TartuNLP at WMT25 LLMs with Limited Resources for Slavic Languages Shared Task

Taido Purason and Mark Fishel
Institute of Computer Science
University of Tartu, Estonia
{taido.purason, mark.fisel}@ut.ee

Abstract

This paper describes the TartuNLP submission to the Upper Sorbian (*hsb*) and Lower Sorbian (*dsb*) tracks of the WMT25 LLMs with Limited Resources for Slavic Languages shared task, which jointly targets machine translation (MT) and question answering (QA). We develop a single multilingual model based on `Qwen2.5-3B-Instruct` by continuing pretraining on Sorbian monolingual and parallel data together with general instruction datasets, combining language acquisition and instruction-following in a single step. The resulting model delivers substantial improvements over the baseline `Qwen2.5-3B-Instruct` model and also achieves the highest ranking for both tasks in the *hsb* and *dsb* shared task tracks.

	DE-HSB		HSB-QA		final
Team	**chrF++**	**points**	**acc**	**points**	**points**
TartuNLP	86.33	4	58.10	4	8
NRC	87.20	4	29.05	1	5
SDKM	75.73	2	55.24	3	5
baseline	13.88	1	42.86	2	3

Table 1: **Upper Sorbian** (hsb) rankings.

	DE-DSB		DSB-QA		final
Team	**chrF++**	**points**	**acc**	**points**	**points**
TartuNLP	78.20	4	57.56	4	8
NRC	78.24	4	32.20	1	5
SDKM	64.34	2	51.71	3	5
baseline	12.21	1	45.85	2	3

Table 2: **Lower Sorbian** (dsb) rankings.

1 Introduction

This paper presents an overview of the TartuNLP systems developed for the WMT25 Limited Resource Slavic Languages shared task (Okabe et al., 2025). This shared task aimed to create a single large language model (LLM) capable of jointly performing both machine translation (MT) and question answering (QA) for less-resourced Slavic languages. The participants of the shared tasks were limited to using the Qwen2.5 model family (Qwen Team, 2024) with a size constraint of 3B parameters. Our team participated in the Upper Sorbian (*hsb*) and Lower Sorbian (*dsb*) tracks, both endangered languages spoken by only about 20,000–30,000 people in total (Moseley, 2007). The taxonomy of Joshi et al. (2020) categorizes both languages as category-1, *the scraping-bys*. We focused on building a single model that supports both tasks and languages simultaneously. This joint objective introduces a specific challenge: while a moderate amount of MT data exists for Sorbian, there is no QA training data, requiring balancing performance across both tasks.

Although one or both of the Sorbian languages have been included in recent massively multilingual models (Imani et al., 2023; Lin et al., 2024; Ji et al., 2025b,a), to our knowledge, no prior work has developed dedicated LLMs for Sorbian.

We build on recent work adapting LLMs to extremely low-resource languages through continued pretraining and instruction tuning (Purason et al., 2025; Etxaniz et al., 2024; Sainz et al., 2025). Our approach follows Sainz et al. (2025), who demonstrated for the Basque language that combining language acquisition and instruction tuning in a single step and starting from an instruction-tuned model is beneficial. We also adopt this joint learning of instruction-following and language for an instruction-tuned base model in our system.

We continually pretrain `Qwen2.5-3B-Instruct` (Qwen Team, 2024) on a mix of monolingual documents and sentences in *hsb* and *dsb*, general instruction-following data (primarily in English), and Sorbian MT instructions. We supplement the data provided by the organizers with document-level texts from Fineweb-2 (Penedo et al., 2025) and Wikipedia articles (Wikimedia

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 1143–1150
November 8–9, 2025.

Foundation). Our final model outperforms the baseline `Qwen2.5-3B-Instruct` and also achieves the highest rank in the shared task (see Tables 1 and 2). **We publish the final model on HuggingFace**[1].

2 Datasets

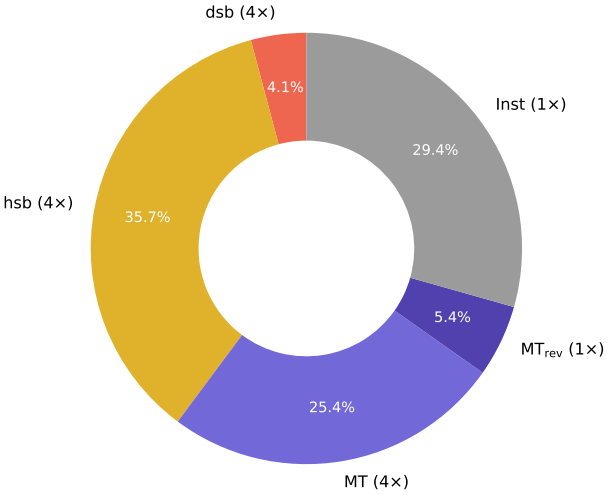

Figure 1: **Data mix** for the final model (% of total training tokens). The number of epochs for each dataset is stated in parentheses. The resulting training set is 1.198B tokens.

Dataset	Texts	Words	Chars
HSB			
Sentence-level (WMT)	1.8M	25.2M	166.2M
Document-level	53.4K	13.1M	90.9M
- Fineweb-2	40.2K	12.1M	83.6M
- Wikipedia	13.2K	1.1M	7.3M
Total (dedup)	1.7M	35.9M	240.6M
DSB			
Sentence-level (WMT)	170.5K	2.5M	15.3M
Document-level	9.5K	2.0M	13.9M
- Fineweb-2	6.3K	1.8M	12.2M
- Wikipedia	3.3K	249.5K	1.7M
Total (dedup)	169.8K	4.4M	28.4M

Table 3: Monolingual dataset statistics.

	de-hsb	de-dsb	dsb-hsb
Sentence pairs	636.3K	212.2K	62.6K

Table 4: The parallel data sentence pair counts.

Monolingual data (see Table 3). We used all of the monolingual data provided by the organizers for this year (Okabe et al., 2025) and the data from previous WMT Sorbian shared tasks (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022), which was sentence-level aligned. Additionally, we used Upper and Lower Sorbian Fineweb-2 (Penedo et al., 2025) documents and Upper and Lower Sorbian Wikipedia documents from the 2025 05 20 dump (Wikimedia Foundation) extracted with WikiExtractor (Attardi, 2015). We also experimented with using German Fineweb-2 (Penedo et al., 2025) documents for pretraining, however, we did not use this data for training the final submission model.

Parallel data (see Table 4). Our parallel data is entirely from this and previous years' shared tasks (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022). The sentence pairs were formatted as instructions in a chat template. Since the task was to translate from German to the Sorbian languages, we trained on translation into the Sorbian languages (de-hsb, de-dsb, hsb-dsb, dsb-hsb) for 4 epochs. We refer to this dataset as MT. In addition to that, we also added one epoch of data in the hsb-de and dsb-de directions as well (referred to as MT$_{rev}$).

The monolingual and parallel data was deduplicated with normalization from Stopes (Pierre Andrews, 2022; NLLB Team et al., 2022). We also removed the held-out and validation data using the same normalization method.

Instructions (see Table 11). We used instruction data from Magpie (Xu et al., 2025), Aya (Singh et al., 2024), EuroBlocks (Martins et al., 2025) OpenAssistant2 (Köpf et al., 2023), and FLAN v2 (Longpre et al., 2023). We removed instructions that were not in English, German, or the closest Slavic languages to the Sorbian languages. Still, 98.6% of instructions in the resulting dataset are in English.

The final mix. The data mix of the submitted model consisted of 1.198B tokens and is displayed in Figure 1.

3 Methodology

We used `Qwen2.5-3B-Instruct` (Qwen Team, 2024) as our base model, motivated by the findings of Sainz et al. (2025), who demonstrated that continued pretraining on already instruction-tuned models is effective for low-resource languages. Following their findings, we combined the language acquisition and instruction-tuning in the same continued pretraining step. Our continued pretraining

[1] huggingface.co/tartuNLP/Qwen2.5-3B-Instruct-hsb-dsb

Model							MT (BLEU / chrF++) de-dsb	MT (BLEU / chrF++) de-hsb	QA (acc) dsb	QA (acc) hsb	Evaluation setting
BASELINES											
`emma-500-llama3.1-8b-bi`							10.8 ± 0.7 / 35.3 ± 0.8	20.0 ± 1.1 / 47.2 ± 0.7	48.0 ± 7.4	41.2 ± 7.2	MT: 5-shot; QA: 3-shot
`Qwen2.5-3B-Instruct`							0.8 ± 0.2 / 12.6 ± 0.3	1.3 ± 0.2 / 17.4 ± 0.3	55.3 ± 7.8	58.0 ± 7.4	MT: 5-shot; QA: 3-shot
CONTINUED PRETRAINING											
	dsb	**hsb**	**MT**💬	**MT$_{rev}$**💬	**Inst**💬	**deu**					
(1)	4x	4x					19.9 ± 0.6 / 45.3 ± 0.5	28.4 ± 1.2 / 54.2 ± 0.6	69.0 ± 6.9	67.9 ± 6.9	MT: 5-shot; QA: 3-shot
(2)	4x	4x	4x				58.6 ± 1.0 / 77.1 ± 0.7	66.6 ± 0.7 / 82.1 ± 0.4	66.1 ± 7.1	70.9 ± 6.8	MT: 0-shot💬; QA: 3-shot
(3)	4x	4x	4x		1x		60.2 ± 1.2 / 78.1 ± 0.7	67.2 ± 0.7 / 82.6 ± 0.4	67.5 ± 7.2	**73.1 ± 6.6**	0-shot💬
(4)	4x	4x	4x	1x	1x		62.0 ± 1.0 / 79.2 ± 0.6	68.3 ± 0.7 / 83.2 ± 0.4	67.5 ± 7.2	70.4 ± 6.7	0-shot💬
(5)	4x	4x	4x	1x	1x	25%	62.1 ± 1.1 / 79.2 ± 0.6	68.6 ± 0.7 / 83.4 ± 0.4	65.7 ± 7.1	65.1 ± 6.9	0-shot💬
(4)	**Final submission**					validation	**62.8 ± 1.1 / 79.8 ± 0.6**	**69.1 ± 0.7 / 83.7 ± 0.4**	**69.3 ± 7.0**	69.3 ± 6.9	MT: 0-shot💬†; QA: 1-shot💬
						test	- / 78.20	- / 86.33	57.56	58.10	

Table 5: Validation set scores for the baselines and fine-tuned models. The shared task's final submission validation and test set scores. 💬 - chat instruction format; † - beam search decoding with beam size 4.

was performed jointly on monolingual documents and sentences, parallel MT data, and instruction-formatted data.

Our initial experiments indicated that the model begins to overfit on Sorbian data around the fourth epoch. A similar finding was also made in Purason et al. (2025), who observed overfitting at 4 epochs for low-resource languages of similar size. With this in mind, we repeated the Sorbian monolingual and parallel data four times during training. The instruction data was repeated once, while the significantly more abundant German monolingual data (not used for the final submission) was limited to 25% of the total training token budget. It should be noted that we did not conduct a thorough investigation into the different data sampling strategies and combinations. A different number of epochs or curriculum learning might provide better results.

For instruction-formatted samples, we applied loss only on the assistant (target) tokens. All datasets were packed into sequences of 4096 tokens, with any overflow tokens carried into the next training sequence. The hyperparameters for the model training are listed in Table 10. The models were trained on either 8 or 16 nodes, each consisting of 4 AMD MI250x GPUs (acting as 8 units) on the LUMI supercomputer. The training of the final model took 139 GPU-hours.

4 Evaluation

We evaluate our models using the `lm-eval-harness` framework (Gao et al., 2024), which we also use to generate the final shared task submissions. Evaluation is conducted using a few-shot and zero-shot prompting strategies, depending on the training strategy. We use a few-shot evaluation without a chat template and a zero-shot evaluation using a chat-style format.

For QA, we follow a multiple-choice setup, selecting the answer option with the highest log-probability from a predefined set of candidates. We calculate the accuracies and report the average across the language levels in the validation set.

For the MT evaluation, we calculated the BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) scores. We also report 95% confidence intervals calculated from standard errors reported by `lm-eval-harness`.

For the final submission, we apply beam search with a beam size of 4 for MT, and use one-shot prompting with the chat template for QA, where each example is treated as a separate turn in a multi-turn conversation. We use greedy decoding in all other evaluation settings (beam size of 1).

5 Results

5.1 Main results

Table 5 summarizes the performance of our models on both machine translation (MT) and question answering (QA) for Upper and Lower Sorbian. We compare our systems against two open-weight baselines: `emma-500-llama3.1-8b-bi` (Ji et al., 2025a), which includes Sorbian in its training data, and `Qwen2.5-3B-Instruct` (Qwen Team, 2024).

As expected, we observe that the `Qwen2.5-3B-Instruct` model performs poorly on Sorbian MT benchmarks before any continued pretraining. In contrast, `emma-500`, which was trained with Sorbian data, performs noticeably better in MT. However, the trend reverses for QA:

`Qwen2.5-3B-Instruct` significantly outperforms `emma-500`, highlighting the strength of instruction tuning for QA even in low-resource settings.

Our continued pretraining configurations substantially outperform the baselines across both MT and QA tasks, demonstrating the effectiveness of the language adaptation. From these results, we find that:

- Including MT data during continued pretraining yields large gains in translation quality compared to relying solely on few-shot prompting.
- Adding instruction-following data provides additional improvements for both tasks.
- Adding a small amount of reverse-direction (into German) MT data (MT_{rev}) appears to slightly boost MT performance without harming QA.
- Allocating 25% of the training budget to monolingual German data does not improve MT and slightly degrades QA.

Despite these trends, due to the small size of the QA evaluation set (n = 162) and the resulting wide confidence intervals, it remains difficult to draw definitive conclusions about which components contributed most to QA performance. Nevertheless, our results confirm the benefit of continued pretraining on task- and language-specific data, particularly when jointly targeting MT and QA with a single model.

Our final submission additionally used one-shot prompting for QA and beam search with a beam size of 4, which slightly increased the scores, although this increase is likely not significant.

Our submission achieved the highest rank in QA and MT tasks for both languages and was also the overall winner in those languages.

5.2 Combined or separate Sorbian training

Since they are closely related, we also investigate how much the Sorbian languages benefited from joint training. From the results in Table 6, we see that both languages benefit from the joint training, especially for the generative MT task. It is also apparent that for the QA task, the model trained on only one of the Sorbian languages can perform quite well for the other, even without the data, suggesting that we gain language understanding from the other Sorbian language.

	MT (BLEU)		**QA** (acc)	
Sorbian data	**de-dsb**	**de-hsb**	**dsb**	**hsb**
hsb + dsb	**60.2 ± 1.2**	**67.2 ± 0.7**	67.5 ± 7.2	**73.1 ± 6.6**
dsb	54.0 ± 1.0	9.7 ± 0.4	**69.6 ± 7.0**	67.3 ± 7.0
hsb	11.8 ± 0.6	65.6 ± 0.7	61.0 ± 7.4	71.8 ± 6.8

Table 6: MT (BLEU) and QA (acc) scores when training the Sorbian languages **together vs separately**. *hsb+dsb* configuration is equal to (3) in Table 5. Zero-shot prompting with a chat template was used.

5.3 Effect of the document-level data

		MT (BLEU)		**QA** (acc)	
Sorbian	N_{chars}	**de-dsb**	**de-hsb**	**dsb**	**hsb**
sent-level	181.5M	11.9 ± 1.2	19.8 ± 1.7	**71.3 ± 6.8**	71.3 ± 6.7
doc-level	104.8M	21.3 ± 0.9	29.3 ± 0.7	**71.3 ± 6.8**	70.4 ± 6.7
combined	269.1M	**24.4 ± 1.0**	**32.5 ± 0.7**	68.8 ± 6.8	**71.9 ± 6.6**

Table 7: Results for training with HSB and DSB data either **document level, sentence-level, or combined** (without MT examples). The training data includes 1 epoch of *INST* and 25% token budget for *deu* (doc-level). 5-shot BLEU scores are reported for MT, and 3-shot accuracy is reported for QA. N_{chars} is the number of characters in the *hsb* and *dsb* datasets

We examine the effect of incorporating document-level data in addition to sentence-level data. No explicit *hsb/dsb* MT training data was used in these experiments. The results in Table 7 indicate that document-level data is more beneficial than sentence-level data for MT when evaluating in a few-shot setting. This is somewhat surprising given that the sentence-level dataset is substantially larger, and the quality of the Fineweb-2 portion of the document-level dataset has not been verified. Nevertheless, combining document- and sentence-level datasets yields the highest MT scores. For QA, we did not observe significant differences between the data setups.

5.4 MT Supervised Fine-tuning

We instruction-tune our model as a separate step with machine-translation examples (4 epochs of *de-hsb*, *de-dsb*, *hsb-dsb*, and *dsb-hsb*) formatted in chat format as instructions, instead of training jointly. In Table 8, this approach shows a slight increase in the BLEU scores of the machine translation benchmarks. However, our model loses its capability to answer questions, so this strategy does not satisfy the goals of the shared task.

Continued pretraining						SFT_{MT} 💬	MT (BLEU)		QA (acc)	
dsb	hsb	MT 💬	MT_{rev} 💬	Inst 💬	deu		de-dsb	de-hsb	dsb	hsb
4x	4x	4x	1x	1x	25%	-	62.1 ± 1.1	68.6 ± 0.7	**65.7 ± 7.1**	**65.1 ± 6.9**
4x	4x			1x	25%	4x	**64.8 ± 1.0**	**70.6 ± 0.6**	36.6 ± 7.1	34.7 ± 7.2

Table 8: **Machine translation SFT** results (validation set) evaluated with a zero-shot chat format. 💬 - conversational instruction format.

	MT (BLEU)		QA (acc)	
Model	**de-dsb**	**de-hsb**	**dsb**	**hsb**
CPT	**62.1 ± 1.1**	**68.6 ± 0.7**	65.7 ± 7.1	65.1 ± 6.9
CPT + 0.1 BASE	61.8 ± 1.1	68.4 ± 0.7	66.6 ± 7.1	67.1 ± 7.0
CPT + 0.3 BASE	58.1 ± 1.1	65.8 ± 0.7	**67.9 ± 7.0**	**71.1 ± 6.8**

Table 9: **SLERP-merged models'** zero-shot chat-formatted validation set scores. +x BASE means that the merging weights of the base model (`Qwen2.5-3B-Instruct`) are x while the Sorbian continually pretrained (CPT) model weight stays 1.0

5.5 Merging

Inspired by TowerLLM (Rei et al., 2025), who reported that merging with the original model improved general results while not harming translation significantly, we also decided to explore merging. SLERP merging with `mergekit` (Goddard et al., 2025) did show a slight increase in the QA scores (see Table 9), although its significance is questionable due to the small size of the validation set. We also noticed that merging started harming the translation quality at 0.3 weight for the base instruction-tuned model. With this in mind, we did not use it for the final model. It is possible that the benefit would be more apparent on other tasks that we did not measure due to the lack of validation data in the Sorbian languages.

6 Conclusion

We presented the TartuNLP submission to the WMT25 Shared Task on Limited Resource Slavic Languages, targeting both machine translation and question answering for Upper and Lower Sorbian. Our approach combined continued pretraining and instruction tuning in a single step, leveraging the Qwen2.5-3B-Instruct model. By integrating task-specific and monolingual Sorbian data, we achieved significant improvements over existing baselines and obtained the highest rank for both tasks in both languages. Our results demonstrate that a unified approach can effectively serve multiple low-resource language tasks, even under resource constraints.

7 Limitations

Our findings are limited by the fact that we only consider two tasks. Also, the MT data seems to have short sentence-level data, limiting the conclusions further. Since the test set of the QA task is relatively small, we have rather low confidence in the minor differences in scores of the approaches. We did not thoroughly explore all the design choices made for this system, and we did not explore data filtering, which could significantly impact the resulting model.

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

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

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

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

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.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2025. Arcee's mergekit: A toolkit for merging large language models. *Preprint*, arXiv:2403.13257.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Indraneil Paul, Hengyu Luo, and Jörg Tiedemann. 2025a. Massively multilingual adaptation of large language models using bilingual translation data. *Preprint*, arXiv:2506.00469.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2025b. Emma-500: Enhancing massively multilingual adaptation of large language models. *Preprint*, arXiv:2409.17892.

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.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Pedro Henrique Martins, João Alves, Patrick Fernandes, , Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, 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.

Christopher Moseley. 2007. *Encyclopedia of the world's endangered languages*. Routledge.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Shu Okabe, Daryna Dementieva, Marion Di Marco, Lukas Edman, Kathy Hämmerl, Marko Měškank, Anita Hendrichowa, and Alexander Fraser. 2025. Findings of the WMT 2025 shared task for LLMs with limited resources for slavic languages: MT and QA. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

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.

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. *Preprint*, arXiv:2506.20920.

Kevin Heffernan Onur Çelebi Anna Sun Ammar Kamran Yingzhe Guo Alexandre Mourachko Holger Schwenk Angela Fan Pierre Andrews, Guillaume Wenzek. 2022. stopes - modular machine translation pipelines. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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 6677–6697, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André

F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, and 1 others. 2025. Instructing large language models for low-resource languages: A systematic study for basque. *arXiv preprint arXiv:2506.07597*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*.

A Hyperparameters

The training hyperparameters for the submitted model are in Table 10.

B Evaluation prompts

Evaluation prompts are presented in Figure 2.

C Instruction-tuning dataset overview

The full overview of the instruction data composition is presented in Table 11.

Hyperparameter	Value
Learning rate	1e-4
Optimizer	`AdamW`
Adam ϵ	1e-8
Adam β_1, β_2	0.9, 0.95
Sequence length	4096
Weight decay	0.1
Scheduler	`warmup-stable-decay`
Warmup steps	256
Decay steps	768
FSDP Strategy	`SHARD_GRAD_OP`
GPUs	64
Precision	`bfloat16`
Batch size (total)	128
Batch size (tokens)	524288
Training steps	2285
Training tokens	1,197,871,104

Table 10: Hyperparameters for the training of the submitted model.

MT

```
Translate the text from German to Lower Sorbian.\n\nGerman: {{de}}\nLower Sorbian:
```

MT (chat)

SYSTEM:

```
You are are a professional translator. Translate the following text from German to Lower Sorbian. Answer with the translated text.
```

USER:

```
{{de}}
```

QA

```
{context}\n\nQuestion:\n{question}\n\nPossible answers:\nA {answer_1}\nB {answer_2}...\nZ {answer_n}\n\nAnswer:
```

Figure 2: Prompts used for evaluation with `lm-eval-harness` (Gao et al., 2024).

Dataset		eng	pol	deu	ces	slk	slv	Total
CohereLabs/aya_dataset	Singh et al. (2024)	3944	1483	241	0	0	0	5668
Magpie-Align/Magpie-Llama-3.1-Pro-MT-300K-Filtered	Xu et al. (2025)	295830	36	228	21	3	3	296121
OpenAssistant/oasst2	Köpf et al. (2023)	22311	155	1785	5	0	0	24256
ai2-adapt-dev/flan_v2_converted	Longpre et al. (2023)	89982	0	0	0	0	0	89982
utter-project/EuroBlocks-SFT-Synthetic-1124†	Martins et al. (2025)	3057	916	1019	0	0	0	4992
Total		415124	2590	3273	26	3	3	421019

Table 11: Overview of the **instruction datasets**. †- *multilingual-synthetic-mmlu*, *synthetic-eurollm-9B*, and *multilingual-synthetic-arc* subsets from EuroBlocks.