

SMUGRI-COPA: Evaluating Commonsense Reasoning for Low-Resource Finnic Languages

Taido Purason¹, Marili Tomingas², Janek Vaab², Mark Fishel¹

¹Institute of Computer Science

²Institute of Estonian and General Linguistics

University of Tartu

{name.surname}@ut.ee

Abstract

We introduce SMUGRI-COPA, a commonsense reasoning benchmark for two heavily under-resourced Finnic languages: Võro and Livonian. Existing evaluation resources for these languages rely on a translated subset of FLORES sentences, limiting linguistic diversity and the evaluation of broader language understanding. To address this gap, we human-translate the COPA (Choice of Plausible Alternatives) benchmark into Võro and Livonian, creating, to our knowledge, the largest discriminative benchmark currently available for these languages. We evaluate a wide range of proprietary and open-weight language models on SMUGRI-COPA and compare the results with Estonian XCOPA. Our results reveal a substantial performance gap between Estonian and the lower-resource languages, with most models failing to significantly outperform random guessing on Livonian. We further find that enabling reasoning for GPT-5 substantially improves accuracy for Võro and Livonian, while its reasoning summaries show frequent reliance on explicit translation of the tasks. Our work highlights the challenges of evaluating endangered language understanding and provides a new resource for multilingual evaluation research.

 [tartuNLP/smugri-copa](https://github.com/tartuNLP/smugri-copa)

1 Introduction

Recent years have seen increasing efforts to support low-resource Finno-Ugric (FiU) languages, with the vast majority of work focusing on machine translation (MT, Yankovskaya et al., 2023; Tars et al., 2022, 2021; Rikters et al., 2022; Pashchenko et al., 2025). The limited research on large language models (LLMs, Purason et al., 2025; Samuel et al., 2025) has been hindered by the lack of suitable evaluation resources. For most low-resource FiU languages, benchmarks are either absent or

derived from a small set of FLORES sentences (NLLB Team et al., 2022) either directly for benchmarking MT (Yankovskaya et al., 2023) or through translated SIB-200 (Adelani et al., 2024, topic classification) and Belebele (Bandarkar et al., 2024, multiple-choice QA) published by Purason et al. (2025). This reliance on translated FLORES is problematic. Since many source sentences originate from English Wikinews, they often contain specialized, international, or culturally distant vocabulary without well-established equivalents in extremely low-resource languages, making translations more difficult, less natural, and less representative of actual language use. As a result, it is challenging to assess LLM performance beyond translation or to track real progress in broader natural language understanding.

To investigate alternatives, we focus on two extremely low-resource FiU languages—Võro and Livonian—both of which currently rely exclusively on FLORES-based evaluation with SMUGRI-FLORES (Yankovskaya et al., 2023), SIB-SMUGRI, and Belebele-SMUGRI (Purason et al., 2025). While Võro and Livonian are both low-resource languages, they differ by orders of magnitude in their speaker populations and available resources. Yet each has an active community and prior work on LLM training and evaluation (Purason et al., 2025) to serve as a point of comparison. These characteristics motivate the need for evaluation benchmarks that are feasible to adapt to extremely low-resource languages while capturing language understanding beyond translation.

To address this, we translate the Choice of Plausible Alternatives (COPA, Roemmele et al., 2011) dataset and investigate whether it can serve as a suitable evaluation benchmark for Võro and Livonian, enabling more meaningful LM evaluation. Since COPA is a discriminative benchmark, we focus on assessing the models’ ability to understand the language, rather than their ability to generate

Dataset		Premise	A	B
COPA	ENG	The girl found a bug in her cereal.	She poured milk in the bowl.	She lost her appetite.
XCOPA	EST	Tüdruk leidis oma helveste seest putuka.	Ta kallas piima kaussi.	Ta kaotas oma isu.
SMUGRI-COPA (ours)				
	VRO	Tüttrik löüd' krõbuskidõ seest mutiga.	Tä pand' piimä kaussi.	Täl läts iso ärkiq.
	LIV	Neitst lēdziz eņš kōpkōd sizāld kukkiz.	Ta valīz semđi koušō.	Ta kōtiz eņš aptīt.

Table 1: An effect-type example from the COPA validation set with XCOPA and our translated versions. A and B are options from which the model must choose the most plausible one (*What happened as a result?* in this case), given the premise.

fluent text in those languages.

Our contributions are as follows: **(1)** we release SMUGRI-COPA¹ (*CC BY 4.0*), a human-translated extension of XCOPA (Ponti et al., 2020) for the endangered low-resource Finnic languages Võro and Livonian, creating, to our knowledge, the largest discriminative language model evaluation benchmark currently available for these languages; **(2)** we provide a systematic evaluation of both proprietary and open-weight instruction-tuned and base language models on SMUGRI-COPA, revealing a substantial performance gap between Estonian and the two low-resource languages and showing that enabling reasoning in GPT-5 substantially improves accuracy scores for Võro and Livonian, while reasoning summaries reveal frequent reliance on translation; and **(3)** we provide a qualitative analysis of challenges involved in translating benchmarks to endangered low-resource languages.

2 Background

Võro and Livonian are closely related both geographically and linguistically, belonging to the Finnic branch of the Uralic languages. Among higher-resource Finnic languages, they are most closely related to Estonian. Like Estonian, both Livonian and Võro are also considered morphologically rich languages.

According to the resource-based taxonomy of languages by Joshi et al. (2020), Livonian is in the lowest-resourced category — *The Left-Behinds* (0/6) — and Võro is in the second-lowest category — *The Scraping-bys* (1/6). Moseley (2010) considers Livonian *critically endangered*, and Võro *definitely endangered*.

According to the 2021 Population and Housing Census, 99,560 people in Estonia have reported proficiency in Võro South Estonian (Statistics Estonia,

2022, including 26,220 Seto speakers). However, some estimates suggest that the number of active, fluent Võro speakers is likely below 50,000 (Plado et al., 2023).

The number of speakers of Livonian as a second language has previously been estimated at 30 (Druviete and Kļava, 2018). However, compared to many other critically endangered languages, as a result of almost two centuries of collecting language data, Livonian is documented relatively well, especially given its small number of speakers (Ernštreits et al., 2022).

COPA (Roemmele et al., 2011) is an English benchmark that evaluates a model’s ability to choose the more plausible of two alternatives given a premise. There are two types of tasks in COPA: 1) *What happened as a result?* 2) *What was the cause of this?* Table 1 demonstrates an example of type 1) task from COPA and the translations from our work and XCOPA (Ponti et al., 2020). The given example demonstrates that, while both options are possible, one is clearly more plausible as the effect (*B*).

3 Related Work

NLP for Võro and Livonian. There has been growing interest in developing language technologies for low-resource Finno-Ugric languages such as Võro and Livonian. Earlier work, including Võro and/or Livonian, has primarily focused on machine translation (Yankovskaya et al., 2023; Tars et al., 2022; Rikters et al., 2022; Tars et al., 2021; Ernštreits et al., 2022). In addition to text-based approaches, speech technologies have also been explored, including the development of neural text-to-speech systems for Võro (Rätsep and Fishel, 2023). More recently, the development of large language models (LLMs) supporting these languages (Purason et al., 2025) has been explored, alongside evaluation efforts (Kuulmets et al., 2025).

¹huggingface.co/datasets/tartuNLP/smugri-copa

Existing benchmarks. One of the first benchmarks for these languages is SMUGRI-FLORES (Yankovskaya et al., 2023), which consists of translations of 250 sentences (out of 1012) from the FLORES (NLLB Team et al., 2022) machine translation benchmark. Since both SIB-200 (Adelani et al., 2024) and Belebele (Bandarkar et al., 2024) are partially constructed from FLORES sentences, Purason et al. (2025) were able to reuse these translations with minimal additional translation to extend those benchmarks to Võro, Livonian, and Komi, enabling automatic discriminative evaluation of language models. In addition, they introduced SMUGRI-MT-Bench for human evaluation of multi-turn conversational performance in LLMs.

Translations of COPA. XCOPA (Ponti et al., 2020) extended COPA (Roemmele et al., 2011) to 11 languages, including Estonian, via translation. Since then, the benchmark has been adapted to a wide range of additional languages (Gonzalez-Agirre et al., 2024; Ljubešić et al., 2024; Wibowo et al., 2024; Doddapaneni et al., 2023; Žagar and Robnik-Šikonja, 2022; Ljubešić and Lauc, 2021; Shavrina et al., 2020). The work most closely related to ours is Ljubešić et al. (2024), which extends COPA to South Slavic dialects and was also used in the VarDial 2024 Dialect-COPA shared task (Chifu et al., 2024).

4 Methodology

4.1 Why was COPA chosen?

(1) Ease of translation. Many existing benchmarks contain domain-specific or culturally distant content that is difficult to translate naturally into extremely low-resource languages. FLORES-based benchmarks such as Belebele and SIB-200 are one example, often requiring specialized vocabulary (e.g., medical terms) that may not exist in the target language. As a result, such benchmarks may be less representative of actual language use and thus difficult to translate. In contrast, COPA focuses on short everyday situations, making translation substantially easier and more natural.

(2) Diversity. Existing automatic benchmarks for these languages all rely on the 250 FLORES sentences, which limits linguistic and topical variation. Translating COPA introduces a new set of sentence contexts, thereby increasing benchmark diversity.

(3) Applicability. COPA can be used to evaluate both encoder-only and decoder-only language mod-

els. For decoder-only models, it is applicable to both base and instruction-tuned models.

(4) Size. In extremely low-resource settings, translating thousands of examples is often infeasible due to the limited availability of translators. At the same time, benchmark size strongly affects evaluation uncertainty. A 127-example benchmark like Belebele-SMUGRI typically yields confidence intervals of roughly $\pm 7\text{--}9\%$ in the 20–80% accuracy range, making comparisons between similarly performing models unreliable. In contrast, the 500-example XCOPA test set has confidence intervals below $\pm 4.5\%$, enabling more reliable comparisons (see Figure 1).

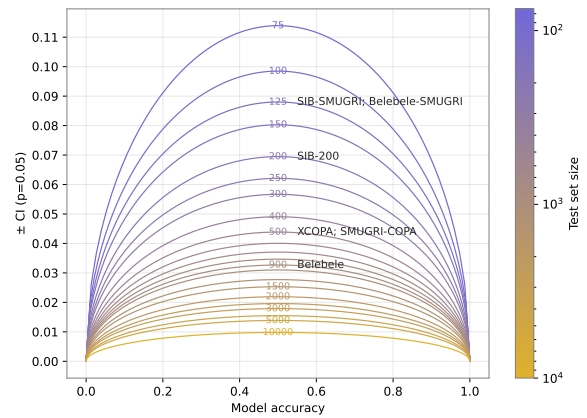


Figure 1: Estimated 95% confidence intervals for various dataset sizes. Following the lm-evaluation-harness (Gao et al., 2024), we estimate confidence intervals using a normal approximation based on the standard error of accuracy.

4.2 Translation

The dataset was translated by a native speaker for Võro and a fluent speaker for Livonian (due to a lack of native speakers), following detailed instructions (see Appendix A.1). The translators were compensated and additionally offered an opportunity to contribute as authors. Translators were asked to maintain the correct meaning of each example, ensuring that the cause-and-effect relationship remained intact, while creating sentences that sound natural in the target language. When a direct translation was not possible, they were allowed to adjust the wording to better fit the language. Translators also flagged any cases where the original answer key seemed incorrect or where the translation would produce incorrect examples in the target-language context. The translators were given both the English source of COPA (Roemmele

et al., 2011) and also the existing Estonian translation from XCOPA (Ponti et al., 2020), since Estonian is closely related to Võro and Livonian. The XCOPA structure, including validation/test split, labels, and the option order, is kept intact. The resulting dataset will be released under *Creative Commons Attribution 4.0* following XCOPA (Ponti et al., 2020). Each language was translated by one translator, and the resulting translations were not independently validated by another speaker (see [Limitations](#)).

We deviate from the translation methodology of XCOPA (Ponti et al., 2020) by giving translators the correct label from the start. This allows them to judge how best to translate each example so that the intended answer remains correct within the linguistic and cultural context of the target languages, and to flag items whose correctness becomes unclear in that context. Because the pool of qualified translators for these low-resource languages is small, providing the label upfront also helps make their time and effort more efficient.

4.3 Experiment setup

We evaluate both open- and closed-weight instruction-tuned models, as well as open-weight base models. COPA can be framed in a multiple-choice or a cloze format (see Figure 2) and evaluated with both log-probability-based evaluation and generative evaluation, depending on whether there is access to the token log-probabilities or not.

The base models were evaluated using `lm-evaluation-harness` (Gao et al., 2024). We evaluate models in 7–9B and 70B ranges, covering predominantly English-centric (Grattafiori et al., 2024, Llama-3), multilingual (Martins et al., 2025, EuroLLM), massively multilingual (Hernández-Cano et al., 2026, Apertus, includes VRO and LIV in training data), Estonian-specialized (Dorkin et al., 2026, EstLLM) and Võro-and-Livonian-specialized (Purason et al., 2025, Llama-SMUGRI-7B) models.

We evaluate OpenAI models through Azure OpenAI API and Gemini through Gemini API with the prompt outlined in Appendix B.1, using *temperature* = 0, except for reasoning modes, where we use the recommended value. We also evaluate competitive open-weight models DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Apertus-70B-Instruct-2509 (Hernández-Cano et al., 2026), and Gemma-4-31B-it (Gemma Team et al., 2026).

Multiple-choice format

```
{premise}, {connector}:
A. {choice_1}
B. {choice_2}
Answer: A/B
```

Cloze format

```
{premise}, {connector} {choice_1}/{choice_2}
```

Figure 2: The prompt formats used for base model evaluation. The model selects the candidate continuation with the highest log-probability. The connector for COPA could be *because* or *therefore* (in the target language).

We evaluate on the test set and use the validation set for few-shot examples. For calculating standard errors and 95% confidence intervals, we use the implementation from `lm-evaluation-harness`. We consider a model to significantly outperform random guessing when its 95% confidence interval excludes the 50% random baseline. The full list of models and their precise versions is provided in Appendix B.3, Table 12. We also report the Estonian XCOPA (Ponti et al., 2020) for comparison.

5 Results

5.1 The translated dataset

Following XCOPA, the translated dataset consists of a 500-example test set and a 100-example validation set. As shown in Table 2, Livonian and Võro require 2.18× and 1.66× as many GPT-5 tokens as English, respectively, indicating substantially less efficient representation under this tokenizer. This implies higher processing costs for the same text with those models for low-resource languages. Further dataset statistics are listed in Appendix C.

	byte	char	word	GPT-5 tokens
ENG	1	1	1	1
EST	0.97	0.93	0.74	1.42
VRO	1.00	0.90	0.78	1.66
LTV	1.08	0.91	0.79	2.18

Table 2: **How many times longer is the same text compared to English?** Premiums in number of bytes, characters, words and tokens compared to English (c_{tgt}/c_{eng}) for the test set ($n = 500$).

5.2 Benchmarking

5.2.1 Instruction-tuned models

Performance across languages. For zero-shot prompting with proprietary instruction-tuned mod-

Model	EST	VRO	LIV
PROPRIETARY INSTRUCT-MODELS			
GPT-4o	98.0 $\pm$ 1.2	70.0 $\pm$ 4.0	50.0 $\pm$ 4.4
GPT-4.1	98.0 $\pm$ 1.2	69.0 $\pm$ 4.1	51.4 $\pm$ 4.4
GPT-5	97.4 $\pm$ 1.4	68.8 $\pm$ 4.1	53.2 $\pm$ 4.4
GPT-5 [✳] (low)	99.0 $\pm$ 0.9	84.4 $\pm$ 3.2	61.0 $\pm$ 4.3
Gemini-3-flash [✳] (low)	99.4 $\pm$ 0.7	94.0 $\pm$ 2.1	79.6 $\pm$ 3.5
OPEN-WEIGHT INSTRUCT-MODELS			
DeepSeek-V3.2 (685B)	82.8 $\pm$ 3.3	53.0 $\pm$ 4.4	49.2 $\pm$ 4.4
Apertus-70B-Instruct [†]	88.0 $\pm$ 2.9	58.2 $\pm$ 4.3	54.2 $\pm$ 4.4
Gemma-4-31B-it	92.6 $\pm$ 2.3	63.0 $\pm$ 4.2	49.8 $\pm$ 4.4

Table 3: COPA benchmark accuracy (zero-shot). For proprietary models with reasoning enabled[✳], reasoning effort is indicated in parentheses. [†]- models including Võro and Livonian in training data.

els, Estonian achieves a nearly perfect score (see Table 3). On the other hand, the Võro accuracy is noticeably lower compared to Estonian for the same model, and for Livonian, none of the non-reasoning models can significantly outperform the random baseline (50%). We can observe this ranking of Estonian, Võro, and Livonian scores across all model types.

The substantially lower performance of Livonian compared to Võro is consistent with both its extremely limited resource availability and its lower similarity to Estonian, which may reduce cross-lingual transfer. In FineWeb-2 (Penedo et al., 2025), Livonian contains only 159 documents compared to 8200 for Võro (nearly 60× fewer words). Table 4 shows that Võro is more similar to Estonian than Livonian in terms of vocabulary overlap and Levenshtein ratio. Similar observations were also reported by Ljubešić et al. (2024), who found that dialects closer to the corresponding higher-resource language achieved higher COPA performance. Prior work has also identified lexical overlap as a contributor to cross-lingual transfer (Lin et al., 2019; Pires et al., 2019).

Effect of reasoning. Comparing the two GPT-5 configurations, enabling reasoning at low effort increases accuracy from 68.8% to 84.4% for Võro and from 53.2% to 61.0% for Livonian. Furthermore, when looking at Gemini-3-flash with similar reasoning effort, we see even higher scores, making it by far the best model benchmarked. Among the evaluated instruction-tuned models, GPT-5 and Gemini-3-flash are the only models to significantly outperform the random baseline for Livonian. With reasoning disabled, GPT-5 per-

	top-100	Levenshtein ratio	
	overlap	char	word
VRO - EST	15	0.457	0.105
LIV - EST	4	0.403	0.097
VRO - LIV	3	0.314	0.042
EST - ENG	0	0.225	0.000

Table 4: **Similarities of translated COPA texts** (test set, $n = 500$) between languages. Following Ljubešić et al. (2024), we report character and word Levenshtein ratios $(1 - d_{lev}(a, b) / \max(|a|, |b|))$ and the number of overlapping words in the top-100 most frequent words. Higher values indicate higher similarity.

	Translation reasoning	Low understanding	Surface-form reliance
VRO	19/23	4/23	3/23
LIV	21/26	16/26	8/26

Table 5: Manual analysis of GPT-5 reasoning summaries for 50 randomly selected parallel examples. Not all examples yielded a reasoning summary. Low understanding denotes a mostly incorrect interpretation of the example’s meaning. See Appendix B.4 for details.

forms at a similar level to other OpenAI models, suggesting that enabling reasoning is the cause behind the increase in accuracy scores.

When examining GPT-5 reasoning summaries for a subset of 50 examples, we observe that the reasoning summaries frequently describe explicit translation in both languages (Table 5). However, the quality of the demonstrated understanding differed substantially between Võro and Livonian. For Võro, the reasoning summaries generally reflected a correct understanding of the task, with only occasional issues. In contrast, for Livonian, many reasoning summaries revealed limited understanding, often relying on only partially correct translations or shallow associations with related languages. These results show that enabling reasoning in GPT-5 increases benchmark accuracy, but does not necessarily demonstrate correct reasoning in the target language.

Open-weight instruct models. The open-weight instruct models show substantial variation in performance that is not clearly correlated with model size (Table 3). None of the evaluated open-weight instruct models significantly outperform the random baseline on Livonian. For Võro, performance also varies considerably: DeepSeek-V3.2 (685B) does not significantly outperform the random baseline, while the strongest open-weight instruct model in

Model	EST		VRO		LIV	
	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot
CLOZE FORMAT						
Llama-SMUGRI-7B [‡]	72.4 ± 3.9	76.6 ± 3.7	58.4 ± 4.3	60.2 ± 4.3	53.0 ± 4.4	52.8 ± 4.4
Llama-3.1-EstLLM-8B-0525	77.6 ± 3.7	77.6 ± 3.7	55.4 ± 4.4	56.8 ± 4.3	49.0 ± 4.4	50.0 ± 4.4
Apertus-8B-2509 [†]	70.4 ± 4.0	76.2 ± 3.7	54.4 ± 4.4	54.6 ± 4.4	47.2 ± 4.4	45.8 ± 4.4
Llama-3.1-8B	56.6 ± 4.3	61.0 ± 4.3	54.8 ± 4.4	51.0 ± 4.4	48.4 ± 4.4	47.6 ± 4.4
EuroLLM-9B	73.2 ± 3.9	78.8 ± 3.6	54.2 ± 4.4	54.6 ± 4.4	47.6 ± 4.4	47.0 ± 4.4
Apertus-70B-2509 [†]	72.4 ± 3.9	79.4 ± 3.5	56.4 ± 4.4	57.8 ± 4.3	47.0 ± 4.4	48.2 ± 4.4
Llama-3.1-70B	68.8 ± 4.1	75.6 ± 3.8	54.6 ± 4.4	57.0 ± 4.3	48.6 ± 4.4	49.4 ± 4.4
CONTINUED PRE-TRAINING ON VRO+LIV:						
Llama-3.1-EstLLM-8B-VRO-LIV [‡] (<i>ours</i>)	73.8 ± 3.9	76.2 ± 3.7	60.8 ± 4.3	63.4 ± 4.2	52.8 ± 4.4	53.4 ± 4.4
MULTIPLE-CHOICE FORMAT						
Llama-SMUGRI-7B [‡]	55.4 ± 4.4	61.6 ± 4.3	50.6 ± 4.4	54.0 ± 4.4	49.8 ± 4.4	48.4 ± 4.4
Llama-3.1-EstLLM-8B-0525	68.6 ± 4.1	90.0 ± 2.6	51.4 ± 4.4	67.0 ± 4.1	51.8 ± 4.4	52.2 ± 4.4
Apertus-8B-2509 [†]	57.0 ± 4.3	90.2 ± 2.6	51.0 ± 4.4	61.4 ± 4.3	50.2 ± 4.4	52.4 ± 4.4
Llama-3.1-8B	65.2 ± 4.2	70.2 ± 4.0	50.4 ± 4.4	56.4 ± 4.4	49.2 ± 4.4	51.0 ± 4.4
EuroLLM-9B	80.0 ± 3.5	88.4 ± 2.8	53.0 ± 4.4	63.8 ± 4.2	51.4 ± 4.4	51.4 ± 4.4
Apertus-70B-2509 [†]	75.4 ± 3.8	93.6 ± 2.1	53.4 ± 4.4	65.8 ± 4.2	48.8 ± 4.4	45.2 ± 4.4
Llama-3.1-70B	85.8 ± 3.1	95.0 ± 1.9	56.4 ± 4.4	74.2 ± 3.8	51.4 ± 4.4	52.4 ± 4.4
CONTINUED PRE-TRAINING ON VRO+LIV:						
Llama-3.1-EstLLM-8B-VRO-LIV [‡] (<i>ours</i>)	81.0 ± 3.4	89.2 ± 2.7	65.8 ± 4.2	77.6 ± 3.7	50.8 ± 4.4	58.8 ± 4.3

Table 6: Accuracy (%) of the **open-weight base models** on the COPA benchmark for Estonian, Võro and Livonian. Log-probability-based evaluation. [†]- models including Võro and Livonian in training data, [‡]- models targeting Võro and Livonian specifically.

our experiments, Gemma-4-31B-it, reaches 63.0%. Furthermore, all proprietary models we tested outperform the open-weight instruct models on Võro.

5.2.2 Base models

When looking at open-weight base models, we again see that Estonian has the highest accuracy, while Võro lags behind, and Livonian is again not significantly better than random guessing.

Prompt format. Gu et al. (2025) found that during pre-training, the cloze format (CF) performs well in early training; however, the multiple-choice format (MCF) outperforms it later.

With the models we tested for Võro, we see that most achieve higher scores with MCF than with CF for 5-shot prompting, whereas for 0-shot prompting, the difference is smaller. One exception is Llama-SMUGRI-7B, which is explicitly adapted for Võro and Livonian, and achieves the highest score for Võro among previously available base models within the CF setting for both 0- and 5-shot prompting, while being unable to take advantage of the MCF, possibly due to being based on less capable Llama-2-7B compared to the newer models we benchmark.

When looking at how similarly different benchmarks rank base models (see Figure 3), we see that CF correlates more with generative perfor-

mance in the target language, as shown by EST-VRO SMUGRI-FLORES, while MCF correlates more with discriminative multilingual understanding, shown by multiple-choice benchmark Belebele-SMUGRI. Full results for previous benchmarks are listed in Appendix F.

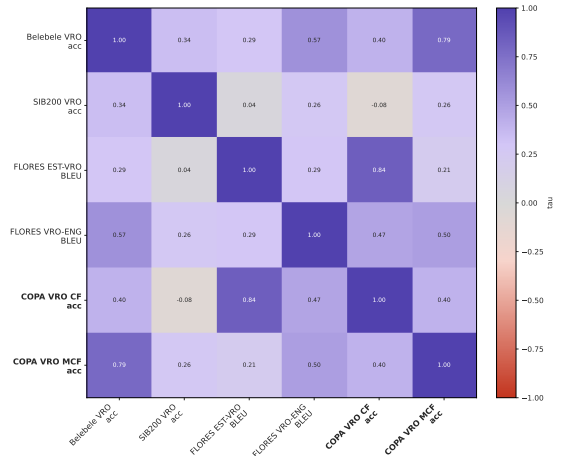


Figure 3: How similarly do different benchmarks rank the 8 base models? Kendall τ rank correlation for various benchmarks in Võro (5-shot).

Continued pre-training. To further assess the suitability of SMUGRI-COPA as a benchmark, we continued pre-training the

strongest Estonian model in the 7–9B range – Llama-3.1-EstLLM-8B-0525 – on *FineWeb-2* Estonian (Penedo et al., 2025), *FineWeb-Edu* (Lozhkov et al., 2024, English) and VRO and LIV data from *SMUGRI*², *Wikipedia* (Wikimedia Foundation, 2025), and *liv4ever* (Rikters et al., 2022), with the target language data making up 50% of training tokens. We list further technical details in Appendix E.

We see that continued pre-training on 22M tokens yields a stronger model that achieves the highest scores among all base models for Võro across all prompting strategies and is the only model to surpass the random baseline on MCF 5-shot prompting for Livonian (see Table 6). This suggests that SMUGRI-COPA can reflect improvements from training on VRO and LIV data.

5.3 Adequacy of Machine Translation

Model	VRO		LIV	
	Human	MT	Human	MT
GPT-4o	70.0 $\pm$ 4.0	79.4 $\pm$ 3.5	50.0 $\pm$ 4.4	55.2 $\pm$ 4.4
GPT-4.1	69.0 $\pm$ 4.1	77.6 $\pm$ 3.7	51.4 $\pm$ 4.4	51.6 $\pm$ 4.4
GPT-5	68.8 $\pm$ 4.1	76.6 $\pm$ 3.7	53.2 $\pm$ 4.4	52.8 $\pm$ 4.4

Table 7: Human-translated vs machine-translated (MT) benchmark accuracies.

Since machine-translated benchmarks have become increasingly common for evaluation, we investigate whether a machine-translated XCOPA could estimate the performance of the human-translated version. The scores of machine-translated³ versions are substantially higher for Võro compared to human translations (see Table 7). These results caution against using machine-translated benchmarks without language-specific validation.

6 Challenges in translating COPA

While in COPA, the scenarios are mostly straightforward and culturally appropriate to translate, there are a few examples where accurate translation can be problematic. We try to consider the **correctness** — if the most plausible choice given the premise remains the same — and how **appropriate** the example is given the culture and the typical usage of the language. We outline the main

translation challenges below and provide additional examples in Appendix D.

6.1 Cultural context

There are only a few potential minor issues with the cultural context. While they generally do not change the correctness of the examples for Võro and Livonian, they are worth considering briefly.

❶ **Example 1** The woman ran out of eggs, therefore: A) she went to the **farm**. B) she went to the **supermarket**.

❷ **Explanation** Getting eggs from a farm in Võro-speaking areas is more common than in Estonian-speaking urban population centers. While getting eggs from a supermarket is still arguably more common, this aspect shifts the balance of the choices and might also be an issue in other languages and cultures.

6.2 Idiomatic expressions

COPA includes some phrases that cannot be translated literally while preserving naturalness. Idiomatic expressions are often realized differently across languages, requiring reformulations to preserve naturalness and correctness.

❶ **Example 2** I **paused** to stop talking, because A) I lost my voice. B) I **ran out of breath**.

❷ **Explanation** *Paused* is hard to translate in a way that implies that the talking is stopped temporarily. Additionally, "*I ran out of breath*" is challenging to translate in a manner that a fluent speaker would use, and was therefore translated as "*I had to breathe*" for Võro.

❶ **Example 3** The accident was my fault, therefore A) I felt guilty. B) I **pressed charges**.

❷ **Explanation** It is hard to translate *I pressed charges* in a way that is natural in Võro, so it was translated as *I went to court*, which loses the detail of being the party to press charges and could affect the correctness of this example.

6.3 Semantic non-equivalence and lexical gaps

Some examples were difficult to translate because the target languages do not always cover the same concepts as English. In Võro, this mainly appeared as semantic non-equivalence, where the closest translation has a slightly different or broader meaning than the original English word, potentially increasing ambiguity. In Livonian, lexical gaps were

²huggingface.co/datasets/tartuNLP/smugri4-data

³translated from Estonian XCOPA using <https://neurotolge.ee/> API on January 31, 2025

more common, especially for modern concepts and international vocabulary without well-established equivalents. In such cases, translations relied on paraphrasing and lexical innovation. The latter has been addressed by borrowing from related or close contact languages such as Estonian and Latvian (e.g., the word *checkers*) or adapting international stems to Livonian orthography and morphology (e.g. *team*, *marathon*, *grill*, *hit song*), in line with established practices in Livonian. Some examples also became more ambiguous due to homonymy and polysemy in Livonian. For example, the word *īel* can mean both *street* and *voice* in the genitive case.

❶ **Example 4** I yanked the imposter’s hair, therefore A) her wig came off. B) she went **bald**.

🔍 **Explanation** In Võro, *bald* was translated as *with a bare head*, which could also mean without a head covering (e.g., a hat or a wig). This could also make the correct choice more ambiguous in Võro, thereby affecting the example’s correctness.

❶ **Example 5** The man’s email inbox was full of **spam**, therefore A) he deleted the spam. B) he sent out a mass email.

🔍 **Explanation** There is no documented equivalent for *spam* in Livonian, so a new compound was formed from existing words meaning *junk* and *mail*.

❶ **Example 6** The item was packaged in **bubble wrap**, because A) it was fragile. B) it was small.

🔍 **Explanation** There is no word for bubble wrap in Livonian, so it was translated as bubble plastic (*bulpplast*).

6.4 Other issues

One of the more challenging issues in translating Võro sentences was choosing the correct demonstrative pronoun from the ternary system (*seo–taa–tuu*) to represent English binary *this–that*. Their interpretation depends not only on spatial and temporal distance but also on additional factors, but the use of these pronouns in written Võro has not been systematically studied. Moreover, the entire demonstrative pronoun system appears to be undergoing change in the context of Estonian–Võro bilingualism (Plado and Reile, 2020;

Pajusalu, 2015).

6.5 Overall correctness

Only a small number of translations were identified as potentially affecting benchmark correctness. Based on a manual review of the translators’ comments, we identified potentially correctness-affecting shifts in fewer than 1% of the Võro (3/500) and Livonian (4/500) examples. Because this estimate is based on translator comments rather than an independent audit, additional ambiguous cases may not have been identified.

7 Discussion

The parallel nature of COPA translations, such as SMUGRI-COPA, is both a blessing and a curse. On the one hand, it enables direct comparisons across languages and language varieties, allowing the study of cross-lingual differences within a shared evaluation framework. On the other hand, this makes the benchmark more general and less sensitive to culture-specific knowledge and language-specific nuances. There remains a need for evaluation benchmarks grounded in local cultural contexts and natural language use. Developing such culturally and linguistically specific resources remains an important direction for future work.

8 Conclusion

In this work, we introduced SMUGRI-COPA, a new commonsense reasoning benchmark for the extremely low-resource Finnic languages Võro and Livonian. By extending XCOPA to these languages, we provide the largest discriminative benchmark for language model evaluation currently available for them and enable direct comparison with existing XCOPA translations across languages with different resource levels.

Our results reveal a substantial gap between higher-resource Estonian and the lower-resource languages, especially Livonian, where most models fail to significantly outperform random guessing. In contrast, reasoning-enabled models can substantially improve accuracy scores in extremely low-resource settings. Nevertheless, our reasoning-summary analysis shows that these gains may rely on translation and surface-form cues and should not be taken as conclusive evidence of target-language understanding. We additionally find that machine-translated benchmarks may inflate evaluation scores, raising concerns about their reliability

in low-resource settings.

Overall, our findings highlight both the importance and the difficulty of evaluating language understanding in endangered languages. We hope SMUGRI-COPA will support future work on multilingual evaluation and language technologies for low-resource and endangered languages.

Limitations

There are multiple language variants that people use in practice (dialects, subdialects, orthographies). We focus on a single version translated by a single person for each language, which in practice could affect results compared to other varieties of the same language. There might be disagreement in the community about the preferred versions of the languages, and we do not contend that we present the only correct version. Beyond checks performed by the translators themselves, we have not conducted independent human validation of translation or task correctness. Furthermore, providing translators with the correct labels helped preserve the intended causal relationships, but may also have influenced wording choices or introduced unintended surface cues. Label-blind validation would therefore be valuable future work.

Since COPA and XCOPA have been publicly available for years, they may have been included in language model training corpora. However, our translations are completely new and were not publicly available before our evaluation. Although this reduces the likelihood of direct string-level contamination, it does not eliminate task-level contamination: a model may translate an item into English or Estonian and recover an answer previously encountered through COPA or XCOPA. Nevertheless, asking the same question in another language can provide useful evidence of the model’s ability to understand the target language and connect information across languages. Consequently, high accuracy provides evidence of target-language understanding, but cannot by itself establish that the model reasoned independently in the target language when its training data is unknown.

The reasoning-summary analysis cannot fully resolve this uncertainty, since OpenAI’s reasoning models provide summaries rather than complete internal reasoning traces. Our analysis is therefore limited to these model-generated summaries and cannot establish what the models’ full internal reasoning contained.

Finally, COPA evaluates only one form of causal commonsense reasoning. Results on SMUGRI-COPA may therefore not generalize to broader language understanding or generation tasks.

Ethics Statement

To avoid implying a single authoritative standard for the languages, we explicitly present the dataset as one possible translation among multiple possible variants. The dataset should therefore not be interpreted as a definitive representation of the languages, but rather as one resource intended to support the evaluation and development of language models.

SMUGRI-COPA will be released under *CC BY 4.0*. To reduce accidental benchmark contamination, we use a gated repository that will automatically grant access after displaying a non-binding request to not use the dataset for training or redistribute it in a way accessible to web-crawlers. This request does not modify the license. We additionally include a canary GUID with each example to support future detection of dataset leakage into training corpora.

During dataset creation, translators were compensated for their work and additionally offered the opportunity to participate in the writing process as co-authors.

Acknowledgments

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

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie 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.
- Giuseppe Attardi. 2015. Wikiextractor. <https://github.com/attardi/wikiextractor>.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman

- Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. 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.
- Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. [VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. [Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.
- Aleksei Dorkin, Taido Purason, Emil Kalbaliyev, Hele-Andra Kuulmets, Marii Ojastu, Mark Fišel, Tanel Alumäe, Eleri Aedmaa, Krister Kruusmaa, and Kairit Sirts. 2026. [Estllm: Enhancing estonian capabilities in multilingual llms via continued pretraining and post-training](#). *Preprint*, arXiv:2603.02041.
- Ina Druviete and Gunta Kļava. 2018. [The role of Livonian in Latvia from a sociolinguistic perspective](#). *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 9(2):129–146.
- Valts Ernštreits, Mark Fišel, Matīss Rikters, Marili Tomingas, and Tuuli Tuisk. 2022. [Language resources and tools for Livonian](#). *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(1):13–36.
- 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, and 5 others. 2024. [The language model evaluation harness](#).
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Bauccells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. [Building a data infrastructure for a mid-resource language: The case of Catalan](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia. ELRA and ICCL.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Hadad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. [OLMES: A standard for language model evaluations](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5020–5048, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Āurech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 83 others. 2026. [Apertus: Democratizing open and compliant LLMs for global language environments](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Hele-Andra Kuulmets, Taido Purason, and Mark Fishel. 2025. [How well do LLMs know Finno-Ugric languages? A systematic assessment](#). In *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, Tallinn, Estonia. University of Tartu Library.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. [DIALECT-COPA: Extending the standard translations of the COPA causal common-sense reasoning dataset to South Slavic dialects](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98, Mexico City, Mexico. Association for Computational Linguistics.
- Nikola Ljubešić and Davor Lauc. 2021. [BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian](#). In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.
- Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. [Fineweb-edu: the finest collection of educational content](#).
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. [Eurollm-9b: Technical report](#). *Preprint*, arXiv:2506.04079.
- Christopher Moseley, editor. 2010. [Atlas of the World's Languages in Danger](#), 3rd edition. UNESCO, Paris.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Renate Pajusalu. 2015. [Võro demonstratives: changing or disappearing? Eesti ja soome-ugri keeleteaduse ajakiri](#). *Journal of Estonian and Finno-Ugric Linguistics*, 6(2):167–190.
- 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.
- Dmytro Pashchenko, Lisa Yankovskaya, and Mark Fishel. 2025. [Paragraph-level machine translation for low-resource Finno-Ugric languages](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 458–469, Tallinn, Estonia. University of Tartu Library.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language](#). In *Second Conference on Language Modeling*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Helen Plado, Liina Lindström, and Sulev Iva. 2023. [Võro South Estonian](#). In Daniel Abondolo and Riitta-Liisa Valijärvi, editors, *The Uralic Languages*, pages 386–432. Routledge.
- Helen Plado and Maria Reile. 2020. [Vähemuskeele uurimine keeleteadusliku katsega. demonstratiivpronoomenite kasutus ruumis viitamisel tänapäeva võro keeles](#). *Emakeele Seltsi Aastaraamat*, 65:142–163.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense 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. 2025. [LLMs for extremely low-resource Finno-Ugric languages](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6692–6712, Albuquerque, New Mexico. Association for Computational Linguistics.
- Liisa Rätsep and Mark Fishel. 2023. [Neural text-to-speech synthesis for Võro](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727, Tórshavn, Faroe Islands. University of Tartu Library.
- Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fishel. 2022. [Machine translation for Livonian: Catering to 20 speakers](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 508–514, Dublin, Ireland. 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 *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.
- David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja Øvrelid, Lucas Georges Gabriel Charpentier, Andrey Kutuzov, and Stephan Oepen. 2025. [Small languages, big models: A study of continual training on languages of Norway](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 573–608, Tallinn, Estonia. University of Tartu Library.
- Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. [RussianSuperGLUE: A Russian language understanding evaluation benchmark](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.
- Statistics Estonia. 2022. RL214462: Estonian speaking population by ability to speak a dialect, age group, sex, educational attainment, and place of residence (settlement region), 31 December 2021. <https://andmed.stat.ee/en/stat/RL214462>. Population and Housing Census 2021, accessed 2026-05-11.
- Maali Tars, Taido Purason, and Andre Tättar. 2022. [Teaching unseen low-resource languages to large translation models](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 375–380, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Maali Tars, Andre Tättar, and Mark Fišel. 2021. [Extremely low-resource machine translation for closely related languages](#). In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasajo, and Alham Aji. 2024. [COPAL-ID: Indonesian language reasoning with local culture and nuances](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.
- Wikimedia Foundation. 2025. [Wikimedia downloads](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. [Machine translation for low-resource Finno-Ugric languages](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.
- Aleš Žagar and Marko Robnik-Šikonja. 2022. [Slovene SuperGLUE benchmark: Translation and evaluation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France. European Language Resources Association.

A Translation Details

A.1 Translator guidelines

The guidelines given to translators are in Figure 4. Additionally, the translators were instructed to prioritize naturalness in the target language and task correctness, and were permitted to make adjustments when sentences were not directly translatable. For example, when translating "bubble wrap", since there is no one-to-one translation, "bülplast" ("bubble plastic") was used.

Each example in the XCOPA test dataset consists of three short sentences: a premise and two choices. The tasks fall into two types: cause and effect. For cause-type tasks, you must choose which of the two options explains the cause of the situation described in the premise. For effect-type tasks, you must choose which option describes the effect of the situation described in the premise. In both cases, you should choose the most plausible/likely of the two options, given the premise.

The goal is to translate the test dataset into {Võro/Livonian} while preserving the correctness of the tasks. To check this, we have included the correct answer number with each task. The dataset was originally in English but has been translated into Estonian. In case the Estonian translation seems unclear, the original English version is also included. We ask you to translate two subsets: a test set of 500 examples and a validation set of 100 examples. If it seems that the originally marked correct answer is wrong, please let us know by tagging the sentence.

Figure 4: Translator guidelines translated into English (original in Estonian).

A.2 Compensation

The translators were compensated €800 gross (before taxes) for translating the dataset. We estimated that translating a single COPA example required

approximately 7 minutes on average. This corresponds to an estimated hourly wage of approximately €11.43. At the time of the work, the mean gross hourly wage in the translators’ country was €11.97/hour, and the median was €9.82/hour (all before taxes). The translators were additionally offered the opportunity to contribute to the paper writing as co-authors.

B Evaluation Details

B.1 Instruction-following model prompt

The instruction-following model prompt is listed in Figure 5. The output to this prompt was parsed with regular expressions created iteratively to cover different model outputs.

```
SYSTEM:
You will receive a premise and two options describing
either the cause or the effect. Choose the most plausible
option and reply only with its letter (A or B).

USER:
{premise}, {connector}\n
A. {choice_1}\n
B. {choice_2}
```

Figure 5: Instruction-following model prompt. The connector for COPA could be *because* or *therefore* depending on the task type (in the target language).

B.2 Technical details

The models were evaluated on 8 GPU compute units (4 AMD MI250x) on the LUMI supercomputer.

B.3 Full list of benchmarked models

The full list of models, their versions, and licenses is provided in Table 12. All evaluated models were used in accordance with their respective licenses and applicable terms of use.

B.4 Manual analysis of reasoning summaries

We randomly selected 50 COPA examples and manually analyzed the GPT-5 reasoning summaries for the corresponding Võro and Livonian versions. As summaries were not returned for every response, the analysis covered 23 Võro and 26 Livonian summaries. We labeled a summary as using *translation reasoning* when it explicitly reconstructed the target-language input in another language. Translation correctness was rated on a five-point scale by comparison with the parallel English source;

ratings of 1 (*fully incorrect*) or 2 (*mostly incorrect*) were categorized as *low understanding*. We labeled *surface-form reliance* when the reasoning used lexical overlap, cognates, shared tokens, or related-language forms as evidence. These categories were not mutually exclusive.

C Dataset Statistics

We list SMUGRI-COPA dataset statistics in Table 8.

Lang.	bytes	words	characters	GPT-5 tokens
TEST SET ($n = 500$)				
ENG	43096	8043	43096	9954
EST	41658	5992	40122	14107
VRO	42887	6268	38813	16539
LIV	46650	6314	39261	21738
VALIDATION SET ($n = 100$)				
ENG	9131	1689	9131	2082
EST	8708	1245	8383	2925
VRO	8811	1290	7995	3400
LIV	9820	1319	8282	4523

Table 8: SMUGRI-COPA test and validation set statistics with Estonian XCOPA and English COPA for comparison.

D Challenges in Translating COPA

We list some additional examples here to give the reader a more comprehensive overview of actual challenges encountered when translating to a low-resource language.

D.1 Cultural context

❶ **Example 7** I emptied my pockets, therefore: A) I retrieved a ticket stub. B) **I found a weapon.**

❷ **Explanation** Finding a weapon in a pocket would be an uncommon scenario for Võro and Livonian. Fortunately, this does not change the correctness of the example.

D.2 Semantic non-equivalence and lexical gaps

❶ **Example 8** The girl ran out of energy, because A) she played **checkers**. B) she jumped rope.

❷ **Explanation** Some game names are not documented in Livonian, so borrowing from Latvian together with Livonian morphology was used as a translation strategy. This follows existing borrowing patterns already used in Livonian

game terminology.

❶ **Example 9** I felt bored, therefore A) I **snorted**. B) I yawned.

🔍 **Explanation** There is no direct translation for *snorted* in Livonian, so it was translated onomatopoeically.

E Continued Pre-training Details

E.1 Setup

The Võro and Livonian training data was collected from SMUGRI4⁴ (CC BY-NC-SA 4.0), Wikipedia (Wikimedia Foundation, 2025) using WikiExtractor (Attardi, 2015), and *liv4ever* (Rikters et al., 2022, CC BY-NC-SA 4.0). The datasets were mixed proportionally using only 1 epoch for training. 50% of the training budget was taken from Estonian FineWeb-2 (Penedo et al., 2025, ODC-BY) and FineWeb-Edu (Lozhkov et al., 2024, ODC-BY) in equal proportion. We list the training dataset language composition in Table 9.

The continued pre-training of Llama-3.1-EstLLM-8B-0525 (Dorkin et al., 2026, Llama-3.1 license) was conducted using Hugging Face transformers (Wolf et al., 2020). The training examples were packed into sequences of 2048 tokens. The learning rate was selected from {2e-5, 4e-5, 8e-5} to minimize the mean of Estonian, Võro, and Livonian validation loss on held-out pre-training data. A single training run took approximately 51 GPU-hours. The training was run on the LUMI supercomputer across 4 nodes (32 GPU compute units, AMD MI250x) with the hyperparameters listed in Table 10.

Our model is trained solely for research purposes, in line with the training dataset and model licenses.

	LIV	VRO	EST	ENG
Tokens	1.3M	9.8M	5.5M	5.5M

Table 9: Continued pre-training dataset language composition in Llama-3 tokens.

E.2 Training results

We examine how SMUGRI-COPA accuracy improves as continued pre-training progresses (see Figure 6). We see steady improvements for MCF prompting, while CF prompting scores saturate

⁴<https://huggingface.co/datasets/tartuNLP/smugri4-data>

Hyperparameter	Value
Learning rate	4e-5
Optimizer	AdamW
Adam ϵ	1e-8
Adam β_1, β_2	0.9, 0.95
Weight decay	0.1
Sequence length	2048
Scheduler	cosine decay to 10%
Warmup ratio	0.04
FSDP Strategy	SHARD_GRAD_OP
Number of GPUs	32
Precision	bfloat16
Batch size (total)	32
Batch size (total, tokens)	65,536
Training steps	337
Training tokens	22,085,632

Table 10: Hyperparameters for the continued pre-training.

after just 84 steps for Võro (roughly 2.4M Võro tokens).

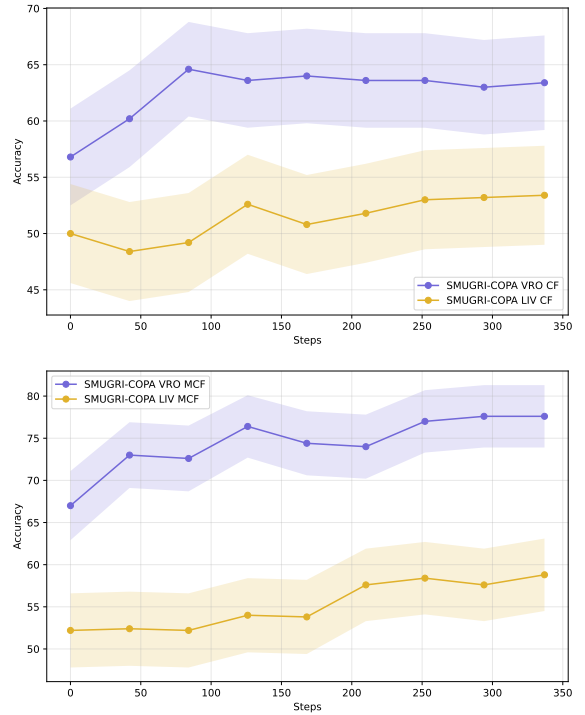


Figure 6: SMUGRI-COPA 5-shot test set accuracy at various continued pre-training steps for cloze format (CF) and multiple-choice format (MCF).

F Additional Benchmark Scores

We list the 5-shot evaluation scores for all benchmarks covering Võro and Livonian in Table 11. We used lm-evaluation-harness configuration files from Kuulmets et al. (2025). We report Belebele-SMUGRI (Purason et al., 2025), SIB-SMUGRI (Purason et al., 2025) and SMUGRI-FLORES

Model	SMUGRI-COPA (CF)		SMUGRI-COPA (MCF)		Belebele		SIB-200	
	VRO	LIV	VRO	LIV	VRO	LIV	VRO	LIV
Llama-SMUGRI-7B	60.2 ± 4.3	52.8 ± 4.4	54.0 ± 4.4	48.4 ± 4.4	36.2 ± 8.4	28.3 ± 7.9	78.4 ± 7.2	64.8 ± 8.4
Llama-3.1-EstLLM-8B-0525	56.8 ± 4.3	50.0 ± 4.4	67.0 ± 4.1	52.2 ± 4.4	56.7 ± 8.7	43.3 ± 8.7	83.2 ± 6.6	76.8 ± 7.4
Apertus-8B-2509	54.6 ± 4.4	45.8 ± 4.4	61.4 ± 4.3	52.4 ± 4.4	53.5 ± 8.7	37.8 ± 8.5	76.0 ± 7.5	71.2 ± 8.0
Llama-3.1-8B	51.0 ± 4.4	47.6 ± 4.4	56.4 ± 4.4	51.0 ± 4.4	51.2 ± 8.7	37.8 ± 8.5	86.4 ± 6.0	72.8 ± 7.8
EuroLLM-9B	54.6 ± 4.4	47.0 ± 4.4	63.8 ± 4.2	51.4 ± 4.4	52.0 ± 8.7	42.5 ± 8.6	83.2 ± 6.6	76.0 ± 7.5
Apertus-70B-2509	57.8 ± 4.3	48.2 ± 4.4	65.8 ± 4.2	45.2 ± 4.4	63.8 ± 8.4	48.0 ± 8.7	84.8 ± 6.3	78.4 ± 7.2
Llama-3.1-70B	57.0 ± 4.3	49.4 ± 4.4	74.2 ± 3.8	52.4 ± 4.4	73.2 ± 7.7	45.7 ± 8.7	87.2 ± 5.9	78.4 ± 7.2
Llama-3.1-EstLLM-8B-VRO-LIV	63.4 ± 4.2	53.4 ± 4.4	77.6 ± 3.7	58.8 ± 4.3	66.1 ± 8.3	46.5 ± 8.7	83.2 ± 6.6	81.6 ± 6.8

FLORES (BLEU)								
	EST-VRO	EST-LIV	ENG-VRO	ENG-LIV	VRO-ENG	LIV-ENG	VRO-EST	LIV-EST
Llama-SMUGRI-7B	28.2 ± 2.1	4.4 ± 0.9	10.5 ± 1.3	2.2 ± 0.5	25.7 ± 2.0	14.4 ± 1.6	48.6 ± 2.4	17.4 ± 1.8
Llama-3.1-EstLLM-8B-0525	15.9 ± 1.4	5.6 ± 1.1	7.5 ± 1.3	2.2 ± 0.5	22.9 ± 2.2	9.0 ± 1.7	42.1 ± 2.3	8.6 ± 1.3
Apertus-8B-2509	13.5 ± 1.4	5.1 ± 1.0	5.2 ± 1.0	1.7 ± 0.5	24.5 ± 1.9	9.2 ± 1.8	37.9 ± 2.7	8.8 ± 1.3
Llama-3.1-8B	13.6 ± 1.5	4.7 ± 1.0	3.1 ± 0.7	1.9 ± 0.4	16.4 ± 2.4	10.3 ± 1.6	26.1 ± 1.9	6.6 ± 1.2
EuroLLM-9B	12.3 ± 1.6	2.2 ± 0.5	5.0 ± 1.0	1.3 ± 0.4	26.9 ± 2.3	11.4 ± 1.6	40.2 ± 2.6	10.6 ± 1.8
Apertus-70B-2509	21.2 ± 1.8	5.7 ± 1.4	8.6 ± 1.4	3.5 ± 1.0	28.2 ± 2.1	12.3 ± 2.3	45.9 ± 3.0	11.4 ± 1.9
Llama-3.1-70B	20.9 ± 2.0	6.9 ± 1.4	9.3 ± 1.4	4.7 ± 1.0	30.4 ± 2.5	16.1 ± 2.7	42.9 ± 2.5	13.2 ± 2.0
Llama-3.1-EstLLM-8B-VRO-LIV	34.0 ± 2.5	7.3 ± 1.1	14.2 ± 1.4	4.3 ± 0.8	29.2 ± 2.4	17.8 ± 1.9	51.2 ± 2.7	19.4 ± 2.0

Table 11: Scores (accuracy or BLEU (Papineni et al., 2002)) for 5-shot evaluation with benchmarks supporting Võro or Livonian.

Model		Version	License
GPT-4o	-	gpt-4o-2024-05-13	Proprietary
GPT-4.1	-	gpt-4.1-2025-04-14	Proprietary
GPT-5	-	gpt-5-2025-08-07	Proprietary
Gemini-3-flash	-	gemini-3-flash-preview	Proprietary
DeepSeek-V3.2	(DeepSeek-AI et al., 2025)	deepseek-ai/DeepSeek-V3.2	MIT
Apertus-70B-Instruct-2509	(Hernández-Cano et al., 2026)	swiss-ai/Apertus-70B-Instruct-2509	Apache-2.0
Gemma-4-31B-it	(Gemma Team et al., 2026)	google/gemma-4-31B-it	Apache-2.0
Llama-SMUGRI-7B	(Purason et al., 2025)	tartuNLP/Llama-SMUGRI-7B	-
Apertus-8B-2509	(Hernández-Cano et al., 2026)	swiss-ai/Apertus-8B-2509	Apache-2.0
Llama-3.1-EstLLM-8B-0525	(Dorkin et al., 2026)	tartuNLP/Llama-3.1-EstLLM-8B-0525	Llama-3.1
Llama-3.1-8B	(Grattafiori et al., 2024)	meta-llama/Llama-3.1-8B	Llama-3.1
EuroLLM-9B	(Martins et al., 2025)	utter-project/EuroLLM-9B	Apache-2.0
Llama-3.1-70B	(Grattafiori et al., 2024)	meta-llama/Llama-3.1-70B	Llama-3.1
Apertus-70B-2509	(Hernández-Cano et al., 2026)	swiss-ai/Apertus-70B-2509	Apache-2.0

Table 12: Models benchmarked in results section.

(Yankovskaya et al., 2023), all of them licensed under *Creative Commons Attribution Share Alike 4.0*, which is in line with our usage of them.