Data-Centric Approach at the LoResMT 2026 Turkic Translation Challenge: Russian-Kyrgyz

Dmitry Novokshanov
HSE University
danovokshanov@gmail.com

Abstract

We describe our submission to the Turkic languages translation challenge at LoResMT 2026, which focuses on translation from Russian into Kyrgyz. Our approach leverages parallel data, synthetic translations, a comprehensive filtering pipeline and a four-stage curriculum learning strategy. We compare our system with contemporary baselines and present the model that achieves a chrF++ score of 49.1 and takes first place in the competition.

1 Introduction

Machine translation (MT) has witnessed remarkable progress with the emergence of neural machine translation (NMT) (Bahdanau et al., 2014). Predominantly driven by transformer-based architectures (Vaswani et al., 2017), this technology has pushed the boundaries of translation quality further, with systems now achieving human-level performance in some domains on high-resource language pairs. However, these breakthroughs remain largely inaccessible to the majority of the world's approximately 7,000 languages, as most language pairs lack sufficient parallel data to train competitive MT systems (Haddow et al., 2022). The Turkic language family presents a particularly compelling case study for low-resource MT research. Comprising about 30 languages spoken by approximately 200 million people across a vast geographic region (Rybatzki, 2020), Turkic languages share rich morphological complexity characterized by agglutinative structure, vowel harmony, and head-final syntax (Johanson and Csató, 2015). Spoken by around 5 million people primarily in Kyrgyzstan, Kyrgyz lacks large-scale parallel corpora and still presents a challenge in the MT task (Alekseev and Turatali, 2024; Mirzakhalov et al., 2021). This paper describes our system for the Russian-Kyrgyz translation track of the Turkic Languages Translation Challenge at LoResMT 2026. Our approach is based on three key steps: (1) comprehensive collection of available data and augmentation with synthetic translations utilizing modern LLMs; (2) complex data filtering pipelines; and (3) a four-stage training methodology.

2 Methodology

2.1 Data

Training an effective neural machine translation system for low-resource language pairs requires careful attention to data quality and diversity. In this section, we describe our data collection, filtering pipeline, and the construction of training, validation, and test sets.

2.1.1 Training Data

Primary Sources. We utilized OPUS (Tiedemann, 2012) as our primary source of parallel data, extracting sentence pairs from three languages in all combinations: Russian, Kyrgyz, Uzbek, and Tajik. The inclusion of auxiliary language pairs follows the established practice of leveraging transfer learning to improve translation quality for low-resource targets (Zoph et al., 2016; Nguyen and Chiang, 2017; NLLB Team et al., 2022). Uzbek and Tajik were selected not only due to their geographic proximity, but also due to typological similarity in the case of Uzbek and a shared writing system (Cyrillic script) in the case of Tajik.

Our preprocessing pipeline consisted of several stages. First, we removed duplicate sentence pairs and entries containing null values. We then applied regular expression-based filtering to eliminate examples consisting solely of special symbols, punctuation marks, or digits, as well as those written in inappropriate scripts. Additionally, we filtered out sentences with disproportionately high ratios of special characters to alphabetic content. We also used the fastText language detection model (Grave et al., 2018) on top of that to ensure we obtained

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 225–230

correct language pairs.

Following basic preprocessing, we scored all parallel sentences using SONAR (Duquenne et al., 2023), specifically the `blaser_2.0` model, which provides cross-lingual semantic similarity scores. We discarded all sentence pairs scoring below 3.0, as these typically indicated misaligned or low-quality translations. The remaining data was partitioned into two quality tiers:

- **Standard quality**: sentence pairs with SONAR scores in the range $[3.0, 3.7)$
- **High quality**: sentence pairs with SONAR scores ≥ 3.7

Supplementary Sources. To enhance domain coverage and conversational fluency, we incorporated additional synthetic data from two sources. First, we leveraged FineWeb-2 (Penedo et al., 2024), a large-scale multilingual web corpus. We sampled 500,000 Kyrgyz-language documents from FineWeb-2; documents exceeding 500 tokens were segmented at sentence boundaries (splitting on end-of-sentence punctuation marks) to produce training-amenable chunks. This process yielded 1,849,234 Kyrgyz examples. These were then back-translated (Sennrich et al., 2015) into Russian using two recent large language models: Gemma-3-27B (Gemma Team, 2025) and Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025). The resulting synthetic parallel data underwent our standard filtering pipeline with the addition of LLM-specific filters to remove artifacts such as model-generated notes, meta-commentary, and cases where the model simply repeated the input sequence. Additionally, we computed the cross-entropy loss of the MADLAD-400-7B-MT model (Kudugunta et al., 2024) on each example pair, retaining only those with loss values in the range (0.1,4.5) to exclude deviations at both ends. After applying the filtering pipeline, we retained 1,566,927 sentences for training.

Second, we utilized the SiberianPersonaChat dataset (Denis Petrov, 2023), a Russian-language dialogue corpus. We selected 56,529 dialogue samples with sequence lengths under 507 tokens and translated them from Russian into Kyrgyz using GPT-4o (OpenAI, 2024). This synthetic parallel data provides coverage of informal, conversational language patterns that are typically underrepresented in web-crawled corpora.

Split	Source	Sentences
Train	OPUS (standard quality)	8,207,314
	OPUS (high quality)	7,069,840
	Synthetic Fineweb-2	3,133,854
	Synthetic dialogue	113,058
Valid	FLORES-200 dev (st. 1-2)	11,964
	FLORES-200 dev (st. 3)	1994
	Synthetic (held-out)	3,902
Test	FLORES-200 devtest	1,012
	Shared task test	2,311

Table 1: Resulting dataset statistics for training, validation, and test splits. The number of sentences includes all translation directions.

2.1.2 Validation Data

We employed two validation sets to monitor training progress and perform hyperparameter selection:

1. **FLORES-200 dev**: The development split of the FLORES-200 dataset (NLLB Team et al., 2022), which has become a standard MT benchmark. We use this set to validate all 12 directions during the first two stages of training and only the Russian-Kyrgyz pair in stage 3.

2. **Synthetic dialogue**: A held-out portion of the GPT-4o translated dialogue data, reserved for validating performance on longer sequences. This set is used to validate stage 4.

2.1.3 Test Data

For evaluation of the target Russian→Kyrgyz direction, we utilized two test sets:

1. **FLORES-200 devtest**: The devtest split of FLORES-200, used for internal evaluation and comparison with baselines.

2. **Shared task test set**: The official blind test set provided by the LoResMT 2026 organizers.

2.1.4 Data Statistics

Table 1 summarizes the statistics of our training, validation, and test datasets.

2.2 Evaluation Methodology

We evaluate our models using two complementary metrics that capture different aspects of translation quality.

Training Data	xCOMET	chrF++
ru-ky only	16.9	21.5
ru-ky + ru-uz	17.4	22.1
ru-ky + ru-uz + ru-tg	**18.7**	**23.2**

Table 2: Ablation study on auxiliary language inclusion using mT0-small (300M). Results on FLORES-200 devtest (ru→ky). Best configuration in **bold**.

chrF++. As the primary metric for the LoResMT 2026 shared task, we report chrF++ (Popović, 2017), a character n-gram F-score metric that additionally incorporates word unigrams and bigrams. chrF++ is particularly well-suited for morphologically rich languages like Kyrgyz, as it captures partial matches at the subword level. All chrF++ scores reported in this paper are computed using the Hugging Face `evaluate` library (von Werra et al., 2022).

xCOMET-XXL. To complement the n-gram-based evaluation, we additionally report scores from xCOMET-XXL (Guerreiro et al., 2024), a learned sentence-similarity evaluation metric. Scores are reported multiplied by 100 for uniformity.

2.3 Training Methodology

Our training approach consists of two main phases: (1) preliminary experiments to validate data composition choices; and (2) a four-stage curriculum training procedure that progressively refines the model on increasingly high-quality data.

2.3.1 Auxiliary Language Selection

Before committing to our full training pipeline, we conducted ablation experiments to verify that incorporating Uzbek and Tajik parallel data alongside Russian-Kyrgyz would benefit translation quality. Using mT0-small (Muennighoff et al., 2023), a 300M-parameter multilingual encoder-decoder model, we trained separate models on different data configurations and evaluated them on FLORES-200 devtest.

As shown in Table 2, the combination of all three language pairs yielded the best performance, confirming that transfer learning from both Uzbek and Tajik provides complementary benefits.

2.3.2 Model Selection and Vocabulary Pruning

Based on the positive results from our preliminary experiments, we selected *mt0-large* (Muennighoff et al., 2023) as our base model. mT0, a variant of mT5 (Xue et al., 2021) additionally tuned on a diverse crosslingual task mixture, is well-suited for cross-lingual transfer to low-resource targets. The original *mt0-large* model contains approximately 1.23 billion parameters, with a significant portion allocated to the embedding layers covering the full mT5 vocabulary of 250,000 tokens. To improve computational efficiency, we applied vocabulary pruning (Zhu and Gupta, 2017). Specifically, we retained only tokens that appear at least once in our combined training corpus, reducing the vocabulary size substantially. This pruning reduced the total model size from 1.23B to approximately 800M parameters. The pruned model maintains identical architecture and pretrained weights for all non-embedding parameters.

2.3.3 Four-Stage Curriculum Training

We employ a curriculum learning strategy (Bengio et al., 2009) that progressively trains the model on data of increasing quality and domain specificity. This approach allows the model to first learn general translation patterns from larger but noisier data, then refine its outputs on cleaner, more targeted examples.

Stage 1: Standard Quality Multilingual Data. The pruned model is first trained on the standard quality tier of our OPUS-derived data (SONAR scores in $[3.0, 3.7)$) across all language pairs. This stage exposes the model to a large volume of parallel text, establishing foundational translation capabilities.

Stage 2: High Quality Multilingual Data. The best checkpoint from Stage 1 is further trained on the high quality tier (SONAR scores ≥ 3.7). This smaller but cleaner dataset helps the model refine its translations and reduce errors introduced by noisy training examples.

Stage 3: Open-Source Synthetic Data. The best Stage 2 checkpoint is trained on machine-translated data derived from FineWeb-2 (Penedo et al., 2024). This web-crawled multilingual corpus provides broad domain coverage and exposes the model to diverse vocabulary, sentence structures, and longer texts.

Model	chrF++	xCOMET
Open-Source Baselines		
Gemma3 27B	40.2	67.1
Qwen3 235B	36.2	55.5
GPT-oss 120B	37.4	61.9
MADLAD-400 7B	40.8	74.9
NLLB-200 54B	<u>41.7</u>	<u>73.4</u>
Our Models		
Ours (Stage 1)	32.2	38.5
Ours (Stage 2)	43.0	73.1
Ours (Stage 3)	44.8	78.8
Ours (Stage 4 / Final)	**44.9**	**80.5**

Table 3: Comparison of our system with baseline models on FLORES-200 devtest (ru→ky). Best open-source result <u>underlined</u>, overall best in **bold**.

Stage 4: High-Quality Synthetic Dialogue Data. Finally, the best Stage 3 checkpoint is fine-tuned on our GPT-4o translated dialogue data from *Siberian-PersonaChat*. This final stage adapts the model to conversational register.

2.3.4 Training Configuration

All training stages were conducted in a consistent environment with hyperparameters determined through preliminary experimentation. Each stage was trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 1×10^{-3}, a label smoothing factor of 0.1, and varying learning rates. Exact hyperparameters for each stage can be provided upon request.

3 Results

3.1 Comparison with Baselines

Table 3 presents a comparison of our system against publicly available baseline models on the FLORES-200 devtest set for Russian-to-Kyrgyz translation.

Our final model outperforms all open-source baselines on both metrics. Notably, we surpass NLLB-200 54B—a model nearly 70 times larger than ours—by 3.2 chrF++ points and 7.1 xCOMET-XXL points. The comparison with general-purpose large language models is particularly striking. The progression across training stages illustrates the effectiveness of the curriculum learning approach. Moreover, our model also shows high performance in the opposite Kyrgyz-to-Russian direction with 42.4 chrF++ and 82.8 xCOMET-XXL scores.

3.2 Shared Task Results

On the official LoResMT 2026 shared task test set (combined public and private portions), our system achieves **49.1 chrF++** and **69.7 xCOMET-XXL**, securing first place in the Russian-to-Kyrgyz translation track.

3.3 Released Resources

As a contribution to the research community and to support further work on Kyrgyz language NLP, we publicly release artifacts developed in this work:

- **Model checkpoint**: The final Stage 4 model is available on Hugging Face.[1]
- **Synthetic parallel data**: Our filtered FineWeb-2 and GPT-4o-translated datasets are released for research purposes.[2]
- **Interactive demo**: A demonstration interface is available on Hugging Face Spaces, allowing users to test Russian-to-Kyrgyz translation without local installation.[3]

4 Conclusion

We presented our winning system for the Russian-to-Kyrgyz translation track of the LoResMT 2026 Turkic Languages Translation Challenge. Our approach combines careful data curation from OPUS corpora with synthetic data generation and complex filtering pipelines. The resulting 800M-parameter model outperforms substantially larger baselines—including NLLB-200 54B, Gemma3 27B, and Qwen3 235B—achieving 44.9 chrF++ on FLORES-200 devtest and 49.1 chrF++ on the official shared task evaluation. We release our model and synthetic datasets to support future research on Kyrgyz and other low-resource Turkic languages.

Acknowledgments

The author is highly grateful to Dmitriy Akimov, German Beyger, Yuliana Sidelnikova, and Anton Polevoi for their support along the author's journey in machine translation and valuable advice in building this system in particular.

[1] https://huggingface.co/Novokshanov/ru-ky-mt0-loresmt2026

[2] https://huggingface.co/datasets/Novokshanov/ru-ky-synthetic-loresmt2026

[3] https://huggingface.co/spaces/Novokshanov/ru-ky-loresmt2026-demo

References

Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: Challenges, progress, and future. *Preprint*, arXiv:2411.05503.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Ivan Ramovich Denis Petrov. 2023. Russian dataset for chat models.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nuno M. Guerreiro, Ricardo Rei, Sara Stymne, Alon Lavie, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Lars Johanson and Éva Á Csató. 2015. *The Turkic Languages*. Routledge.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusber, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Raber, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15991–16111. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2024. GPT-4o system card.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Thomas Wolf, and Leandro von Werra. 2024. FineWeb-2: A 14 trillion token multilingual web dataset. *Hugging Face Technical Report*.

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

Volker Rybatzki. 2020. The altaic languages: Tungusic, mongolic, turkic. In *The Oxford Guide to the Transeurasian Languages*, pages 22–28. Oxford University Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 5998–6008. Curran Associates, Inc.

Leandro von Werra, Lewis Tunstall, Nandan Thakur, Joe Davison, Yacine Jernite, Tristan Moran, and Thomas Wolf. 2022. Evaluate: A library for easily evaluating machine learning models and datasets.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Wang, Bowen Lin, Bwen Hui, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *Preprint*, arXiv:1710.01878.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.