DevLake at LoResMT 2026: The Impact of Pre-training and Model Scale on Russian-Bashkir Low-Resource Translation

Vyacheslav Tyurin
DevLake Team
1voldis1l@gmail.com

Abstract

This paper describes the submission of Team **DevLake** for the LoResMT 2026 Shared Task on Russian-Bashkir machine translation. We conducted a comprehensive comparative study of three distinct neural architectures: NLLB-200 (1.3B), M2M-100 (418M), and MarianMT (77M). Our primary goal was to evaluate whether parameter-efficient fine-tuning (PEFT) of massive models outperforms full training of compact architectures in a low-resource setting. To achieve this, we employed QLoRA for large models and vocabulary expansion techniques for the smaller MarianMT. We also implemented a rigorous data filtering pipeline using a domain-specific BERT classifier. Our results demonstrate that model scale and "native" pre-training coverage are decisive factors: our best system, a fine-tuned **NLLB-200-1.3B** model, achieved a CHRF++ score of **52.67**, significantly outperforming the compact baseline (43.15) despite the latter's extensive training on a larger dataset. We release our code and trained models to facilitate further research.

1 Introduction

Machine translation (MT) for low-resource languages remains a challenging frontier in Natural Language Processing. Bashkir, a Turkic language with rich agglutinative morphology and approximately 1.2 million speakers, suffers from a scarcity of high-quality parallel corpora compared to high-resource languages like English or Russian.

The LoResMT 2026 Shared Task provided a training dataset of approximately 1.2 million Russian-Bashkir sentence pairs. While the volume of data appeared substantial, preliminary analysis revealed mixed quality, including noise, misalignments, and code-switching. Our participation was driven by a practical engineering question: *Is it better to fine-tune a massive "generalist" model using quantization or to train a specialized "lightweight" model from scratch?*

Using a single **NVIDIA RTX 3080 (10GB)**, we explored both directions. We found that "vocabulary surgery" on small models (MarianMT) leads to grammatical fluency but semantic hallucinations, whereas quantized fine-tuning of large models (NLLB) yields superior translation quality.

2 Related Work

Recent advances in multilingual NMT have shifted focus from training bilingual models to fine-tuning massive pre-trained transformers. NLLB-200 (Costa-jussà et al., 2022) sets the state-of-the-art for many low-resource languages by leveraging a Mixture-of-Experts architecture and massive data mining. Similarly, M2M-100 (Fan et al., 2021) demonstrated that direct translation between non-English pairs is viable without English as a pivot.

For efficient training, techniques like LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) have democratized access to large models, allowing 1B+ parameter models to be trained on consumer hardware. Our work builds on these foundations, applying them specifically to the Cyrillic Turkic context.

3 Data Preparation

The official training dataset contained noise. To ensure model stability, we implemented a strict semantic filtering pipeline using a specialized metric: **slone/bert-base-multilingual-cased-bak-rus-similarity** (Slone Team, 2023). This BERT-based model predicts whether a Russian and Bashkir sentence pair carries the same meaning.

We created two distinct splits to test different hypotheses:

- **High-Precision Set (≥ 0.80):** Approximately 486,000 pairs. This high-quality subset was crucial for fine-tuning our largest model (NLLB) to refine its style without polluting it with noise or hallucinations.

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 209–212

- **Massive Set ($\geq$ 0.10):** Approximately 923,000 pairs. Used for training smaller models (MarianMT) to maximize their exposure to rare vocabulary and morphological forms.

4 Methodology

We experimented with three distinct architectures, representing different scales and pre-training strategies. Table 1 details our training configuration.

Parameter	NLLB	MarianMT
Base Model Size	1.3B	77M
Fine-Tuning	QLoRA	Full FT
Precision	4-bit	FP32
Learning Rate	$2e^{-4}$	$5e^{-5}$
Epochs	1	3
Data Size	486k	923k

Table 1: Comparison of training configurations.

4.1 System 1: NLLB-200 (1.3B)

Our primary system is based on **NLLB-200-1.3B-Distilled**. This model is particularly suitable because it explicitly includes Bashkir (`bak_Cyrl`) in its pre-training data.

Optimization: Fitting a 1.3B parameter model into 10GB VRAM is impossible with standard training. We utilized **QLoRA**. The base model was frozen and loaded in 4-bit precision ('nf4'). Trainable LoRA adapters were attached to the attention modules.

- **LoRA Config:** $r = 64$, $\alpha = 64$, dropout = 0.1.
- **Targets:** We targeted all linear layers: $q, k, v, o, gate, up, down$ projections.

We trained for 1 epoch on the High-Precision Set using the AdamW optimizer.

4.2 System 2: M2M-100 (418M)

We fine-tuned **facebook/m2m100_418M**. While smaller than NLLB, it offers a robust baseline. We trained it for 3 epochs on the medium-quality data split ($\geq$ 0.60). This model served as a stable fallback for our ensemble experiments.

4.3 System 3: MarianMT (77M)

We conducted an extensive experiment with **Helsinki-NLP/opus-mt-en-trk** (Junczys-Dowmunt et al., 2018). This 77M parameter model is efficient but was trained for English-Turkic translation and does not know Russian or the Bashkir Cyrillic script.

Vocabulary Adaptation: Instead of replacing the tokenizer entirely, we manually extended the existing vocabulary with missing Bashkir Cyrillic characters (e.g., 'Ҕ', 'Ө', 'Ҫ') and resized the embedding layer to accommodate the new tokens.

Training: We performed **Full Fine-Tuning** (all parameters unfrozen) in FP32 precision for 3 epochs on the Massive Set (923k pairs). To facilitate transfer learning, we prepended the '»bak«' token to all source sentences.

5 Post-Processing

Translation models often suffer from specific artifacts. We implemented a post-processing pipeline to address them.

5.1 Exact Match Retrieval

We hypothesized that the test set might overlap with the training data. We indexed the entire training corpus and checked for exact matches with the test source sentences. We found **7 exact matches**. For these cases, we bypassed the model and injected the ground truth translation, guaranteeing 100% accuracy for these samples.

5.2 Inference Heuristics

NLLB models are prone to "repetition loops" (generating the same word indefinitely). To counter this, we enforced 'no_repeat_ngram_size=3' and a repetition penalty of 1.2 during Beam Search ($k = 5$).

6 Results and Analysis

We evaluated our models using the Corpus CHRF++ metric on the official leaderboard. Additionally, we performed a manual inspection of the outputs to understand the qualitative differences between architectures.

6.1 Quantitative Results

Table 3 summarizes the performance. The correlation between model scale and performance is evident.

6.2 Qualitative Analysis: The Impact of Scale

Our manual analysis revealed critical differences in robustness between the large and small models. Table 2 demonstrates specific failure modes encountered during testing.

Source (Russian)	MarianMT (77M)	NLLB-200 (1.3B)
Яблочный сидр (Apple cider)	Яблочный ултырма. (Apple **do not sit / planting**)	Алма сидары (Apple cider)
Яблочный сидр, (Apple cider,)	Яблочный ултыра, (Apple **is sitting**,)	Алма сидары, (Apple cider,)
Яблочный сидр, пожалуйста! (Apple cider, please!)	Япраҡлы ултырғыс, зинһар! (**Leafy chair**, please!)	Алма сидары, зинһар! (Apple cider, please!)
Через пару часов, окей? (In a couple of hours, okay?)	Ике сәғәттән (In two hours)	Бер-ике сәғәттән, окей? (In a few hours, okay?)

Table 2: Comparison of model robustness. MarianMT exhibits severe hallucinations and input sensitivity, while NLLB remains stable.

Model	Params	Strategy	CHRF++
MarianMT	77M	Full FT (923k)	43.15
M2M-100	418M	LoRA (900k)	48.80
NLLB-200	**1.3B**	**QLoRA (486k)**	**52.67**

Table 3: Official leaderboard results. Despite using less training data (High-Precision Set), the NLLB model achieved the highest score due to its pre-training quality.

1. Model Brittleness and Input Sensitivity: As shown in Table 2, the MarianMT model is highly unstable. Adding a comma or an exclamation mark completely changes the semantic output.

- The phrase "Apple cider" was hallucinated as "planting" or "sitting" depending on punctuation.
- Adding "please" triggered a complete semantic collapse, generating "Leafy chair" (Япраҡлы ултырғыс).

This suggests that the small model relies heavily on surface-level statistics and subword combinations, failing to capture the robust semantic representation of the source sentence.

2. Lexical Precision: NLLB-200 consistently produced the correct terminology ("Apple cider" → Алма сидары) regardless of punctuation changes. It also correctly handled the colloquial "okay" and the approximate time expression "couple of hours" (бер-ике сәғәттән), whereas MarianMT reverted to literal translations.

7 Reproducibility

To facilitate future research in low-resource Turkic languages, we release our code and trained adapters. We ensured that all experiments can be reproduced on consumer-grade hardware.

- **Codebase:** The complete training and inference pipeline is available on GitHub: https://github.com/Voldisoriginal/LoResMT-2026-Russian-Bashkir.
- **Model Checkpoints:** We uploaded the fine-tuned models to the Hugging Face Hub:
  - **NLLB-1.3B:** https://huggingface.co/Voldis/nllb-1.3b-rus-bak
  - **M2M-100:** https://huggingface.co/Voldis/m2m100-rus-bak
  - **MarianMT:** https://huggingface.co/Voldis/marian-rus-bak
- **Technical Details:** All models were trained using **PyTorch 2.5.1**, **Transformers 4.46.0**, **PEFT 0.12.0**, and **bitsandbytes 0.49.0**. We used a fixed random seed (42) for data splitting and initialization to ensure deterministic results.

8 Conclusion

Our participation in LoResMT 2026 highlights that for low-resource languages, leveraging massive pre-trained models (like NLLB) via quantization is significantly more effective than training smaller, specialized architectures from scratch. The "knowledge" embedded in the 1.3B parameters of NLLB regarding Bashkir morphology outweighed the agility of the 77M MarianMT model, even when the latter was trained on twice as much data.

Limitations

While our NLLB-based approach yielded the best results, it comes with computational constraints. The inference of a 1.3B model, even in 4-bit quantization, requires approximately 2GB of VRAM and has higher latency compared to the CPU-friendly MarianMT (77M). Additionally, our filtering pipeline

relies on a multilingual BERT model; biases inherent in BERT could potentially exclude valid but rare dialectal variations of Bashkir from the training set.

References

Marta R. Costa-jussà and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers and 1 others. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Angela Fan and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Edward J Hu and 1 others. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Marcin Junczys-Dowmunt and 1 others. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*.

Slone Team. 2023. Bert-base-multilingual-cased-bak-rus-similarity. slone/bert-base-multilingual-cased-bak-rus-similarity.