How Well do LLMs know Finno-Ugric Languages? A Systematic Assessment

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

Abstract

We present a systematic evaluation of multilingual capabilities of open large language models (LLMs), specifically focusing on five Finno-Ugric (FiU) languages. Our investigation covers multiple prompting strategies across several benchmarks and reveals that Llama 2 7B and Llama 2 13B perform weakly on most FiU languages. In contrast, Llama 3.1 models show impressive improvements, even for extremely low-resource languages such as Võro and Komi, indicating successful cross-lingual knowledge transfer inside the models. Finally, we show that stronger base models outperform weaker, language-adapted models, thus emphasizing the importance of the choice of the base model for successful language adaptation.

1 Introduction

Large language models (LLMs) have recently made significant advances in multilingual settings. For instance, GPT-4 achieves 80.9% accuracy for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark (OpenAI et al., 2024). For some time, strong multilingual capabilities were mainly limited to proprietary models, such as ChatGPT[1] and Claude[2], whose weights, training details, and inference processes are kept private. These models outperformed open LLMs[3] like Llama 2 models (Touvron et al., 2023), on non-English tasks. However, open-weight LLMs have recently begun to close this gap (Dubey et al., 2024; Jiang et al., 2024), even though the officially supported languages of these models remain limited and the primary focus is on those with significantly more data available than for Finno-Ugric (FiU) languages.

On the other hand, it has been observed that even models optimized solely for English, such as the Llama 2 family models (Touvron et al., 2023), demonstrate some understanding of a wide range of languages beyond their intended use (Holtermann et al., 2024). In experiments conducted by Holtermann et al. (2024), the Llama 2 7B chat model correctly answered 14% and 40% of basic open-ended questions in Estonian and Finnish, respectively, even though only 0.03% of the Llama 2 training data was in Finnish and less than 0.005% in Estonian (Touvron et al., 2023).

This work evaluates the multilingual capabilities of open LLMs on five FiU languages: Finnish, Estonian, Livonian, Võro, and Komi. Among these, Finnish and Estonian are the most well-resourced, making it easier to adapt existing LLMs for these languages through continued pretraining (Kuulmets et al., 2024; Luukkonen et al., 2023). In contrast, Võro, Livonian, and Komi are extremely low-resource languages, making language-specific adaptation considerably more challenging.

The aim of this work is to clarify the capabilities of open LLMs in understanding FiU languages. While it is evident that open LLMs can understand these languages to some degree (Holtermann et al., 2024), their proficiency and comparative performance across models remain largely unexplored. We focus on Llama models, which have demonstrated state-of-the-art performance and competitiveness with proprietary models (Dubey et al., 2024; Touvron et al., 2023) and have been widely used in non-English adaption (Kuulmets et al., 2024; Etxaniz et al., 2024; Lin et al., 2024; Fujii et al., 2024; Dima et al., 2024; Basile et al., 2023). Another reason for focusing on Llama

[1] https://openai.com/index/chatgpt/
[2] https://www.anthropic.com/claude
[3] Models that have publicly accessible weights available for use, modification, and research.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 340–353
March 3-4, 2025

models is that the newer Llama 3.1 models are natively multilingual, potentially improving performance on unsupported languages as well. For further insights, we compare Llama models with Mistral NeMo (Jiang et al., 2024), another natively multilingual open model shown to be competitive with Llama 3.1 model of the same size.

We evaluate only base models rather than chat-optimized models, as most knowledge is acquired during pretraining (Zhou et al., 2023; Lin et al., 2023). In other words, a stronger base model offers greater potential for developing a strong chat model. Consequently, the performance of base models on different FiU languages can serve as a relative estimate of the chat model's quality.

The evaluation is conducted using several existing benchmarks that include one or more Finno-Ugric languages. We examine both the zero-shot and few-shot capabilities of these models. Additionally, we explore whether chain-of-thought prompting, which involves first translating the input to English, could improve results on Finno-Ugric languages. In summary, we seek to answer the following research questions:

1. How well can open LLMs solve tasks in Finno-Ugric languages?
2. What is the expected improvement from few-shot prompting over zero-shot prompting in solving tasks in Finno-Ugric languages?
3. Can chain-of-thought prompting, where the model first translates the input into English, improve the performance of open LLMs on Finno-Ugric languages?

2 Related Work

2.1 Multilingual LLMs

While state-of-the-art LLMs are typically trained on English-centric data, they exhibit some multilingual capabilities (Brown et al., 2020; Holtermann et al., 2024), even for languages with minimal representation in the training data (Holtermann et al., 2024; Touvron et al., 2023). This suggests that knowledge transfer from high-resource languages to low-resource languages must occur at least to some extent within the model. These multilingual capabilities can be further enhanced through continued pretraining in the target languages, even with just a few billion tokens of data (Pires et al., 2023; Cui et al., 2024; Kuulmets et al., 2024; Etxaniz et al., 2024).

Recent open LLMs such as Llama 3.1 (Dubey et al., 2024), Mistral NeMo (Jiang et al., 2024), and Tower (Alves et al., 2024) are specifically optimized for multilingual performance. For example, Llama 3.1 models officially support seven non-English languages (Dubey et al., 2024), Mistral NeMo is particularly strong in ten languages other than English (Jiang et al., 2024), and Tower is trained on a multilingual dataset consisting of ten languages, including English. According to Dubey et al. (2024), the strong performance in non-English languages is achieved by increasing the proportion of multilingual data in the pretraining dataset and incorporating high-quality target language instructions into the instruction-tuning data.

However, neither Mistral NeMo nor Llama 3.1 models officially support Finno-Ugric languages. The amount of Finno-Ugric data in their pretraining corpora is unknown but is likely very limited. For example, Purason et al. (2024) presented experiments on adapting LLMs to FiU languages, but gathered only 2.6 million characters of pretraining data for Livonian, 14 million for Võro, and 579 million for Komi.

2.2 In-context Learning

In-context learning (ICL) (Brown et al., 2020) is a method where a pretrained language model *learns* to generate the desired output for a given task from the context of the prompt, without any gradient updates. One of the most common applications of ICL is few-shot prompting, where a few example question-answer pairs are provided in the prompt to guide the model in solving the task.

2.2.1 Chain-of-thought Prompting

Chain-of-thought (CoT) prompting (Wei et al., 2023) is a prompting technique that improves upon few-shot prompting. With CoT, the example demonstrations provided in the prompt include a series of intermediate reasoning steps that conclude with an answer as opposed to being just question-and-answer pairs. While initially proposed to improve English reasoning in LLMs, Shi et al. (2022) showed that CoT prompting turns English-centric PaLM and GPT-3 into multilingual reasoners, achieving strong results even in languages whose proportion in the training data is as small as 0.01%. Notably, they achieve an accuracy of 91% on the Estonian subset of the multilingual commonsense reasoning benchmark XCOPA

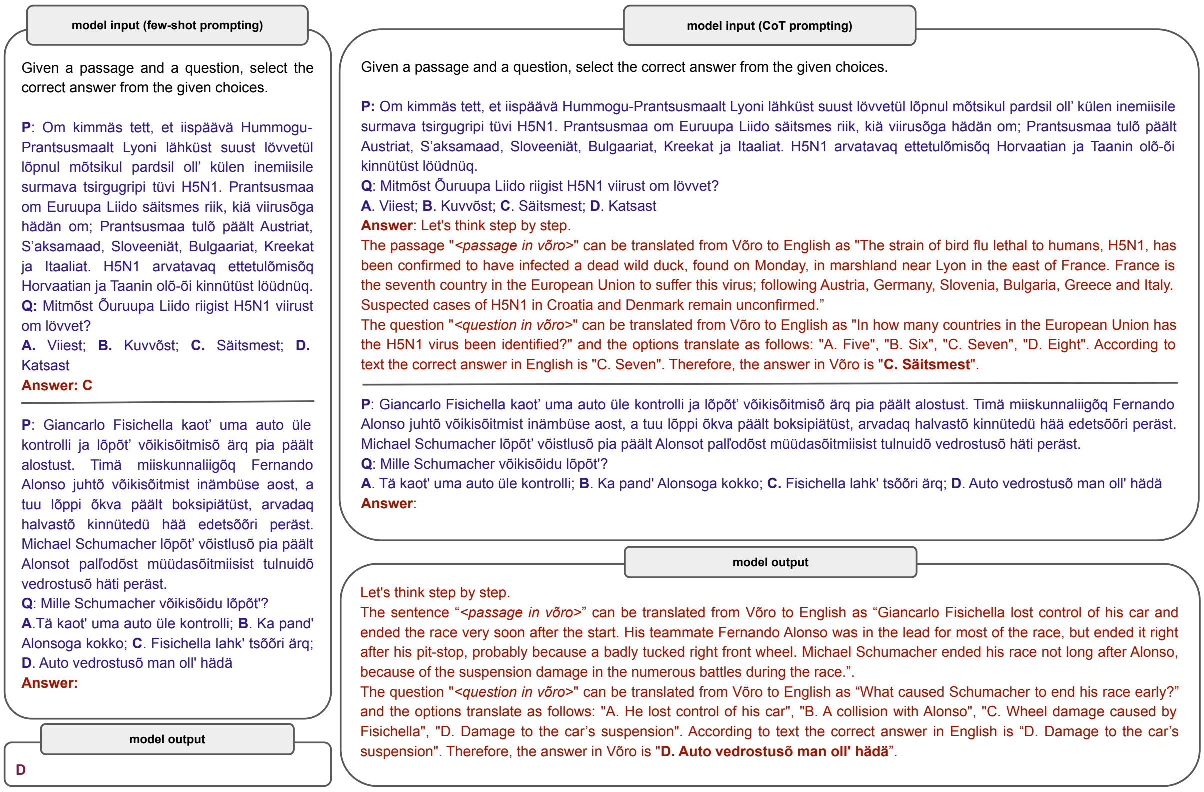

Figure 1: Model input and expected output for few-shot prompting (left) and for CoT prompting where the intermediate step involves translating the input from the source language (Võro) to English. The example is taken from the Belebele benchmark.

(Ponti et al., 2020) (average accuracy 89.9%) with PaLM. Their observation that there is no strong correlation between performance and language frequency in the training corpora leads them to suggest that, to some extent, language models can transfer knowledge from high-resource to low-resource languages, and that this ability is mainly facilitated by scale.

2.3 English as Pivot Improves Multilingual Capabilities of LLMs

One of the findings of Shi et al. (2022) is that CoT prompting with intermediate reasoning steps in English outperforms native CoT prompting with steps in the target language. Huang et al. (2023) show that conversational models such as ChatGPT and Llama-2 also benefit from using English as a pivot language – asking the model to first retell the request in English improves performance on non-English tasks. Notably, this strategy eliminates the need for few-shot examples, meaning that the ability to translate between English and the target language must have been learned during (pre)training rather than from parallel examples provided in the context. Zhang et al. (2024) instruction-tune pretrained LLMs to first process instructions in the pivot language English and then produce responses in the target language.

The phenomenon has been explicitly studied by Zhang et al. (2023), who show that ChatGPT behaves similarly to subordinate bilinguals whose representation of knowledge is strongly biased toward English and, as a consequence, translates all non-English inputs to English. Wendler et al. (2024) investigate the latent representations of token embeddings of LLaMA 2 and find that in the middle layers, these are closer to English tokens, and only in the final layers shift towards target language tokens. They interpret this result as the "concept space" being closer to English.

3 Datasets

The selection of benchmark tasks is determined by the availability of datasets for our target languages. In total, we evaluate the models on five tasks using nine datasets. These datasets primarily originate from cross-lingual benchmarks that include multiple languages. For our experiments, we

task	datasets	est	fin	vro	kpv	liv
machine translation	FLORES-200 (NLLB Team, 2022), SMUGRI-FLORES (Yankovskaya et al., 2023)	✓	✓	✓	✓	✓
multiple choice QA	Belebele (Bandarkar et al., 2024), Belebele-smugri (Purason et al., 2024)	✓	✓	✓	✓	✓
text classification	SIB-200 (Adelani et al., 2024), SIB-smugri (Purason et al., 2024)	✓	✓	✓	✓	✓
extractive QA	EstQA (Käver, 2021), TyDiQA (Clark et al., 2020)	✓	✓			
commonsense reasoning	XCOPA (Ponti et al., 2020)	✓				

Table 1: Tasks and datasets used for benchmarking the models.

use only the subsets that correspond to the selected target languages. A summary of the datasets, tasks and their language coverage is provided in Table 1.

Machine Translation (MT) Our evaluation includes translation tasks from low-resource FiU languages to English. For this purpose, we use the FLORES-200 benchmark (NLLB Team, 2022), which includes Estonian and Finnish, and the FLORES-SMUGRI dataset (Yankovskaya et al., 2023), which translates the first 250 sentences from FLORES-200 to ten low-resource FiU languages, including Komi, Võro, and Livonian. To ensure consistency, we use only the first 250 sentences of FLORES-200 for Estonian and Finnish as well.

Multiple choice QA This task involves selecting the correct answer from a set of options, given a passage, a question, and possible answer choices. We use the Belebele dataset (Bandarkar et al., 2024), which augments paragraphs from the FLORES-200 benchmark with corresponding questions and answer choices. Among its 122 languages, Belebele includes Estonian and Finnish. Purason et al. (2024) further extend the dataset to cover Võro, Livonian, and Komi, resulting in a total of 127 examples per language. For consistency, we use the same number of examples for Estonian and Finnish.

Topic classification We use the massively multilingual text classification benchmark SIB-200 (Adelani et al., 2024), which bases on the FLORES-200 benchmark and comprises 125 examples per language. This benchmark involves classifying sentences from FLORES-200 into seven categories. Purason et al. (2024) extend it to include Võro, Livonian, and Komi.

Extractive QA It is a task in which the objective is to identify a snippet from a given passage that answers a given question. There exists an Estonian dataset for this task, EstQA (Käver, 2021) which includes 603 test examples, each potentially featuring multiple golden answers. In our evaluation, however, we consider only the first answer for each example. Finnish is included into the multilingual dataset TyDiQA (Clark et al., 2020) covering eight typologically diverse languages. Both of these datasets are translation-free, meaning they are created directly in the target language rather than translated from English. In our experiments, we use Finnish samples from the `secondary-task` subset of TyDiQA, where the task format is similar to EstQA. This subset contains 782 Finnish test examples.

Commonsense reasoning Reasoning skills have been observed to be less trivially transferable across languages than question-answering abilities (Kuulmets et al., 2024; Zhu et al., 2024; Huang et al., 2023). To avoid creating a misleading impression of the models' capabilities, it is essential to include reasoning datasets in our evaluation benchmarks. To the best of our knowledge, only one such benchmark incorporates a Finno-Ugric language: XCOPA (Ponti et al., 2020), which includes Estonian. XCOPA requires models to identify which of two answer choices most plausibly represents the cause or effect of a given premise. The test dataset comprises 500 examples.

4 Methodology

For tasks that do not require open-ended text generation (e.g., Belebele, SIB, XCOPA), performance is evaluated by calculating the log likelihood of each possible answer choice and selecting the most likely one as the prediction. In contrast, tasks requiring open-ended text generation, such as FLORES, extractive QA, we use greedy decoding to generate predictions.

We report the results both in zero-shot and few-shot setting where we add either 1, 3 or 5 input-output pairs to the prompt to provide the model with task-specific guidance. Additionally, we investigate the impact of CoT prompting, which guides the model to generate intermediate reasoning steps before producing the final answer. Drawing inspiration from Shi et al. (2022), the intermediate steps require translating the input into English, identifying the answer in English, and translating it back to the target language. CoT prompting can also be used both in zero-shot[4] and few-shot settings. In the zero-shot setting, the prompt ends with *"Let's think step-by-step"* (Kojima et al., 2022), while in the few-shot setting, this is followed by explicit reasoning steps. Figure 1 illustrates model input and output in one-shot setting with and without CoT.

We use regexes to extract answers from the generated text in tasks requiring decoding. Although this approach may occasionally produce false negatives, the models generally adhere well to the output format in few-shot settings. We implement all evaluation strategies with `lm-eval-harness` framework (Gao et al., 2024) and make the task configurations publicly available.[5]

5 Results

5.1 Main Results

Table 2 shows 5-shot results (without CoT) across all tasks and models. In general, Llama 2 7B and Llama 2 13B perform significantly worse on the observed FiU languages than the Llama 3.1 family models. The exception is Finnish, on which the Llama 2 models are notably better than on the other FiU languages. This may be due to the larger amount of Finnish data in the Llama 2 training dataset (Touvron et al., 2023) when compared to data in other FiU languages. However, both Llama-2 7B and Llama 2 13B still appear weak on Finnish when compared to other models.

Llama-2 70B shows notable improvements over Llama 2 7B and Llama 2 13B on Estonian and Finnish across all tasks. The results for Belebele and SIB also indicate improvement for Võro, though the improvement in machine translation (FLORES) is less pronounced. Additionally, SIB appears to be generally too easy of a benchmark for the models, as Llama 2 7B already achieves 86% accuracy for Finnish. For other languages, the benchmark saturates with Llama 2 70B. For this reason, we exclude SIB from further analysis. Finally, we observe that Llama 2 models are the weakest on Komi and Livonian.

	L2-7B	L2-13B	L2-70B	L3.1-8B	L3.1-70B
SIB					
liv	64.8	61.6	83.2	74.4	77.6
kpv	68.0	59.2	83.2	77.6	87.2
vro	64.8	59.2	85.6	86.4	86.4
est	69.6	68.0	88.8	89.6	89.6
fin	85.6	81.6	91.2	87.2	89.6
Belebele					
liv	26.23	35.25	36.89	37.70	42.62
kpv	27.87	31.15	34.43	52.46	73.77
vro	27.05	32.79	44.26	50.82	73.77
est	28.69	36.07	66.39	68.03	88.52
fin	44.26	54.92	86.89	74.59	91.80
XCOPA					
est	49.2	51.8	67.6	69.2	92.6
FLORES (FiU → En)					
liv	6.8	9.3	12.0	10.5	16.1
kpv	5.4	6.0	7.3	10.3	21.9
vro	7.8	9.1	12.9	16.7	30.3
est	12.6	17.8	26.9	35.3	41.0
fin	29.6	31.9	34.6	32.0	37.1
Extractive QA					
exact match					
est	21.89	34.33	49.25	50.75	52.74
fin	51.66	48.34	53.45	58.31	47.06
F1					
est	35.35	51.39	66.72	70.87	73.76
fin	70.63	70.36	74.65	75.44	72.98
BERTScore F1 (Zhang* et al., 2020)					
est	76.88	82.95	88.86	91.76	93.02
fin	88.50	87.95	89.60	90.63	88.67

Table 2: 5-shot results on all tasks. Accuracy is reported for SIB, Belebele and XCOPA. BLEU is reported for FLORES. BERTScore F1 was calculated using `bert-base-multilingual-cased`.

We notice that on Estonian and Finnish, Llama 2 70B is competitive with Llama 3.1 8B despite the latter being nearly nine times smaller, although Llama-3.1 8B appears to slightly underperform on Finnish, as indicated by the results of Belebele and FLORES.

When comparing Llama-3.1 8B to Llama-3.1 70B, the larger model clearly outperforms the smaller one on Belebele, FLORES, and XCOPA.

[4] We leave zero-shot CoT for future research.

[5] https://github.com/TartuNLP/smugri-lm-eval-configs

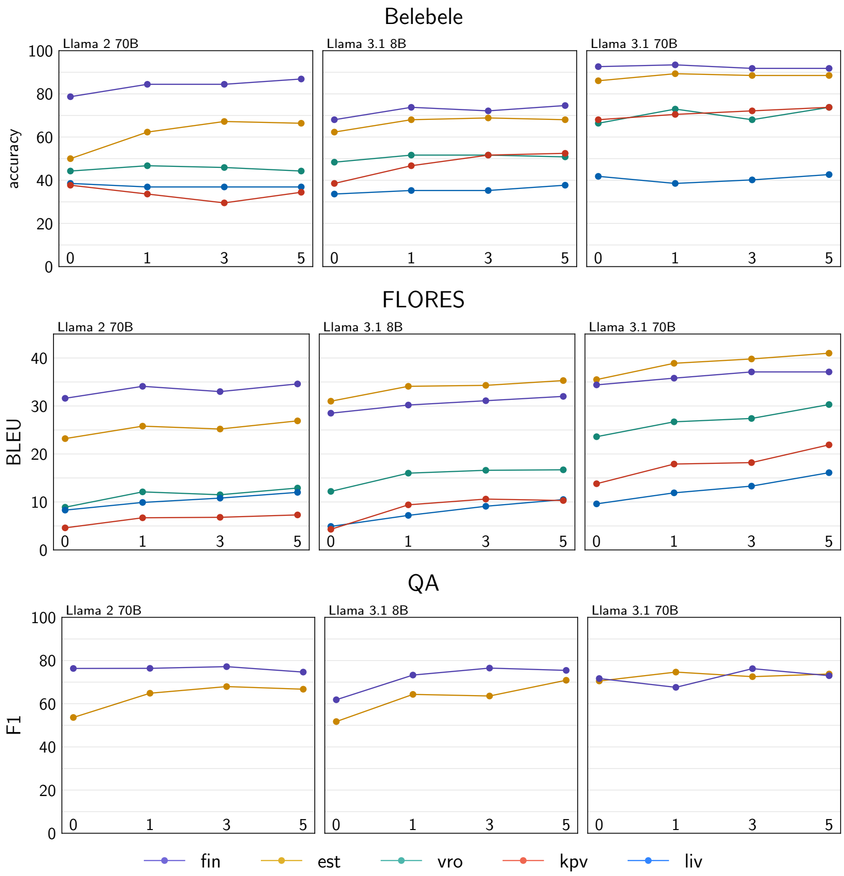

Figure 2: Effect of few-shot examples in 0, 1, 3 and 5-shot setting.

For Estonian and Finnish, the Llama-3.1 70B achieves nearly 90% accuracy on Belebele and XCOPA, along with very strong BLEU scores on the FLORES dataset. The improvements are also significant for extremely low-resource languages Võro, Komi and Livonian.

5.2 The Effect of Few-Shot Examples

We analyze the impact of few-shot examples on the models' ability to solve tasks in FiU languages. We limit this analysis to three models: Llama 2 70B, Llama 3.1 8B, and Llama 3.1 70B due to their superior performance.

Figure 2 illustrates the results. For Belebele and QA tasks, one-shot prompting generally improves performance compared to zero-shot prompting. However, the gains from adding three or five examples vary significantly across tasks and languages. Notably, the improvements from few-shot examples are particularly inconsistent on the Finnish QA task with Llama-3.1 70B.

In contrast, on FLORES benchmark, the improvements are more consistent as the number of examples increases. Notably, Llama-3.1 70B shows substantial gains when translating from Võro, Livonian, and Komi to English, with improvements of 6.6 BLEU points for Võro, 6.6 for Livonian, and 8.1 for Komi when using five examples compared to zero-shot prompting.

To conclude, few-shot prompting can yield notable gains in some cases—such as a 17% improvement for Estonian on Belebele with three examples and using Llama 2 70B as the base model. However, these gains are inconsistent and smaller compared to the improvements achieved by using a stronger base model. For instance, the zero-shot performance for Estonian on Belebele with Llama 3.1 70B surpasses the 3-shot performance of Llama 2 70B. This highlights the greater potential of stronger base models over prompt engineering the weaker models.

5.3 The Effect of CoT Prompting

We analyze the impact of CoT prompting across three tasks: Belebele, QA, and XCOPA. Due to the significant increase in the input length with additional examples, we only compare one-shot prompting with one-shot CoT prompting for Bele-

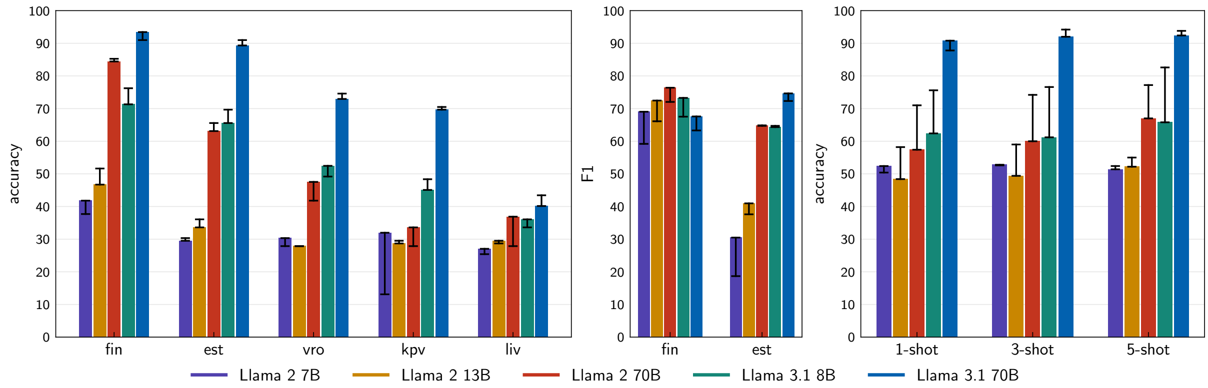

Figure 3: Comparison of CoT prompting and few-shot prompting on Belebele (left, 1-shot), QA (middle, 1-shot) and XCOPA (right, 1-, 3- and 5-shot). The bars shows the scores with few-shot prompting. Horizontal line (–) indicates the score with few-shot CoT prompting with the same number of shots.

bele and QA. For XCOPA we consider 1-, 3-, and 5-shot scenarios.

Figure 4 shows the results. In Belebele task, Llama 2 13B, Llama 2 70B and Llama 3.1 8B benefit from CoT prompting in case of Estonian and Finnish. With the same models the effect of CoT prompting to Võro, Livonian and Komi is mostly negative. Llama 2 7B shows negative or minimal positive gains on all languages. Thi can be explained with the weak translation skills of Llama 2 7B. On the other hand, Llama 3.1 70B has very strong translation skills, yet CoT prompting yields smaller positive improvement than weaker models. This suggests the strong cross-lingual capabilities of Llama 3.1 70B that mitigate the need for CoT prompting.

For the QA task, CoT prompting consistently results in lower performance. This could be attributed to the nature of the extractive QA task, which requires the output to precisely match the correct text snippet. The intermediate translation steps involved in CoT prompting may lead to slight alterations in the morphological form of the answer, causing a mismatch with the expected output.

In XCOPA, we see mostly positive improvements from CoT prompting, with even Llama 2 13B benefiting, while Llama 2 7B does not. The average improvement across all shots for Llama 2 70B and Llama 3.1 8B is 14%. However, the benefit of CoT prompting decreases significantly for Llama 3.1 70B, following the trend observed in the Belebele task.

These observations naturally raise the question of whether there is a correlation between a model's translation capability and its ability to benefit from CoT prompting. To answer that question, we plot the 1-shot BLEU scores of FiU $\rightarrow$ English translation direction against the gains from 1-shot CoT prompting over 1-shot prompting (Figure 4). As shown in the plot, there is no strong correlation between machine translation quality and CoT gains. Interestingly, CoT prompting can provide improvements over few-shot prompting, even for models with weak translation capabilities. However, it also appears that CoT prompting is more likely to degrade performance than enhance it.

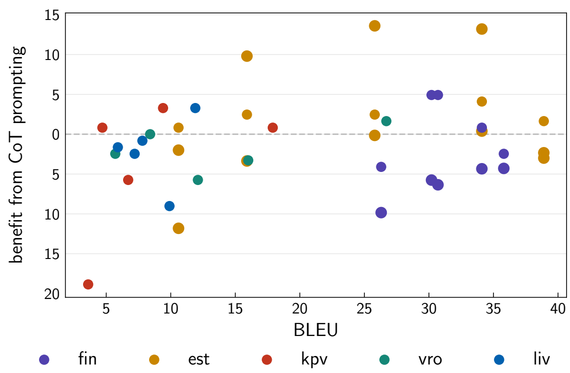

Figure 4: 1-shot BLEU scores for FiU $\rightarrow$ English translation (x-axis) compared with gains from 1-shot CoT prompting over 1-shot prompting (y-axis). Each dot represents a specific Llama model on a specific task and language. Tasks include Belebele, QA, and XCOPA.

Our findings align with Sprague et al. (2024), whose experiments and extensive meta-analysis of existing studies show that CoT provides significant benefits on tasks involving math and logic but offers much smaller gains for other types of tasks.

	Belebele			FLORES			XCOPA			QA		
	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	L3.1
liv	26.23	23.77	**37.70**	6.76	7.70	**10.50**	-	-	-	-	-	-
vro	27.05	31.97	**50.82**	7.83	16.23	**16.72**	-	-	-	-	-	-
kpv	27.87	24.59	**52.46**	5.36	3.64	**10.32**	-	-	-	-	-	-
est	28.69	36.89	**68.03**	12.65	34.29	**35.28**	49.20	68.20	**69.00**	35.35	63.76	**70.87**
fin	44.26	27.87	**74.59**	29.63	18.36	**31.97**	-	-	-	70.63	56.32	**75.44**
avg	30.82	29.02	**56.72**	12.44	16.04	**20.96**	49.20	68.20	**69.00**	52.99	60.04	**73.16**

Table 3: Comparison of five-shot results of **L**lama **2** 7B, **Llam**mas-base and **L**lama **3.1** 8B. F1 score is reported for QA.

6 Comparison With Other Models

6.1 Mistral NeMo

We compare Llama 3.1 8B with its competitor, the 12B-parameter model Mistral NeMo (Jiang et al., 2024), across all tasks except SIB. Both models are evaluated in zero-shot and five-shot settings to assess their ability to perform with and without examples. Results for the zero-shot setting are shown in Table 4, while the five-shot results are presented in Table 5. Note that zero-shot results for the QA task are not reported, as this task is typically evaluated in a few-shot setting due to significantly lower performance in zero-shot scenarios.

	Belebele		FLORES		XCOPA	
	L3.1	MN	L3.1	MN	L3.1	MN
liv	33.61	**35.25**	4.91	**5.85**	-	-
vro	48.36	**50.82**	**12.19**	8.18	-	-
kpv	**38.52**	36.89	**8.18**	3.45	-	-
est	62.30	**74.59**	31.00	**33.04**	**56.80**	56.40
fin	68.03	**74.59**	28.54	**30.39**	-	-
avg	50.16	**54.43**	**16.96**	16.18	**56.80**	56.40

Table 4: Comparison of zero-shot results of **L**lama-**3.1** 8B and **M**istral **N**eMo.

	Belebele		FLORES		XCOPA		QA	
	L3.1	MN	L3.1	MN	L3.1	MN	L3.1	MN
liv	**37.70**	**37.70**	**10.50**	10.10	-	-	-	-
vro	**50.82**	50.00	**16.72**	12.55	-	-	-	-
kpv	**52.46**	34.43	**10.32**	6.01	-	-	-	-
est	68.03	**83.61**	**35.28**	32.28	69.20	**71.60**	70.87	**71.86**
fin	74.59	**78.69**	31.97	**33.24**	-	-	75.44	**77.39**
avg	56.72	**56.89**	**20.96**	18.83	69.20	**71.60**	73.16	**74.63**

Table 5: Comparison of five-shot results of **L**lama-**3.1** 8B and **M**istral **N**eMo. F1 score is reported for QA.

The results show that Mistral NeMo and Llama 3.1 8B perform similarly on FiU languages in the zero-shot setting, though Mistral NeMo is over 4% better on the Belebele task. In the five-shot setting, Mistral NeMo outperforms Llama 3.1 8B on three out of four tasks, except for machine translation, where Llama 3.1 8B demonstrates a stronger ability to learn from examples. Overall, Mistral NeMo excels in Finnish and Estonian, while Llama 3.1 8B appears slightly stronger in extremely low-resource FiU languages. Notably, Llama 3.1 8B consistently outperforms Mistral NeMo in Komi, which, unlike the other languages, uses the Cyrillic script.

	Belebele			FLORES			XCOPA		
	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	L3.1
liv	24.59	**38.52**	33.61	4.74	4.62	**4.91**	-	-	-
vro	23.77	33.61	**48.36**	4.61	9.92	**12.19**	-	-	-
kpv	26.23	29.51	**38.52**	2.88	1.44	**8.18**	-	-	-
est	22.95	39.34	**62.30**	8.53	28.90	**31.0**	48.80	**56.60**	**56.60**
fin	32.79	34.43	**68.03**	27.16	11.57	**28.54**	-	-	-
avg	26.07	35.08	**50.16**	9.59	11.29	**16.96**	48.80	**56.60**	**56.60**

Table 6: Comparison of zero-shot results of **L**lama **2** 7B, **Llam**mas-base and **L**lama **3.1** 8B.

6.2 Llammas

We compare Llama 2 7B with Llammas (Kuulmets et al., 2024), which is an adaptation of Llama 2 7B to Estonian with additional pretraining of 5B tokens of Estonian-centric data. We also include comparative size Llama 2.1 8B in this comparison. The results are presented in Table 6 and Table 3.

Unsurprisingly, Llammas outperforms Llama 2 7B on Estonian by a significant margin; however, its performance on Finnish, in general, decreases substantially. As indicated in the tables presented in Section 5.1, Llama 2 7B already demonstrates some capability in solving tasks in Finnish, unlike in other FiU languages. This suggests that continued pretraining on Estonian notably damages this capability.

Llammas consistently outperforms Llama 2 7B on Võro, which is not surprising given the linguistic similarities between Võro and Estonian. The comparison between Livonian and Komi is less clear in determining which model performs better. However, Llama 3.1 8B surpasses both models by a large margin, except on the Belebele task in Livonian. Notably, Llama 3.1 8B outperforms Llammas even on Estonian, demonstrating that language-specific adaptation of a weaker base model cannot compete with a stronger, unadapted base model.

7 Conclusion

We evaluated the Llama 2 and multilingual Llama 3.1 family models on five Finno-Ugric languages with varying amounts of available resources. Our results show that Llama 2 7B and 13B perform poorly on most languages, except for Finnish, where they achieve moderate results. In contrast, the Llama 3.1 family models demonstrate impressive performance, even for extremely low-resource languages like Võro and Komi.

The comparison of zero-shot and few-shot prompting indicates that few-shot prompting is beneficial across all languages. However, increasing the number of examples does not always lead to better performance. Similarly, few-shot CoT prompting brings substantial benefits for tasks like commonsense reasoning but negatively affects others, such as QA. Notably, the strongest model, Llama 3.1 70B, benefits less from CoT prompting on tasks where it helps weaker models, suggesting that strong cross-lingual capabilities reduce reliance on CoT prompting.

Outstanding results in MT, XCOPA, and Belebele for Estonian and Finnish highlight the need for stronger benchmarks to better assess the capabilities and limitations of these models. The surprisingly strong results from Llama 3.1 70B on Komi and Võro, despite extremely limited resources, demonstrate effective cross-lingual knowledge transfer and reduce the dependence on large target-language datasets for reasonable performance.

Finally, our comparison with Mistral NeMo suggests that the latter outperforms Llama 3.1 8B in Estonian and Finnish. Furthermore, our analysis of Llama models versus Llammas shows that a stronger, general-purpose base model consistently outperforms a weaker base model adapted to a specific language, emphasizing the critical role of the base model in successful language adaptation.

Acknowledgements

This work was partially supported by the Estonian Research Council grant PRG2006 as well as the National Programme of Estonian Language Technology grant EKTB104. All computations were performed in the High Performance Computing Center of the University of Tartu.

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.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario

Amodei. 2020. Language models are few-shot learners.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

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

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities.

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. 2024. A framework for few-shot language model evaluation.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4476–4494, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Albert Jiang, Alexandre Sablayrolles, Alexis Tacnet, Alok Kothari, Antoine Roux, Arthur Mensch, Audrey Herblin-Stoop, Augustin Garreau, Austin Birky, Bam4d, Baptiste Bout, Baudouin de Monicault, Blanche Savary, Carole Rambaud, Caroline Feldman, Devendra Singh Chaplot, Diego de las Casas, Eleonore Arcelin, Emma Bou Hanna, Etienne Metzger, Gaspard Blanchet, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Harizo Rajaona, Henri Roussez, Hichem Sattouf, Ian Mack, Jean-Malo Delignon, Jessica Chudnovsky, Justus Murke, Kartik Khandelwal, Lawrence Stewart, Louis Martin, Louis Ternon, Lucile Saulnier, Lélio Renard Lavaud, Margaret Jennings, Marie Pellat, Marie Torelli, Marie-Anne Lachaux, Marjorie Janiewicz, Mickaël Seznec, Nicolas Schuhl, Niklas Muhs, Olivier de Garrigues, Patrick von Platen, Paul Jacob, Pauline Buche, Pavan Kumar Reddy, Perry Savas, Pierre Stock, Romain Sauvestre, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Sophia Yang, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Wang, Théophile Gervet, Timothée Lacroix, Valera Nemychnikova, Wendy Shang, William El Sayed, and William Marshall. 2024. Mistral-nemo-base-2407.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

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

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning.

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

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.

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.

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.

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

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

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2024. Llms for extremely low-resource finno-ugric languages.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource finno-ugric languages. In *The 24rd Nordic Conference on Computational Linguistics*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.