and Tiedemann, 2016), and Europarl (Tiedemann, 2012). We filter the data with OpusFilter (Aulamo et al., 2020) using long word, sentence length, source-target length-ratio, character score, language-ID, terminal punctuation, and non-zero numerals filters.

We use a setup in which 75% of instructions prompt translation from English to Estonian, and 25% prompt translation in the opposite direction. The goal of including a small amount of Estonian-English is to maintain the quality of English generation. We refer to this translation task instructions dataset as **TrTask**.

We supplement the relatively low-quality **TrTask** dataset with high-quality parallel data from WMT18 dev set (Bojar et al., 2018) and MTee (Tättar et al., 2022) held-out validation dataset. We refer to it as **HQTrTask**. In **HQTrTask** WMT18 dev set is given in a document-level format with documents exceeding 900 tokens split into multiple parts. To convert the translation examples to instructions we utilize 32 English and 13 Estonian prompt templates as Sanh et al. (2021) has demonstrated the importance of using a diverse set of prompts.

3.3 Pretraining Data

For pretraining, we use a subset of Estonian and English data from CulturaX (Nguyen et al., 2023) to make the base model more familiar with Estonian but not forget English. Although the data in CulturaX has already gone through an extensive cleaning pipeline, we expand it by only allowing Estonian data that comes from websites ending with either .ee, .org, or .net. The pretraining is done with up to 5B tokens, 75% of which are Estonian and the rest are English, to prevent English knowledge forgetting.

4 Experimental Setup

4.1 Base Model

To obtain the base model, we continue pretraining Llama-2-7B (Touvron et al., 2023b) with the additional 5B tokens of pretraining data described in Section 3.3. We call the base model Llammas-base. We use packing for pretraining which means that the training examples are concatenated to fill the model context. The training setup and parameters are outlined in Appendix A. We publish our training code[4].

4.2 Instruction-tuned Models

Models instruction-tuned only with Alpacas or translation task instructions use the Alpaca prompting format (Taori et al., 2023). The models relying on high-quality instructions (**HQI** or **HQTrTask**) are trained as conversational models with conversation format following Wang et al. (2023a, see Table 12).

During the training, we mask the user prompts, including the whole user input in the conversational format (including multi-turn). The models are trained for 3 epochs. We picked the best epoch according to the validation loss, which was always the first epoch in our experiments. See Appendix A for other training details.

4.3 Evaluation Datasets

Following Ranaldi et al. (2023); Zhu et al. (2023), we use EstQA (Käver, 2021), an Estonian version of SQUAD (Rajpurkar et al., 2016) as one of the evaluation datasets. Since the original EstQA does not include a validation split, we create one ourselves by separating a small subset of training data for that purpose.

We also evaluate our models on Estonian commonsense reasoning (CSR) and grammatical error correction (GEC) tasks. For commonsense reasoning, we use EstCOPA (Kuulmets et al., 2022),

[3] https://anonymous.4open.science/r/alpaca-est

[4] https://anonymous.4open.science/r/llammas

which is an Estonian version of the COPA task (Roemmele et al., 2011). EstCOPA includes both machine-translated and manually post-edited versions of COPA. We use the latter for our evaluations. Grammatical error correction is evaluated with EstGEC-L2 dataset[5].

Finally, results for English-Estonian and Estonian-English translation (MT) tasks are reported using FLORES-200 devtest (NLLB Team, 2022). It is important to note that, depending on the model, the translation task may be included into the training process, while the models are never exposed to any other evaluation tasks.

4.4 Perfomance on English

Ideally, our model should also perform reasonably well in English. If that was not the case it would mean that we might have washed out the pre-existing knowledge from the models. That could happen, for example, with overly extensive training on task-specific datasets. Naturally, it would be an indication that the model is not using its knowledge in English to generate answers in Estonian. To verify that our models can still understand English, we evaluate our best models on COPA, on an English subset of XQuAD (Artetxe et al., 2020), and an English grammatical error correction task using the W&I+LOCNESS test set (Bryant et al., 2019).

4.5 Evaluation Metrics

To evaluate commonsense reasoning and question-answering we use the assessments of GPT-4 Turbo[2]. More precisely, we employ LLM-as-a-Judge (Zheng et al., 2023) with reference-guided grading where the model is asked to assess the correctness of the predicted answer given the reference answer and the task itself. We modified the evaluation prompt from Zheng et al. (2023) to align with our tasks. We chose GPT-4 Turbo as the evaluator over ChatGPT[2] to ensure the reliability of the results, as it demonstrated a significant improvement in assessment quality (specifically, a reduction in false positives) in our preliminary experiments. To reduce API usage costs, we base our QA accuracy report on 100 randomly chosen samples from the corresponding datasets and splits. When evaluating the commonsense reasoning task, we feed to GPT-4 Turbo only answers that we were not able to classify with a simple string comparison.

We also report standard metrics for most of the tasks. For question answering and grammatical error correction we report F1 and M2 scorer[6] (Dahlmeier and Ng, 2012) or ERRANT (Bryant et al., 2017) F0.5, respectively. For translation tasks we calculate BLEU[7] (Papineni et al., 2002) and chrF++[8] (Popović, 2017) using sacreBLEU (Post, 2018), and COMET (Rei et al., 2020) scores using the `unbabel-wmt22-comet-da` model (Rei et al., 2022).

4.6 Evaluation Prompts

During the development phase, the performance on EstCOPA, EstQA, and their English equivalents is measured with 8 different prompts. The English prompts are from Wei et al. (2021), while prompts for Estonian tasks are written by the authors. On development datasets, we report the best score across the 8 prompts, while on test datasets, we only report the scores obtained with the best prompt according to the development datasets. For machine translation and grammatical error correction tasks, we use the same single prompt during the development and test phases (see Table 10).

5 Experiments and Results

Our experiments are divided into two main sections. In the first section, we pretrain Llama-2-7B on different amounts of pretraining data and investigate the effect of it on cross-lingual instruction-tuning that is done with translation task and general task instructions (Alpacas).

In the second section, we study the influence of supplementing Alpacas with high-quality English instructions, translations, and conversations to the results on Estonian.

5.1 Continued Pretraining of Llama 2

We compare three base models. First, Llama-2-7B without any additional pretraining. Second, the checkpoint of LLAMMAS-BASE that has seen 1B tokens of pretraining data. Third, LLAMMAS-BASE trained on the entire pretraining dataset of 5B tokens. We instruction-tune all three models on Alpacas that consisting of Estonian and English general task instructions. The results of the three models are compared in Figure 1. We observe

[5] https://github.com/tlu-dt-nlp/EstGEC-L2-Corpus

[6] https://github.com/TartuNLP/estgec/tree/main/M2_scorer_est

[7] sacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`

[8] sacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1`

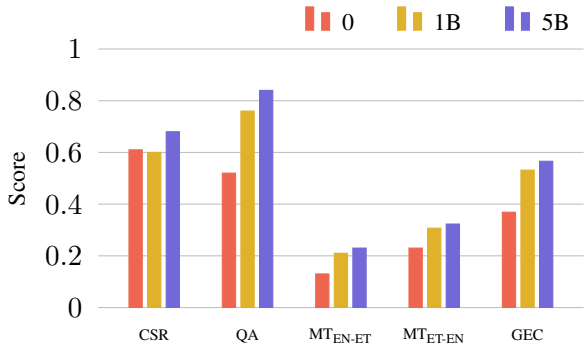

Figure 1: Results on Estonian tasks after fine-tuning Llama-2-7B with cross-lingual instruction-tuning dataset Alpacas. The colors of the bars indicate the size of the pretraining dataset.

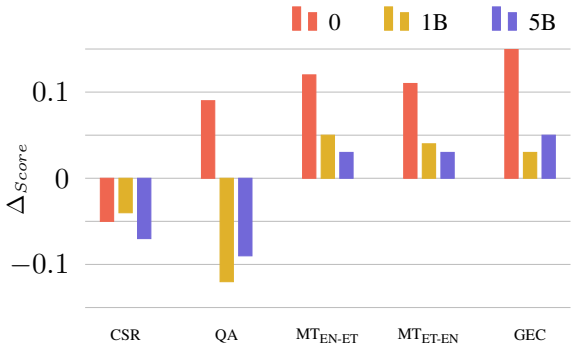

Figure 2: Performance gained or lost on Estonian tasks after fine-tuning Llama-2-7B first on translation task and then on Alpacas compared to when translation task is omitted (Figure 1). The colors of the bars indicate the size of the pretraining dataset.

performance gains on all Estonian tasks as the size of the pretraining dataset increases.

In our preliminary experiment (included into the ablation study, Section 6.1) we observed that after additional pretraining of Llama-2-7B with 1B tokens the benefits of using translation task during fine-tuning diminished. To assess whether this trend persists with even larger pretraining, we instruction-tune the base models with a dataset that consists of both translation and general task instructions, i.e., **TRTASK** and Alpacas. We adopt sequential training based on our preliminary experiment (Section 6.1), which indicated that this setup has a milder negative impact on performance in zero-shot tasks.

Figure 2 shows the performance gained or lost for each task and base model with the translation task used as the first step during instruction-tuning. We can see that without additional pretraining, the translation task significantly improves the results for QA, machine translation, and GEC. However, the benefit diminishes greatly when the pretraining step is introduced. For QA and commonsense reasoning, omitting the translation task after pretraining tends to produce stronger results compared to models where pretraining is followed by the translation task.

5.2 Beyond Alpacas: Knowledge Transfer via High-Quality English Instructions

Instruction-tuning datasets generated with Self-Instruct (Wang et al., 2023b) might suffer from various issues that lower the overall quality of the dataset[1]. Meanwhile, it has been shown that it is possible to achieve remarkably strong performance with just 1,000 high-quality training examples (Zhou et al., 2023). In light of this, we hypothesize that supplementing the Alpacas dataset with a set of high-quality instructions could improve the models. However, as there are no high-quality instruction datasets available for Estonian, we use only high-quality English instructions (**HQI**). For comparison, we train a model where high-quality English instructions are supplemented with high-quality translation task instructions (**HQTRTASK**).

The results are shown in Table 1. Compared to the baseline model (1) that is trained on just Alpacas, we observe a somewhat surprising increase in all scores when Alpacas is supplemented with high-quality English instructions (model (3)). This suggests that there is a positive cross-lingual knowledge transfer from the added high-quality English instructions into Estonian. Moreover, combining high-quality English instructions with high-quality translation tasks further enhances the knowledge transfer (model (4)). We call this model LLAMMAS. However, we observe that the best results for EN→ET, ET→EN, and GEC are obtained with a model that is trained sequentially, with HQTRTASK as the first step of fine-tuning (model (5)). We call this model LLAMMAS-TRANSLATE.

Models (3) – (5) are trained with the data in chat format (see Table 12), since HQI contains English conversational data from Open Assistant 1. Through manual evaluation with 5 conversations (up to 6 turns), we determine that model (4) (LLAMMAS) can adequately engage in multi-turn conversations. It can recall content from previous turns and respond to user requests fairly well. However, we also see that the model sometimes makes grammatical mistakes and uses words or

Model		CSR	QA		MT_{ET-EN}	MT_{EN-ET}	GEC
		acc.	F1	acc.	BLEU	BLEU	F0.5
LLAMMAS-BASE fine-tuned:							
(1)	Alpacas	63.6	46.53	81	22.5	32.3	56.6
(2)	1) TRTASK 2) Alpacas	59.2	46.15	73	25.0	34.5	59.4
(3)	Alpacas + HQI	**66.4**	52.86	82	23.1	32.4	59.4
(4)	Alpacas + HQI + HQTRTASK	**66.4**	**54.75**	**84**	22.6	34.6	60.3
(5)	1) TRTASK 2) **(4)**	62.2	43.46	76	**26.9**	**36.9**	**61.2**
Commercial baselines:							
	GPT3.5	86.0	34.17	93	37.5	26.0	63.4
	GPT4	98.4	35.14	97	37.7	28.5	67.4

Table 1: Fine-tuning LLAMMAS-BASE on different cross-lingual instruction datasets. We call **(4)** LLAMMAS and **(5)** LLAMMAS-TRANSLATE.

phrases that a native Estonian speaker would not use. Many of these phrases sound like translations from English. An example conversation can be seen in Table 13. The model's conversational ability suggests that the model has learned to hold a multi-turn conversation in Estonian through cross-lingual transfer, however, more experiments would be needed to confirm that.

5.3 Results on Translation Task

Conventional neural machine translation (NMT) models leverage tens of millions of parallel sentences along with the use of monolingual corpora. In contrast, LLAMMAS-TRANSLATE uses a modest 1 million sentence pairs from relatively low-quality parallel data sources and a small number of sentences from high-quality sources. In combination with general task instructions, this results in a competitive translation model, as presented in Table 2. We can see that LLAMMAS-TRANSLATE outperforms LLAMMAS although, in terms of COMET, which is more highly correlated with human judgments (Freitag et al., 2022), LLAMMAS still seems competitive.

When comparing LLAMAS-TRANSLATE to the open-source encoder-decoder models MTee and NLLB-MoE, LLAMAS-TRANSLATE achieves better scores on COMET and similar scores on BLEU and chrF++. On ET→EN LLAMMAS-TRANSLATE is outperformed by NLLB-MoE, however, it outperforms MTee on COMET and achieves a similar score in chrF++. We can also see that LLAMMAS-TRANSLATE is competitive with GPT-3.5-turbo, however it is outperformed by GPT-4-turbo[9].

[9]Prompt used for OpenAI models: `"Translate the following {src_lang} text into {tgt_lang}:\n{sentence}"`

5.4 Results on Grammatical Error Correction

LLMs are good at text correction, yet they frequently make extensive edits that diverge from traditional GEC metrics, known for preferring minimal modifications (Coyne et al., 2023). This tendency is apparent in English, where the models exhibit higher recall than precision (see Table 3). For Estonian, in contrast, the models show higher precision but reduced recall, indicating a different correction pattern from Estonian. We leave further exploration of that phenomenon for future work. Finally, we can see that translation task instructions (TRTASK, used for training LLAMMAS-TRANSLATE) enhance performance in Estonian which is in accordance with our earlier experiments.

5.5 Results on XQUAD and COPA

The results on English QA and commonsense reasoning tasks are shown in Table 4. On the QA task, LLAMMAS achieves similar accuracy in English and Estonian (83% vs 84%). However, we observed that LLAMMAS is more chatty in English, resulting in longer answers and therefore lower F1 score when compared to Estonian. Finally, we observe that LLAMMAS solves commonsense reasoning problems significantly better in English than in Estonian (80.6% vs 66.4%) This indicates that LLAMMAS is still not able to utilize all the reasoning capabilities it has in English when the input is given in Estonian.[10]

[10]Hence the name LLAMMAS, as in Estonian, the word *lammas* means *sheep*.

Model	Param.	ET→EN			EN→ET		
		BLEU	chrfF++	COMET	BLEU	chrfF++	COMET
MTee (Tättar et al., 2022)	227M	36.7	61.3	88.48	27.6	56.9	89.18
NLLB-MoE (NLLB Team, 2022)	54.5B	38.8	62.6	89.25	27.1	56.1	91.44
GPT-3.5-turbo	-	37.5	63.0	89.52	26.0	56.3	91.67
GPT-4-turbo	-	**37.7**	**63.8**	**89.74**	**28.5**	**58.4**	**92.55**
LLAMMAS (ours)	7B	34.6	59.2	89.00	22.6	51.8	91.00
LLAMMAS-TRANSLATE (ours)	7B	36.9	61.2	89.09	26.9	56.4	91.92

Table 2: Translation metric scores on FLORES-200 devtest (NLLB Team, 2022).

Model	ET			EN		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
GPT-3.5-turbo	69.58	46.66	63.36	53.63	70.13	56.28
GPT-4	**74.31**	**49.21**	**67.43**	**56.68**	**71.57**	**59.14**
LLAMMAS (ours)	67.6	42.2	60.3	58.01	59.45	58.29
LLAMMAS-TRANSLATE (ours)	68.0	43.6	61.2	55.94	59.34	56.59

Table 3: GEC scores on EstGEC-L2 and W&I+LOCNESS test sets.

5.6 Robustness on Diverse Prompts

We look into the distribution of metric scores on 8 development prompts (Table 1) to assess the robustness of our models when encountering various input prompts.

EstCOPA shows an increase in robustness and average scores with various prompts when high-quality English instructions are used (see Figure 3). This is even further increased by the addition of high-quality translation instructions. While having lower scores than the models without a translation step, Llammas-translate still displays good robustness. On EstQA, however, we don't see the same trend. There is an increase in the median of the metric score, yet the robustness does not increase. For models involving the use of high-quality data, the lowest-scoring prompts still achieve higher F1 scores than the median of the model fine-tuned on Alpacas.

6 Ablation study

6.1 Instruction-Tuning: Sequentially or with a Combined Dataset?

Previous research has explored approaches that combine translation and general task instructions for cross-lingual instruction-tuning (Ranaldi and Pucci, 2023; Ranaldi et al., 2023; Zhu et al., 2023). However, these approaches combine both types of instructions into a single dataset for model fine-tuning. We hypothesize that such setup, especially when a significantly larger translation task dataset is used (e.g. by Zhu et al., 2023), may diminish the contribution of general task instructions during the training, adversely impacting the model's ability to generalize to new tasks.

To test the hypothesis we compare fine-tuning Llama-2-7B on a combined dataset to fine-tuning it with sequential training. The latter involves first training the model on the translation task and then on general task instructions. We replicate the experiment with Llama-2-7B further pretrained on 1B tokens, to validate the consistency of results when the pretraining step is included. We use context size of 224 and, following Zhu et al. (2023), only English to target language translations (TRTASK$_{\text{EN}\rightarrow\text{ET}}$). We compare the results with baselines where translation task data is entirely omitted.

The results in Table 9 show that fine-tuning Llama-2-7B on translation task improves most results (except commonsense reasoning). Combined training is particularly beneficial for EN→ET and grammatical error correction. The latter aligns with the improvement in EN→ET as MT and GEC are similar tasks and often approached in a similar way (Junczys-Dowmunt et al., 2018). However, QA and ET→EN gain more from sequential training. It is especially notable for ET→EN where general task instructions recover the performance after the initial degradation.

However, we observe that when pretraining Llama-2-7B on 1B tokens is included, the performance generally suffers when translation task instructions are used. Exceptions are English-

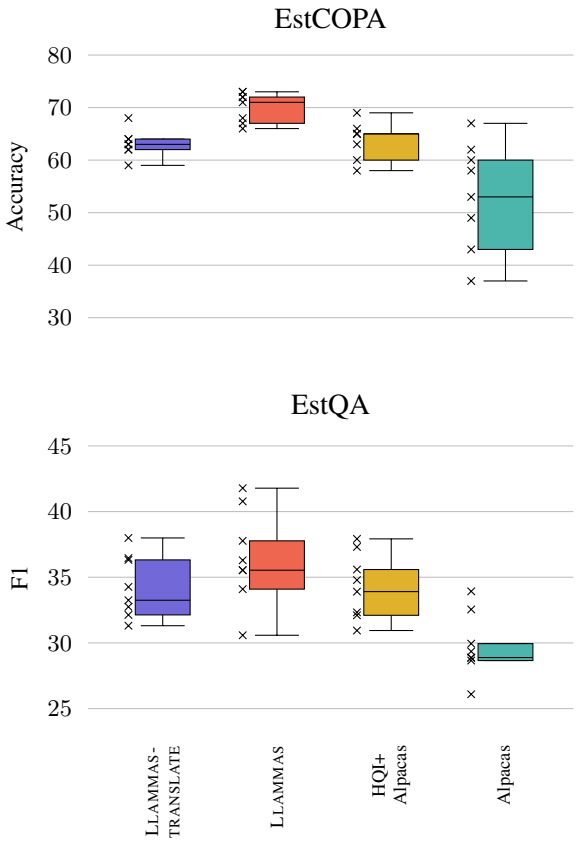

Figure 3: EstCOPA development set accuracy and EstQA development set F1-score of 8 prompts on models fine-tuned from LLAMMAS-BASE (see Table 1).

Estonian and grammatical error correction that naturally benefit from the translation task.

Finally, we can see that EN→ET is rather weak on pretrained Llama-2-7B after fine-tuning on just Alpacas. However, including the task drastically hurts the performance of ET→EN translation task.

6.2 Translation Data: The Impact of Quality and Quantity

In Section 6.1 we found that language-specific pretraining of Llama-2-7B followed by fine-tuning on just Alpacas outperforms the same base model fine-tuned on both translation and general task instructions. Combining the datasets (**TRTASK$_{\text{EN}\rightarrow\text{ET}}$ + Alpacas**) yielded weaker scores than sequential training (**1) TRTASK$_{\text{EN}\rightarrow\text{ET}}$ 2) Alpacas**). To address the potential negative influence from the imbalanced dataset, where translation instructions outnumber general task instructions by about 10 times, we conduct an experiment with a balanced dataset. We fine-tune the base model with a dataset combining general task instructions with 100K translation task instructions (similar in size to Alpacas) from the data mix described in Section 3.2. Table 8 shows that the model does not outperform the Alpacas baseline.

Model	CSR acc.	QA F1	QA acc.
Alpacas	63.4	30.43	85
1) TrTask 2) Alpacas	70.2	29.48	81
Alpacas + HQI	78.6	33.30	**87**
LLAMMAS	**80.6**	**40.96**	83
LLAMMAS-TRANSLATE	73.6	31.35	82
GPT3.5	95.2	30.67	95
GPT4	99.8	33.16	96

Table 4: Results on English commonsense reasoning and question answering.

Additionally, we train the base model with a dataset combining general task instructions with a small set of high-quality translation task instructions from MTee held-out validation sets (Tättar et al., 2022) and WMT18 development set (Bojar et al., 2018). This model also does not outperform the baseline model, except in GEC which seems to benefit from high-quality translation task.

6.3 Translation Data: Single Translation Direction or Both?

We investigate the effect of EN→ET : ET→EN translation direction proportion in our data. From Table 7, we can see that for all tasks, having only EN→ET direction is not optimal when translation data is used. For MT$_{\text{ET}\rightarrow\text{EN}}$ and GEC 25% ET→EN seems to offer the best scores, while for other tasks 50% offers the highest scores. For CSR, having no translation data at all offers the highest accuracy.

7 Conclusion

We successfully adapt Llama 2 to Estonian by creating LLAMMAS - an instruction-following model for Estonian. Additionally, we release Alpaca-est, an Alpaca-style general task instruction dataset for Estonian. Our work has shown competitive results for tasks such as question-answering, machine translation, and grammatical error correction in Estonian while keeping solid results for English. We have also identified signs of cross-lingual transfer from English to Estonian and investigated the effects of translation bitexts in the fine-tuning process. This work marks the first step towards open-source LLMs for Estonian.

Limitations

The key limitation of this work is the dependence on data generated with OpenAI's proprietary LLMs. As Gudibande et al. (2023) have found, these generated datasets result in the imitation of the proprietary LLM's style but not necessarily factuality. Secondly, due to the limited number of benchmarks for Estonian, our evaluation is limited to a rather small number of NLP tasks. Because of the early stages of the research on capabilities and harmlessness, the model will be limited to research purposes.

Ethics

We believe that extending open-source large language models to previously uncovered languages poses a net positive impact as it allows more people access to them. However, the currently released model lacks safety evaluation, meaning that it should be used only for research purposes. Furthermore, the self-instruct style generated instructions have not been manually checked, increasing the risks (for example bias) even more. Further research into evaluating the harmlessness and helpfulness of LLMs for Estonian is needed, as this has not been done for proprietary LLMs that support Estonian either.

Acknowledgements

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction.

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

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

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

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, New Orleans, Louisiana. Association for Computational Linguistics.

Hele-Andra Kuulmets, Andre Tättar, and Mark Fishel. 2022. Estonian language understanding: a case study on the copa task. In *Proceedings of Baltic HLT 2022*, volume 10, page 470–480, Riga, Latvia. Baltic Journal of Modern Computing.

Anu Käver. 2021. Extractive question answering for estonian language. Master's thesis, Tallinn University of Technology.

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

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

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. 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.

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 Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

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.

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*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

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.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. *arXiv preprint arXiv:2308.14186*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Andre Tättar, Taido Purason, Hele-Andra Kuulmets, Agnes Luhtaru, Liisa Rätsep, Maali Tars, Mārcis Pinnis, Toms Bergmanis, and Mark Fishel. 2022. Open and competitive multilingual neural machine translation in production. *Baltic Journal of Modern Computing*, 10(3):422–434.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. 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 Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How far can camels go? exploring the state of instruction tuning on open resources.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv 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*.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, et al. 2023a. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023b. Plug: Leveraging pivot language in cross-lingual instruction tuning. *arXiv preprint arXiv:2311.08711*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

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. *arXiv preprint arXiv:2308.04948*.

A Training Parameters

The context length in our training experiments is 1024 tokens with the overlapping examples truncated. The models are trained with `bf16` precision using DeepSpeed. A learning rate of 2e-5 is used and is linearly decayed to 2e-6. During pretraining a batch size of 256 is used and during instruction-tuning the batch size is 128. We train our models on 4 AMD MI250x GPUs (acting as 8 GPUs) on the LUMI supercomputer.

The pretraining on 5B tokens took 1184 GPU-hours (LLAMMAS-BASE). Instruction-tuning of LLAMMAS took 80 GPU-hours (3 epochs). Instruction-tuning on translation data (TRTASK) for LLAMMAS-TRANSLATE took 190 GPU-hours (3 epochs), in addition to the instruction-tuning on the general instructions (i.e, fine-tuning LLAMMAS).

B Sizes of Datasets

	Validation	Test
Question Answering		
EstQA (Käver, 2021)	85	603
XQuAD (Artetxe et al., 2020)	1190	-
Commonsense Reasoning		
EstCOPA (Kuulmets et al., 2022)	100	500
COPA (Roemmele et al., 2011)	100	500
Grammatical Error Correction		
EstGEC-L2[5]	879	2029
W&I+LOCNESS (Bryant et al., 2019)	4385	4477
Machine Translation		
FLORES-200 (NLLB Team, 2022)	997	1012

Table 5: Sizes of evaluation and test datasets (number of examples). The entire XQUaD was used for both validation and testing.

General instructions	
Alpaca-cleaned (Taori et al., 2023)	52,000
AlpacEst (ours)	52,006
HQI	
CoT (Chung et al., 2022; Ivison et al., 2023)	10,000
FlanV2 (Chung et al., 2022; Ivison et al., 2023)	10,000
Open Assistant 1 (Köpf et al., 2023)	2,363
Translation instructions	
TRTASK	
CCMatrix (Schwenk et al., 2021b)	500,000
WikiMatrix (Schwenk et al., 2021a)	400,000
Europarl (Tiedemann, 2012)	50,000
OpenSubtitles (Lison and Tiedemann, 2016)	50,000
HQTRTASK	
WMT18 dev (doc. level) (Bojar et al., 2018)	245
MTee valid held-out (general) (Tättar et al., 2022)	1528
Additional HQ translation data	
MTee valid held-out (all) (Tättar et al., 2022)	4353
WMT18 dev (sent. level) (Bojar et al., 2018)	2000

Table 6: Sizes of instruction datasets (number of examples).

C Ablation Study Tables

D Evaluation prompts

Model	TrTask ET→EN	CSR acc.	QA acc.	$MT_{EN \to ET}$ BLEU	$MT_{ET \to EN}$ BLEU	GEC F0.5
$TrTask_{100k}$ + Alpacas	50%	59	**76.47**	20.4	**32.7**	56.2
$TrTask_{100k}$ + Alpacas	25%	55	74.12	**21.2**	32.6	**58.1**
$TrTask_{100k}$ + Alpacas	0%	56	71.76	21.1	1.6	56.2
Alpacas	**-**	**66**	74.12	20.8	32.4	50.0

Table 7: Fine-tuning Llama-2-7B further pretrained on 1B token. Translation task ET→EN direction proportion is modified. 0% means that all of TrTask data is in EN→ET direction. The amount of translation task data is fixed at 100k sentence-pairs. Results are reported on development datasets.

Model	TrTask size	CSR acc.	QA acc.	$MT_{EN \to ET}$ BLEU	$MT_{ET \to EN}$ BLEU	GEC F0.5
$TrTask_{EN \to ET}$ + Alpacas	1M	53	63.53	**24.4**	1.50	**57.5**
$TrTask_{EN \to ET}$ + Alpacas	100K	56	71.76	21.1	1.60	56.2
$TrTask_{\text{High quality EN} \to ET}$ + Alpacas	6K	57	69.41	22.2	3.60	**57.5**
Alpacas	-	**66**	**74.12**	20.8	**32.40**	50.5

Table 8: Quantity vs quality: examining the impact of translation task dataset composition. Results are reported on development datasets.

	CSR acc.	QA acc.	GEC F0.5	$MT_{EN \to ET}$ BLEU	$MT_{ET \to EN}$ BLEU
Llama-2-7B					
$TrTask_{EN \to ET}$ + Alpacas	**0.58**	0.61	**0.55**	**24.6**	1.50
1) $TrTask_{EN \to ET}$ 2) Alpacas	**0.58**	**0.65**	0.51	24.5	**27.40**
Alpacas	**0.61**	0.52	0.34	13.9	24.80
Llama-2-7B pretrained on 1B tokens of Estonian-centric data					
$TrTask_{EN \to ET}$ + Alpacas	0.53	0.64	**0.57**	24.4	1.50
1) $TrTask_{EN \to ET}$ 2) Alpacas	**0.55**	**0.71**	0.56	**25.7**	**23.00**
Alpacas	**0.66**	**0.74**	0.50	20.8	**32.40**

Table 9: Comparison of cross-lingual training strategies across two different base models. Results are reported on development datasets.

MT EVALUATION PROMPT:
Translate the following `{source language}` text into `{target language}`.
GEC EVALUATION PROMPT:
Reply with a corrected version of the input sentence in `{language}` with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence.
ESTCOPA EVALUATION PROMPT (the best on development set):
Olgu eelduseks lause "{premise}" Mis on selles lauses kirjeldatud olukorra {question}? VALIKUD: - {choice1} - {choice2}
COPA EVALUATION PROMPT (the best on development set):
{premise} What is the {question} of the preceding sentence? OPTIONS: - {choice1} - {choice2}
ESTQA EVALUATION PROMPT (the best on development set):
Tekst: {paragraph} Küsimus: {question}
XQUAD EVALUATION PROMPT (the best on development set):
Here is a question about this article: {paragraph} What is the answer to this question: {question}

Table 10: Instruction prompts for evaluating LLAMMAS. The prompts will be integrated into Alpaca-style formatting (see Table 11). In the case of conversational model, the Alpaca-formated prompts will be inserted into the chat format (see Table 12).

Below is an instruction that describes
a task, paired with an input that provides
further context. Write a response that
appropriately completes the request.
 	
Instruction:
`{instruction}`
 	
Input:
`{input}`
 	
Response:

Table 11: Alpaca instruction format (Taori et al., 2023)

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

Table 12: Chat format following (Wang et al., 2023a). After **<|assistant|>** the model generates the response. </s> is the end-of-string token.

User:
Tere! Kuidas läheb?

Assistant:
Tänan küsimast, täitsa hästi. Kuidas sinul läheb?

User:
Hästi, kuigi viimased päevad on olnud väsitavad.

Assistant:
Mis teid väsinuna tunneb? Kas teil on midagi konkreetset, mis seda põhjustab? Võite jagada oma muret ja ma aitan leida lahendusi.

User:
Keeruline öelda. Võimalik, et asi on selles, et olen liiga vähe puhanud.

Assistant:
Võib-olla peaksime proovima erinevaid stressileevendamise meetodeid nagu meditatsioon või lõdvestusharjutused. Samuti võiksime planeerida rohkem aega endale ning luua struktuurse rutiini, et tagada piisav uni ja puhkus. Räägime sellest veelgi edasi, kui see on vajalik.

User:
Paku mulle üks stressileevendusmeetod

Assistant:
Proovi regulaarselt praktiseerida joogat, mis aitab sul rahus ja harmoonias olla. Sellel on mitmeid tervistavaid efekte, sealhulgas parem kehaline painduvus, suurenenud energia ja vaimne selgus. Jooga sobib ideaalselt igapäevaselt harjumuseks muutmisena, aidates sul stressirohke elustiiliga toime tulla. Alusta tasakaaluharjutuste, venituste ja hingamisharjutustega ning avasta joogamaailmas uued võimalused endaga paremini kontaktis olemiseks. Leia sobiv stiil ja õpetaja ning alusta oma joogateekonda juba täna.

User:
Tahaks midagi lihtsamat.

Assistant:
Kirjuta endale märkmikusse positiivsed eneseütlemised, näiteks "olen tugev" või "olen nutikas". Lugege neid regulaarselt ja visualiseerige end heade omadustega inimesena. See aitab teil end motiveerida ja uskuda iseendasse. Lisage ka motivatsiooni tsitaate, nagu "Elu pole mitte ainult selle kohta, millistes asjaoludes meie juhtumeid saame, vaid ka selles, kuidas me reageerime." - Walt Disney. Tegele ka füüsiliselt aktiivsema eluga, näiteks treeningu või jalutuskäikudega värskes õhus. Need meetmed aitavad teil stressi maandada ja enda eest hoolitseda. Ärge unustage naeratada ja leida huumorit igapäevaelus. Naeratused levitavad head energiat ja loovad positiivse atmosfääri.

User:
Tänan!

Assistant:
Olete teretulnud! Ma olen siin, et sind toetada ja aidata. Palun anna mulle teada, kui sul on mingeid muid küsimusi või abisoove. Soovin sulle imelisi hetki ja palju edu!

Table 13: An Example conversation with the LLAMMAS model. Generated using sampling with top-p of 0.9, temperature of 0.6, repetition penalty of 1.2, and beam size of 1 (no beam search)