

How multilingual are multilingual LLMs? A case study in Northern Sámi-Finnish Translation

Jonne Sälevä and Constantine Lignos

Michtom School of Computer Science

Brandeis University

{jonnesaleva,lignos}@brandeis.edu

Abstract

We use Finnish and Northern Sámi as a case study to investigate how suitable multilingual LLMs are for low-resource machine translation and how much performance can be improved using supervised finetuning with varying amounts of parallel data. Our experiments on zero-shot translation reveal that mainstream multilingual LLMs from a variety of model families are unsuitable for translation between our chosen languages as-is, regardless of the generation hyperparameters. On the other hand, our experiments on supervised finetuning reveal that even relatively small amounts of parallel data can be very useful for improving performance in both translation directions.

1 Introduction

This paper focuses on benchmarking the usability of multilingual large language models (LLMs) for machine translation (MT) between less-studied languages. We evaluate MT between two Uralic languages, Finnish (`fin`) and Northern Sámi (`sme`).

Finnish is primarily spoken in Finland, whereas Northern Sámi is spoken by the Indigenous Sámi people primarily located in Finland, Sweden, and Norway. Despite their relatedness and similar linguistic characteristics such as agglutinative morphology, Finnish and Northern Sámi are vastly different in terms of number of speakers and language technology support.

Finnish is often regarded as a “medium-resourced” language given its speaker base of approximately 6 million and relatively large amount of available digital resources. In contrast, Northern Sámi remains extremely low-resourced with only approximately 25,000 speakers and far fewer digital resources. The lack of resources in Northern Sámi is due to many factors, including its status as a minority language whose speakers often also speak another majority language which is used in online communication.

This paper provides a case study of how well modern multilingual LLMs support extremely low-resourced languages like Northern Sámi and how easily they can be adapted via supervised finetuning to perform better at MT. While we only focus on one language pair, our hope is that the findings of this paper can be generalized to further languages.

This paper makes the following contributions. First, we assess how modern multilingual LLMs perform as zero-shot translators between Finnish and Northern Sámi. Second, we demonstrate how MADLAD-400, a recent multilingual encoder-decoder model, can be adapted using only small amounts of parallel data to perform well at Finnish-Northern Sámi MT. Our results show that, despite low zero-shot scores, MADLAD-400 can be adapted to rival the performance of recent bilingual models for Finnish-to-Northern Sámi MT and to reach a new state-of-the-art in the Northern Sámi-to-Finnish direction. Finally, we also discuss the utility of various data augmentation schemes such as mixing translation directions during finetuning as well as the use of synthetic data. All code is available at <https://www.github.com/j0ma/how-multilingual-is-multilingual>.

2 Related Work

Developing language technology applications for Sámi languages has a long history and ranges from rule-based language documentation with FSTs (e.g. Rueter and Hämäläinen, 2020) to more modern NLP tasks like text classification (Alnajar et al., 2024) and non-text applications like automated speech recognition (e.g. Getman et al., 2024; Hiovain and Suni, 2025; Gamboni, 2025).

Recently there has also been an interest in MT for Northern Sámi, starting with Aulamo et al. (2021) who developed RNN and Transformer-based translation models for Finnish and Northern Sámi and whose work was later expanded by

Corpus	Sentences
<i>Full data sets (Aulamo et al., 2021)</i>	
Parallel data	25,106
Bidirectional parallel data	50,212
Parallel data + synthetic data	487,862
<i>Test and validation sets</i>	
UiT test set	2,000
YLE test set	151

Table 1: Finetuning and evaluation data set sizes (number of sentences) across our experiments.

Sälevä and Lignos (2024). In addition to bilingual models, there has also been interest in massively multilingual MT models for Northern Sámi (e.g. Tars et al., 2022a,b; Yankovskaya et al., 2023).

With the advent of LLMs, there has also been interest in developing multilingual models that support Northern Sámi. A survey of this landscape is provided by Paul et al. (2024). The MADLAD-400 dataset and associated T5-style models introduced by Kudugunta et al. (2023) were among the first openly-available LLMs to do so. A related trend is developing geography-specific models that support Northern Sámi in addition to majority languages of a given country, such as *NorMistral* (Samuel et al., 2025) which supports variants of Norwegian in addition to Northern Sámi, but not Finnish.

3 Datasets and Models

Datasets All of our experiments rely on the training and evaluation data originally introduced by Aulamo et al. (2021) and also used by Sälevä and Lignos (2024). All corpus sizes can be seen in Table 1. For evaluation, we use the general domain UiT test set of 2,000 sentence pairs as well as the smaller YLE test set of 151 sentences which is restricted to the news domain. Our finetuning experiments use the UiT training set which contains 25,106 parallel sentences. Our “bidirectional parallel data” set consists of the UiT training set repeated twice, once with Finnish as the source language and another time with Northern Sámi as the source language. Finally, our “parallel + synthetic data” includes 487,862 sentence pairs, consisting of the original 25,106 parallel sentences from the UiT train set as well as 462,756 pseudo-parallel sentences with human-generated Northern Sámi text and synthetic Finnish text generated using a rule-based MT system. It corresponds to the RBMT-all-bt training set of Aulamo et al. (2021).

Models Our experiments use models from five multilingual language model families: MADLAD-400 (Kudugunta et al., 2023), Gemma 3 by Google (Team et al., 2025), Aya 101 (Üstün et al., 2024) and Aya Expanse (Dang et al., 2024) by Cohere, and Minstral 3 by MistralAI.¹

Out of the five model families, Gemma 3, Aya Expanse, and Minstral 3 are causal, decoder-only language models whereas MADLAD-400 and Aya 101 use an encoder-decoder architecture. The models we use range in size from relatively small to medium to large. MADLAD-400 and Aya 101 are the smallest model families, containing only 3B/10B parameters for MADLAD-400 and 13B for Aya 101. Minstral 3 also belongs to the medium-sized category with 14B parameters. Gemma 3 and Aya Expanse are the largest models we use with 27B and 32B parameters, respectively. In terms of language support, only MADLAD-400 explicitly supports both Finnish and Northern Sámi whereas Aya 101 provides support for Finnish only. The language support of Gemma 3 is unknown but the technical report claims support for 140 languages (Team et al., 2025). Minstral 3 and Aya Expanse are multilingual LLMs but do not explicitly support either Finnish or Northern Sámi.

Experimental setup We experiment with both zero-shot translation as well as post-training using supervised finetuning. We run all zero-shot translation evaluations using the lm-evaluation-harness library by EleutherAI (Gao et al., 2024). For the post-training experiments with MADLAD-400, we rely on HuggingFace Transformers (Wolf et al., 2019) for both training and inference. For the Gemma 3 finetuning experiments we use the Unslloth framework (Han et al., 2023). For evaluation, we use BLEU (Papineni et al., 2002) and chrF2 (Popović, 2015) computed using the sacrebleu library (Post, 2018).

For zero-shot translation, uncertainty around observed BLEU and chrF2 scores is quantified using the bootstrapping mechanism implemented in lm-evaluation-harness and given as the standard deviation in our tables. In experiments that do not use lm-evaluation-harness, we compute the bootstrap standard deviation using a custom script based on sufficient statistics obtained from sacrebleu.

¹<https://huggingface.co/mistralai/Minstral-3-14B-Instruct-2512>

Direction	Model	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	MADLAD-400-3B-MT	5.20 \pm 0.26	26.59 \pm 0.47	5.39 \pm 1.05	26.63 \pm 1.33
	MADLAD-400-10B-MT	9.04 \pm 0.36	38.77 \pm 0.39	5.87 \pm 0.99	30.48 \pm 1.64
	Gemma 3 27B	0.00 \pm 0.00	0.98 \pm 0.01	0.00 \pm 0.00	0.63 \pm 0.03
	Aya 101 13B	2.13 \pm 0.19	17.39 \pm 0.19	0.88 \pm 0.33	15.80 \pm 0.70
	Aya Expanse 32B	0.92 \pm 0.09	14.59 \pm 0.21	0.71 \pm 0.21	15.83 \pm 0.63
	Minstral 3 14B	0.39 \pm 0.06	11.47 \pm 0.16	0.13 \pm 0.03	13.82 \pm 0.63
Sámi \Rightarrow Finnish	MADLAD-400-3B-MT	18.85 \pm 0.58	50.78 \pm 0.50	15.32 \pm 1.45	39.28 \pm 1.50
	MADLAD-400-10B-MT	24.24 \pm 0.59	56.40 \pm 0.51	18.04 \pm 1.55	42.44 \pm 1.73
	Gemma 3 27B	0.00 \pm 0.00	0.73 \pm 0.01	0.00 \pm 0.00	0.53 \pm 0.02
	Aya 101 13B	2.73 \pm 0.20	24.88 \pm 0.29	2.40 \pm 0.73	22.48 \pm 0.75
	Aya Expanse 32B	0.45 \pm 0.04	16.84 \pm 0.16	0.17 \pm 0.05	12.54 \pm 0.50
	Minstral 3 14B	0.62 \pm 0.10	11.77 \pm 0.18	0.19 \pm 0.05	12.66 \pm 0.62

Table 2: Results of zero-shot MT experiments between Finnish and Northern Sámi using greedy decoding. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

4 Results

4.1 Zero-shot Translation

As our first experiment, we investigate the zero-shot translation capability in both translation directions. We focus on two settings: first, generation using greedy decoding (the default setting in `lm-evaluation-harness`) as well as sampling-based generation. For sampling, we use a temperature of $T = 1$ for all models except MADLAD-400 for which we use $T = 0.1$.

The results using greedy decoding and sampling can be seen in Tables 2 and 3, respectively. Regardless of the decoding algorithm, MADLAD variants perform best out of all the models considered. In the greedy decoding condition, MADLAD-10B achieves BLEU scores of 9.04 (UiT) / 5.87 (YLE) on Finnish-to-Northern Sámi and 24.24 (UiT) / 18.04 (YLE) on Northern Sámi-to-Finnish.

Sampling-based decoding yields similar results, and MADLAD-10B achieves BLEU scores of 9.17 (UiT) / 7.13 (YLE) on Finnish-to-Northern Sámi and 24.10 (UiT) / 18.04 (YLE) on Northern Sámi-to-Finnish. The 3B model lags behind the larger 10B model by 2-6 BLEU points in most conditions. The exception to this is the YLE test set on Finnish-to-Northern Sámi where the gap between the two models is less than 1 BLEU point.

The other models generally perform significantly worse than either of the MADLAD variants. Aya 101 places third when using greedy decoding and scores around 2-3 BLEU on all test sets and in all translation directions. Aya Expanse and Minstral achieve very low translation scores across all tasks, which makes sense given their lack of support for either Finnish or Northern Sámi.

Gemma 3 is a notable exception, performing significantly better using sampling compared to greedy decoding. With greedy decoding, it scores between 0 and 1 in terms of BLEU in both translation directions. However, when sampling is used, the BLEU scores are within two standard deviations of MADLAD-3B on Finnish-to-Northern Sámi for both UiT and YLE test sets and for the YLE test set on Northern Sámi-to-Finnish. This led us to experiment with the generation temperature of Gemma 3 in order to interpolate between greedy decoding and temperature-based sampling; the results can be seen in Appendix A.

4.2 Post-training for MADLAD-400

As MADLAD-400 variants were the best performing models across our zero-shot translation experiments, we next focused on post-training the 3B and 10B variants with supervised finetuning (SFT).

Motivated by the past experiments of [Aulamo et al. \(2021\)](#) and [Sälevä and Lignos \(2024\)](#), our central research question was how much data would be needed to achieve good performance. This led us to experiment with three types of SFT data sets, as explained in Section 3. All SFT experiments were trained for 3 epochs, with the exception of the “synthetic” condition which we trained for 1 epoch due to computational constraints. For decoding, we used beam search with a beam width of 4.

The results of this experiment can be seen in Table 4. Without finetuning, both MADLAD variants do relatively poorly. The 3B variant achieves BLEU scores of 6.22 (UiT) / 5.39 (YLE) on the Finnish-to-Northern Sámi direction, and the 10B variant achieves BLEU scores of 10.75 (UiT) / 9.59 (YLE). In the Northern Sámi-to-Finnish direction,

Direction	Model	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	MADLAD-400-3B-MT	5.16 \pm 0.28	26.44 \pm 0.40	4.85 \pm 0.90	26.42 \pm 1.34
	MADLAD-400-10B-MT	9.17 \pm 0.34	38.67 \pm 0.36	7.13 \pm 1.20	31.09 \pm 1.28
	Gemma 3 27B	4.49 \pm 0.25	30.42 \pm 0.25	3.37 \pm 0.78	26.98 \pm 0.83
	Aya 101 13B	1.88 \pm 0.20	17.00 \pm 0.13	0.93 \pm 0.48	16.57 \pm 0.66
	Aya Expanse 32B	0.32 \pm 0.03	10.43 \pm 0.14	0.14 \pm 0.05	9.12 \pm 0.40
	Minstral 3 14B	0.38 \pm 0.05	11.47 \pm 0.15	0.12 \pm 0.03	13.67 \pm 0.70
Sámi \Rightarrow Finnish	MADLAD-400-3B-MT	19.16 \pm 0.57	51.20 \pm 0.51	15.22 \pm 1.62	39.15 \pm 1.71
	MADLAD-400-10B-MT	24.10 \pm 0.64	56.33 \pm 0.56	18.04 \pm 1.80	42.33 \pm 1.89
	Gemma 3 27B	9.96 \pm 0.42	42.71 \pm 0.42	6.02 \pm 1.17	31.91 \pm 1.33
	Aya 101 13B	2.40 \pm 0.19	25.46 \pm 0.21	1.84 \pm 0.62	22.15 \pm 0.79
	Aya Expanse 32B	0.46 \pm 0.04	15.82 \pm 0.15	0.18 \pm 0.05	11.45 \pm 0.57
	Minstral 3 14B	0.62 \pm 0.10	11.47 \pm 0.19	0.27 \pm 0.07	12.79 \pm 0.59

Table 3: Results of zero-shot MT experiments between Finnish and Northern Sámi using sampling-based decoding. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

Model	Direction	Finetune type	UiT (general)		YLE (news)	
			BLEU	chrF	BLEU	chrF
MADLAD-400-3B-MT	Finnish \Rightarrow Sámi	None	6.22 \pm 0.29	29.88 \pm 0.47	5.39 \pm 0.93	28.99 \pm 1.43
		Parallel	23.22 \pm 0.51	58.01 \pm 0.39	12.44 \pm 1.28	40.59 \pm 1.35
		Bidirectional	24.19 \pm 0.53	58.69 \pm 0.41	13.47 \pm 1.29	41.95 \pm 1.28
		Parallel + synthetic	6.20 \pm 0.30	29.70 \pm 0.45	5.33 \pm 0.96	28.97 \pm 1.34
	Sámi \Rightarrow Finnish	None	21.00 \pm 0.57	54.72 \pm 0.45	16.00 \pm 1.53	41.13 \pm 1.43
		Parallel	36.15 \pm 0.64	66.53 \pm 0.47	19.37 \pm 1.72	45.76 \pm 1.49
		Bidirectional	35.64 \pm 0.65	66.25 \pm 0.47	20.69 \pm 1.77	45.89 \pm 1.48
		Parallel + synthetic	21.07 \pm 0.54	54.58 \pm 0.46	16.04 \pm 1.54	41.15 \pm 1.46
MADLAD-400-10B-MT	Finnish \Rightarrow Sámi	None	10.75 \pm 0.38	42.45 \pm 0.39	9.59 \pm 1.11	34.50 \pm 1.43
		Parallel	28.28 \pm 0.54	62.06 \pm 0.44	14.88 \pm 1.26	45.42 \pm 1.25
		Bidirectional	29.37 \pm 0.58	63.02 \pm 0.42	15.32 \pm 1.31	45.29 \pm 1.30
		Parallel + synthetic	22.44 \pm 0.54	57.45 \pm 0.40	12.35 \pm 1.72	41.55 \pm 1.75
	Sámi \Rightarrow Finnish	None	25.68 \pm 0.63	59.31 \pm 0.48	17.64 \pm 1.98	43.32 \pm 1.77
		Parallel	42.00 \pm 0.72	70.79 \pm 0.48	23.08 \pm 1.67	50.54 \pm 1.52
		Bidirectional	41.97 \pm 0.71	70.75 \pm 0.47	23.68 \pm 1.82	51.10 \pm 1.58
		Parallel + synthetic	19.03 \pm 0.53	56.03 \pm 0.44	11.79 \pm 1.15	40.69 \pm 1.32

Table 4: Results for the MADLAD-3B and 10B supervised finetuning experiments. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

both models perform better, with the 3B variant scoring 21.00 (UiT) / 16.00 (YLE) BLEU, and the 10B achieving BLEU scores of 25.68 (UiT) / 17.64 (YLE). The better performance when translating into Finnish is attributable to the larger amount of Finnish training data used in the pretraining.

Even small amounts of finetuning data seem to yield meaningful improvements. When translating into Northern Sámi, MADLAD-3B achieves BLEU scores of 23.22 (UiT) / 12.44 BLEU (YLE), an improvement of over 100% over the no finetuning setting. The 10B model benefits as well, scoring 28.28 (UiT) / 14.88 (YLE) BLEU when translating into Northern Sámi and 42.00 BLEU (UiT) and 23.08 BLEU (YLE) when translating into Finnish, all near 50% improvement. The “bidi-

rectional” SFT data is also somewhat helpful but to a lesser extent; while it does yield small improvements of approximately 0.5-1 BLEU at times, it is practically tied with the regular finetuning setting. Most promisingly, the Northern Sámi-to-Finnish results represent a state-of-the-art improvement over those achieved by Sälevä and Lignos (2024) (42.00 vs. 28.21 BLEU on UiT, 23.68 vs. 11.03 BLEU on YLE). We also note that while the standard deviations of the YLE results are quite large at roughly 1.5 points, our improvements are far larger in magnitude, indicating their significance.

The use of synthetic data seems less helpful. When translating into Northern Sámi, the 10B model seems to benefit from it on both the UiT and YLE translation tasks (22.44 vs. 10.75 BLEU

for UiT, 12.35 vs. 9.59 BLEU for YLE), while the performance of the 3B model remains essentially unchanged (6.20 vs. 6.22 BLEU for UiT, 5.33 vs. 5.39 BLEU for YLE). In the Northern Sámi-to-Finnish direction, the 10B model’s performance degrades substantially when using synthetic data (19.03 vs. 25.68 BLEU for UiT, 11.79 vs. 17.64 BLEU for YLE), while the 3B model remains stable (21.07 vs. 21.00 BLEU for UiT, 16.04 vs. 16.00 BLEU for YLE). This differs from Sälevä and Lignos (2024), who observe that synthetic target-side data seem to aid in generating backtranslation data which ultimately yields better downstream performance. This suggests that previous findings regarding synthetic data on bilingual Transformer models trained from scratch may not extend to finetuning multilingual models, particularly larger ones.

4.3 Post-training for Gemma 3

Despite its poor performance in our zero-shot experiments, the performance improvement Gemma 3 showed with sampling-based decoding (c.f. Tables 2 and 3) motivated us to investigate how effectively further support for the Finnish–Northern Sámi language pair could be added to the model. To experiment with this, we opted for supervised finetuning similar to the earlier MADLAD-400 experiments. We used the 4B and 12B versions of the instruction-tuned Gemma 3 variant² and finetuned it using the 25,106 parallel sentences. We evaluated finetuning at 3 and 10 epochs.

The results of this experiment can be seen in Table 9 in the Appendix. To mimic the earlier experiments, we evaluated each finetuned model using both greedy and sampling-based decoding. Counterintuitively, greedy decoding consistently outperformed sampling which contrasts with our zero-shot findings where sampling was necessary for the model to produce reasonable output. The 12B model trained for 10 epochs achieved the best results, scoring 12.49 (UiT) / 5.93 (YLE) BLEU on Finnish-to-Northern Sámi and 14.66 (UiT) / 6.12 (YLE) BLEU on Northern Sámi-to-Finnish. While these scores represent a substantial improvement over the near-zero baseline, they still fall short of the finetuned MADLAD models. Training for 10 epochs yielded modest gains over 3 epochs, with improvements of approximately 1 BLEU point on the UiT test set in both translation directions. Performance changes on the out-of-distribution YLE

²<https://huggingface.co/google/gemma-3-4b-it> and <https://huggingface.co/google/gemma-3-12b-it>

test set were more modest and essentially within the margin of error.

4.4 TranslateGemma

After this paper was submitted for review, Google released the TranslateGemma family of models (Google Translate Research Team et al., 2026). These models are specifically designed for translation tasks and were exposed to both Finnish and Northern Sámi during training. For completeness, we performed a post-hoc experiment to evaluate the performance of these models using the 4B, 12B and 27B variants of the model family.

The results of our zero-shot evaluation are shown in Table 6 in the Appendix. All models regardless of size and translation direction perform quite poorly, achieving only single-digit BLEU scores regardless of data set. The 27B variant of TranslateGemma achieved the best performance among these models, scoring 4.47 (UiT) / 4.34 (YLE) BLEU on Finnish-to-Northern Sámi and 8.61 (UiT) / 5.55 (YLE) BLEU on Northern Sámi-to-Finnish.

These scores are lower than those achieved by zero-shot Gemma 3 in both translation directions, as well as both the finetuned Gemma 3 and MADLAD models. This suggests that the translation-specific pretraining of TranslateGemma does not generalize well to low-resourced language pairs, even if the model was exposed to those languages at training time. As TranslateGemma was released after the submission of this article, we were unable to explore finetuning it and leave it for future work.

5 Conclusion

We have shown how even small amounts of parallel data can be leveraged to achieve significant performance improvements on machine translation between Finnish and Northern Sámi. Given that Northern Sámi, like many other less-resourced languages, lacks good decoder-only language model support, this finding is important as SFT post-training can be employed with encoder-decoder models as well. As we only experiment with one language pair, a central open question is to what extent our findings generalize to other language pairs. This paper represents a small, focused contribution that we hope others with more computational resources can expand on. Future work should include exploring whether models like TranslateGemma provide a good candidate for further post-training.

Limitations

The main limitation of our paper is that we only experiment with a single language pair. As such, there is the risk that readers may generalize particular results to other languages/domains without caution.

Ethical Considerations

Our experiments involve Northern Sámi, an Indigenous minority language, and as such can be prone to issues of, for example, unethical or otherwise non-participatory data collection (Wiechetek et al., 2024). We believe the broader impact of our work will be improved access of Northern Sámi speakers to large language models and the various benefits such technologies offer. As non-speakers of Northern Sámi, we are unable to ascertain to what extent this is feasible using the current models and data.

We work with machine translation technology, which may have unethical applications and potential negative impact, particularly towards minoritized communities. However, given that we are working with existing models and data sets, we do not believe that the present work significantly exacerbates the risks of such negative impacts.

Acknowledgments

This work was supported by the grant *Improving Relevance and Recovery by Extracting Latent Query Structure* by eBay to Brandeis University.

References

- Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2024. *Leveraging transformer-based models for predicting inflection classes of words in an endangered Sami language*. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 41–48, Helsinki, Finland. Association for Computational Linguistics.
- Mikko Aulamo, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. *Boosting neural machine translation from Finnish to Northern Sámi with rule-based backtranslation*. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 351–356, Reykjavík, Iceland (Online). Linköping University Electronic Press, Sweden.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. *Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier*. *Preprint*, arXiv:2412.04261.
- Enzo Gamboni. 2025. *Fine-tuning Whisper for Kildin Sami, A Low-Resource Endangered Language*. Master’s thesis, University of Eastern Finland.
- 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. *A framework for few-shot language model evaluation*.
- Yaroslav Getman, Tamás Grósz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. *Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi*. In *Interspeech 2024*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 2539–2543, United States. International Society for Computers and Their Applications (ISCA). Publisher Copyright: \textcopyright 2024 International Speech Communication Association. All rights reserved.; Interspeech ; Conference date: 01-09-2024 Through 05-09-2024.
- Google Translate Research Team, Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. *TranslateGemma Technical Report*. *arXiv preprint*.
- Daniel Han, Michael Han, and Unsloth team. 2023. *Unsloth*.
- Katri Hiovain and Antti Suni. 2025. *Does multilingual and multi-speaker modeling improve low-resource TTS? Experiments on Sámi languages*. In *Proceedings of the 13th edition of the Speech Synthesis Workshop*, pages 196–201. ISCA - International Speech Communication Association.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. *Madlad-400: A multilingual and document-level large audited dataset*. *Preprint*, arXiv:2309.04662.
- 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.

- Ronny Paul, Himanshu Buckhash, Shantipriya Parida, and Dilip K. Prasad. 2024. Towards a more inclusive AI: Progress and perspectives in large language model training for the Sámi language. *Preprint*, arXiv:2405.05777.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Jack Rueter and Mika Hämäläinen. 2020. FST morphology for the endangered Skolt Sami language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 250–257, Marseille, France. European Language Resources association.
- Jonne Sälevä and Constantine Lignos. 2024. Language model priors and data augmentation strategies for low-resource machine translation: A case study using Finnish to Northern Sámi. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12949–12956, Bangkok, Thailand. Association for Computational Linguistics.
- David Samuel, Vladislav Mikhailov, Erik Veldal, 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.
- Maali Tars, Taido Purason, and Andre Tättar. 2022a. 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 Tattar, and Mark Fishel. 2022b. Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, 10(3):435–446.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report.
- Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and Sjur Nørstebø Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15922–15931, Torino, Italia. ELRA and ICCL.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pieric Cistac, Tim Rault, Rémi 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. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing.
- 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.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model.

A Additional Tables

A.1 Gemma 3 Temperature Sweep

Table 5 shows the results of varying the generation temperature for Gemma 3.

A.2 Zero-shot evaluation of TranslateGemma

Table 6 shows the zero-shot performance of the TranslateGemma family of models.

A.3 MADLAD-400 Finetuning

Table 7 shows the supervised finetuning results for MADLAD-400-3B-MT. Table 8 shows the supervised finetuning results for MADLAD-400-10B-MT.

A.4 Gemma 3 Finetuning

Table 9 shows the results of our SFT experiments with the Gemma 3 model family.

Direction	Temperature	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	1.0	4.47 ± 0.25	30.43 ± 0.26	2.84 ± 0.73	26.33 ± 0.83
	0.8	3.48 ± 0.21	29.03 ± 0.22	2.77 ± 0.67	26.59 ± 0.88
	0.6	4.08 ± 0.24	29.67 ± 0.24	2.92 ± 0.76	26.69 ± 0.77
	0.4	4.49 ± 0.25	30.20 ± 0.26	3.75 ± 0.77	27.27 ± 0.78
	0.2	4.30 ± 0.26	30.31 ± 0.24	3.51 ± 0.77	26.73 ± 0.88
	0.1	4.34 ± 0.25	30.29 ± 0.24	3.14 ± 0.76	26.31 ± 0.87
Sámi \Rightarrow Finnish	1.0	10.13 ± 0.41	42.84 ± 0.40	5.71 ± 1.12	31.71 ± 1.34
	0.6	9.82 ± 0.40	42.62 ± 0.39	6.45 ± 1.18	32.23 ± 1.20
	0.8	9.47 ± 0.40	42.34 ± 0.37	4.88 ± 0.94	31.48 ± 1.22
	0.4	10.06 ± 0.39	42.59 ± 0.39	6.31 ± 1.13	31.79 ± 1.25
	0.2	9.93 ± 0.40	42.67 ± 0.40	6.05 ± 1.07	32.14 ± 1.23
	0.1	10.23 ± 0.40	42.84 ± 0.40	5.64 ± 0.93	32.10 ± 1.25

Table 5: Temperature sweep results for Gemma 3 in both the Finnish-to-Northern Sámi and Northern Sámi-to-Finnish directions. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

Direction	Model	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	TranslateGemma 4B	2.09 ± 0.22	17.80 ± 0.19	1.38 ± 0.48	18.28 ± 0.80
	TranslateGemma 12B	2.50 ± 0.18	25.30 ± 0.21	2.05 ± 0.56	24.17 ± 0.83
	TranslateGemma 27B	4.47 ± 0.25	31.81 ± 0.25	4.34 ± 0.82	28.44 ± 0.84
Sámi \Rightarrow Finnish	TranslateGemma 4B	3.66 ± 0.26	26.66 ± 0.21	1.90 ± 0.55	22.33 ± 0.75
	TranslateGemma 12B	6.64 ± 0.32	37.34 ± 0.31	5.48 ± 0.81	29.00 ± 1.00
	TranslateGemma 27B	8.61 ± 0.33	41.54 ± 0.33	5.55 ± 0.89	30.88 ± 1.04

Table 6: Zero-shot performance of instruction-tuned TranslateGemma models on UiT and YLE test sets. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

Direction	Finetune type	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	None	6.22 ± 0.29	29.88 ± 0.47	5.39 ± 0.93	28.99 ± 1.43
	Parallel	23.22 ± 0.51	58.01 ± 0.39	12.44 ± 1.28	40.59 ± 1.35
	Bidirectional	24.19 ± 0.53	58.69 ± 0.41	13.47 ± 1.29	41.95 ± 1.28
	Parallel + synthetic	6.20 ± 0.30	29.70 ± 0.45	5.33 ± 0.96	28.97 ± 1.34
Sámi \Rightarrow Finnish	None	21.00 ± 0.57	54.72 ± 0.45	16.00 ± 1.53	41.13 ± 1.43
	Parallel	36.15 ± 0.64	66.53 ± 0.47	19.37 ± 1.72	45.76 ± 1.49
	Bidirectional	35.64 ± 0.65	66.25 ± 0.47	20.69 ± 1.77	45.89 ± 1.48
	Parallel + synthetic	21.07 ± 0.54	54.58 ± 0.46	16.04 ± 1.54	41.15 ± 1.46

Table 7: Supervised finetuning results for MADLAD-400-3B-MT. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

Direction	Finetune type	UiT (general)		YLE (news)	
		BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	None	10.75 ± 0.38	42.45 ± 0.39	9.59 ± 1.11	34.50 ± 1.43
	Parallel	28.28 ± 0.54	62.06 ± 0.44	14.88 ± 1.26	45.42 ± 1.25
	Bidirectional	29.37 ± 0.58	63.02 ± 0.42	15.32 ± 1.31	45.29 ± 1.30
	Parallel + synthetic	22.44 ± 0.54	57.45 ± 0.40	12.35 ± 1.72	41.55 ± 1.75
Sámi \Rightarrow Finnish	None	25.68 ± 0.63	59.31 ± 0.48	17.64 ± 1.98	43.32 ± 1.77
	Parallel	42.00 ± 0.72	70.79 ± 0.48	23.08 ± 1.67	50.54 ± 1.52
	Bidirectional	41.97 ± 0.71	70.75 ± 0.47	23.68 ± 1.82	51.10 ± 1.58
	Parallel + synthetic	19.03 ± 0.53	56.03 ± 0.44	11.79 ± 1.15	40.69 ± 1.32

Table 8: Supervised finetuning results for MADLAD-400-10B-MT. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.

Direction	Model	Epochs	Decoding	UiT (general)		YLE (news)	
				BLEU	chrF	BLEU	chrF
Finnish \Rightarrow Sámi	4B	3	Greedy	11.25 \pm 0.28	51.31 \pm 0.39	4.49 \pm 0.60	34.22 \pm 1.23
	4B	3	Sample	6.88 \pm 0.22	42.53 \pm 0.37	2.74 \pm 0.37	29.60 \pm 0.92
	4B	10	Greedy	11.34 \pm 0.28	51.33 \pm 0.39	5.30 \pm 0.58	35.99 \pm 1.20
	4B	10	Sample	7.76 \pm 0.23	44.31 \pm 0.38	3.58 \pm 0.45	30.84 \pm 0.94
Finnish \Rightarrow Sámi	12B	3	Greedy	11.31 \pm 0.28	51.52 \pm 0.39	4.94 \pm 0.59	36.39 \pm 1.22
	12B	3	Sample	7.02 \pm 0.22	43.36 \pm 0.37	3.18 \pm 0.52	30.59 \pm 1.05
	12B	10	Greedy	12.49 \pm 0.30	53.17 \pm 0.41	5.93 \pm 0.65	36.96 \pm 1.19
	12B	10	Sample	8.58 \pm 0.24	46.34 \pm 0.37	4.02 \pm 0.57	32.03 \pm 1.16
Sámi \Rightarrow Finnish	4B	3	Greedy	12.17 \pm 0.30	53.46 \pm 0.42	5.51 \pm 0.66	35.02 \pm 1.21
	4B	3	Sample	7.80 \pm 0.23	46.10 \pm 0.37	3.46 \pm 0.58	29.52 \pm 1.02
	4B	10	Greedy	13.02 \pm 0.30	54.84 \pm 0.41	5.00 \pm 0.54	35.37 \pm 1.15
	4B	10	Sample	8.99 \pm 0.25	48.32 \pm 0.41	3.43 \pm 0.51	30.58 \pm 1.05
Sámi \Rightarrow Finnish	12B	3	Greedy	13.93 \pm 0.31	55.90 \pm 0.40	6.03 \pm 0.62	37.42 \pm 1.19
	12B	3	Sample	9.56 \pm 0.27	49.31 \pm 0.42	3.59 \pm 0.48	32.46 \pm 1.14
	12B	10	Greedy	14.66 \pm 0.33	56.51 \pm 0.45	6.12 \pm 0.62	37.74 \pm 1.21
	12B	10	Sample	10.86 \pm 0.29	50.89 \pm 0.43	4.18 \pm 0.51	32.79 \pm 0.99

Table 9: Results for the Gemma 3 supervised finetuning experiments. Uncertainty intervals represent 1 standard deviation computed using bootstrap resampling.