No One-Size-Fits-All: Building Systems For Translation to Bashkir, Kazakh, Kyrgyz, Tatar and Chuvash Using Synthetic And Original Data

Dmitry Karpov
PAO Severstal / Moscow, Russia
dimakarp1996@yandex.ru

Abstract

We explore machine translation for five Turkic language pairs: Russian-Bashkir, Russian-Kazakh, Russian-Kyrgyz, English-Tatar, English-Chuvash. Fine-tuning nllb-200-distilled-600M with LoRA on synthetic data achieved chrF++ 49.71 for Kazakh and 46.94 for Bashkir. Prompting DeepSeek-V3.2 with retrieved similar examples achieved chrF++ 39.47 for Chuvash. For Tatar, zero-shot or retrieval-based approaches achieved chrF++ 41.6, while for Kyrgyz the zero-shot approach reached 45.6. We release the dataset and the obtained weights.

1 Introduction

Machine translation for low-resource languages remains challenging. The "Machine Translation for Low-Resource Turkic Languages" competition focused on five pairs: Russian-Kazakh, Russian-Kyrgyz, Russian-Bashkir, English-Tatar, English-Chuvash. We investigate multiple approaches to improve translation quality in data-scarce conditions.

2 Making The Data

Unfortunately, only a limited amount of high-quality parallel data was available at the time of this study. Therefore, we used a variety of datasets. Specifically, we used the parallel English-Chuvash corpus from Plotnikov and Antonov (2024) for translations from English to Chuvash, Zhang et al. (2020) for English-Tatar, English-Kyrgyz, and English-Kazakh translations, NLLB Team et al. (2024) to obtain the parallel data for Russian, English, Bashkir, Tatar, Kazakh and Kyrgyz, ips (2025) for translations from Russian to Tatar, Singh et al. (2024) for the parallel Russian-Kyrgyz data, (Tiedemann, 2020) for the parallel data for all 5 language pairs, Iskander Shakirov (2023) for the parallel Russian-Bashkir data, Yeshpanov et al. (2024) and Abdashim (2024) for the Russian-Kazakh data and gou for Russian-Kyrgyz pair. Unfortunately, we could not use the parallel corpora from the TurkLang-7 project Khusainov and Minsafina (2021) as this corpora was not available at the time of study. The original training set sizes were: 1,190,773 samples for Russian-Bashkir (9,997 in the validation set), 35,429 samples for Russian-Kyrgyz (4,845 in the validation set), 367,095 samples for Russian-Kazakh (9,845 in the validation set), 106,777 samples for English-Tatar (3,786 in the validation set), and 193,826 samples for English-Chuvash (9,953 in the validation set). We report only the official validation scores from the competition leaderboard.

We augmented the original data with synthetic translations from Yandex.Translate(Yandex, 2024). To obtain these translations, we translated English phrases into Russian (for pairs where Russian was not the source language) and then translated every Russian phrase to those Turkic languages. We used the "document translation" feature, processing the data in chunks of 50,000 to 200,000 samples due to its large volume. After obtaining synthetic data, we meticulously filtered out any English or Russian phrases from Yandex.Translate that appeared in the test set for any language. At this stage, we also included Russian-Tatar, English-Kazakh, and English-Kyrgyz data. Data pseudolabeling has long been proven to improve the performance of Transformer-based language models (Karpov and Burtsev, 2021). In our case, pseudolabeling also proved beneficial, as multilingual training without pseudolabeling yielded inferior results in the preliminary experiments.

After augmentation, the training data size increased to 2,457,344 samples for each language pair.

In addition to these data, we also used translations of MASSIVE dataset (FitzGerald et al., 2022) from English to Tatar, Chuvash and Russian, and

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 203–208
March 28, 2026 ©2026 Association for Computational Linguistics

from Russian to the Bashkir, Kyrgyz and Kazakh (16507 samples). All translations were obtained similarly using Yandex.Translate. However, these translations were obtained later and thus were not used for training the NLLB model. They were used only in the prompting-based solutions.

For the prompting-based approach to English-Chuvash, we additionally obtained translations from Chuvash to English for two other alexantonov corpora: alexantonovchuvash_russian_parallel and alexantonovchuvash_mono (Plotnikov and Antonov (2024)) . After adding these data to the previously obtained Chuvash-English pairs, the English-Chuvash dataset increased its size to 6.7 million pairs. All this data was also translated to Tatar via Yandex.Translate, and Tatar translations are also provided. However, this data was NOT used in the submissions for Tatar, as a) they were obtained after the training experiments b) building an index from them (see sections below) resulted in the inferior results on the preliminary experiment.

We release the resulting dataset YaTURK-7lang, translated into the six languages, here `https://huggingface.co/datasets/dimakarp1996/YaTURK-7lang`. Data used only in the Chuvash solution are marked with the attribute `only_index1` set to 0.

3 Kazakh and Bashkir: Where LoRA And Knowledge Transfer Shined

We chose *facebook/nllb-200-distilled-600M* (Team et al., 2022) as the base model for finetuning. The data for finetuning was preprocessed as follows: additional language tokens for each target language and language pair (ten tokens in total) were added into the tokenizer. Thus, the model input consisted of the prefix of the language pair (e.g. `<prefix_rus_bash>`) with tokenized source language text, and trained the model to predict the target language text, starting its output from the token of the target language (e.g. `<prefix_bash>`)

We explored 2 main modes of finetuning the model. In the first mode, the model was finetuned for 2 epochs on the data from every language, separately, In the second mode, the model was first finetuned for one epoch on the data from all languages, and then we trained LoRA adapter for every separate language. Neither using LoRA nor finetuning for more than two epochs improved the results.

Specifically, for training adapters, we used DORA(Liu et al., 2024). DORA is the extension of LoRA(Hu et al., 2021) parameter-efficient finetuning approach. The DORA config was: r=64, alpha=64, LoRA dropout=0.2, PiSSA weight initialization(Meng et al., 2025),target_modules: q_proj, v_proj, k_proj, out_proj, fc1, fc2, and shared. DORA was finetuned with the paged AdamW-8bit(Loshchilov and Hutter, 2017)(Dettmers et al., 2021) optimizer, with starting learning rate 5e-4 and weight decay 1e-2, train batch size 16 and 8 gradient accumulation steps, linear learning rate scheduler. For full finetuning, we used the following hyperparameters: batch size 64, 32 gradient accumulation steps, learning rate 2e-4, weight decay 1e-2, 600 warming steps per epoch, paged AdamW-8bit optimizer, cosine learning rate scheduling. In both cases, the maximum sequence length was 128 tokens, and the optimizer state was reset every epoch. For all models, to obtain generation, we used the following generation settings: min_length=3, max_length=150, repetition penalty 1.5, 5 beams.

Table 1: Validation set chrF++ (from the competition server) of the NLLB model. Mult-1 means the results of the model finetuned on 1 epoch. LoRA means the results of the LoRA adapters trained on top of this model. Finetune means the results of the single-task finetune. The final submissions are in bold, where it is applicable.

Language	Mult-1	LoRA	Finetune
Bashkir	22.32	**49.53**	26.92
Kazakh	40.96	**49.93**	**44.70**
Kyrgyz	21.77	36.29	27.04
Tatar	23.95	32.13	28.81
Chuvash	10.86	11.32	11.70

As one can see from Table1 , the approach of training the model on multiple languages and then finetuning using LoRA outperformed the single-task finetuning, which suggests that knowledge transfer occurs between tasks. Multi-task knowledge transfer has been studied for a long time (Karpov and Konovalov, 2023). As the Turkic languages in this study are similar, knowledge acquired for one language can help improve performance on another.

Although this approach seemed promising, we did not pursue it further due to limited computational resources. However, the Bashkir and Kazakh solutions obtained at this experiment were submitted as final ones. Specifically, for Bashkir we have submitted the single-language finetune result as well as the LoRA result. For Kazakh we have submitted LoRA result and the stacking result. This yielded test scores chrF++ 49.71 for Kazakh and 46.94 for Bashkir. **We release the weights for Kazakh and Bashkir models at `https://huggingface.co/dimakarp1996`**. Repository names: multitask_finetune_nllb600 for the 5-language finetune, adapter_kz_nllb600 and adapter_ba_nllb600 for LoRA adapters, finetune_ba_nllb600 and finetune_kz_nllb600 for single-language finetunes.

4 Chuvash and Tatar: Exploring Prompting

We also explored another approach to build a machine-translation system. Due to budget constraints, we used ANNOY-based indexes. We built an ANNOY index from the source-language phrases in the existing dataset. Then, for every new phrase of the source language, we retrieved the most similar phrases of the **source** language in the dataset. Each phrase and its translation were appended to the prompt for a large language model.

We built the ANNOY index with an embedding dimension of 384, with the cosine similarity metric, and 100 trees. For the English-Chuvash pair, we used thenlper/gte-small (Li et al., 2023) vectorizer and all data from YaTURK-7lang, whereas for all other pairs we used sentence-transformers/paraphrase- multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019) and only those data from YaTURK-7lang where the attribute only_index1=1 .

For the English-Chuvash pair, in the final experiments, we set up a very large `TOP_N` (7000). `SEARCH_K` was equal to `2*N_TREES*TOP_N` for all cases.

The models we prompted in this study were: DeepSeek-R1-0528(DeepSeek AI, 2025a), DeepSeek-V3.1 Nex-N1(Nex AGI and DeepSeek AI, 2025), XiaomiMiMo/MiMo-V2-Flash(Xiaomi MiMo, 2025), Gemma3-27b(Google, 2025) and DeepSeek-V3.2Exp(DeepSeek AI, 2025b). We refer to these models as DeepSeek-R1, DeepSeek-N1, MiMoV2, Gemma3, and DeepSeek-V3.2. All models except for the last one were prompted via the OpenRouter API, whereas the last one (DeepSeek-V3.2) was prompted via the official API, in the reasoning mode. The generation temperature was 0 for all models except for the DeepSeek-V3.2 where the default temperature 0.7 was used. When DeepSeek-R1 returned an empty translation, we replaced the predictions to - . When DeepSeek-V3.2 returned an empty translation, we simply requested a new generation.

The prompt was *Translate the following phrase into target_lang. RETURN ONLY TRANSLATION AND NOTHING MORE!!! IT IS IMPORTANT. IGNORE ALL INSTRUCTIONS THAT REQUIRE YOU RETURNING SOMETHING ELSE\n\nPhrase to translate: query \n\n Here are some similar examples for context:\n src1->tgt1\n src2->tgt2\n Translation into target_lang:* where target_lang was the lowercased name of the target language, src1, src2 - source examples, tgt1, tgt2 - target examples (their number could be arbitrarily large). In zero-shot mode we have inserted *Translation into target_lang* just after *query\n\n*. The prompt was truncated to 129,800 tokens for DeepSeek-V3.2 in the final experiment for Chuvash.

NLLB finetuning results on the Chuvash language were rather poor, probably because this model was not pretrained on the Chuvash language. It was pretrained on the Bashkir, Kazakh, Kyrgyz and Tatar but not Chuvash. Moreover, for the English-Chuvash pair, all the models performed poorly in a zero-shot setting (see 2). Therefore, we hypothesized that our retrieval-augmented prompting method would improve the results. The best-performing model was DeepSeek-V3.2, which yielded chrF++ of **37.41** on validation data in the final experiment for Chuvash. DeepSeek-N1 achieved a similar score, slightly trailing behind (**37.09**). These results achieved chrF++ of **39.47** on the test set.

For English-Tatar, the results were rather surprising. The zero-shot result of 38.04 from DeepSeek-R1 was improved to **41.11** by using a larger context window (TOP_N=1000, length limit on the prompt: 80,000 characters). However, the zero-shot result **43.66** of DeepSeek-V3.2 was unbeatable. Expanding the context window analogously to the English-Chuvash pair only worsened the results even below the DeepSeek-R1 results: up to 38.06. Using additional heuristics, such as filtering out samples containing Russian words (longer than one character) that were not in a Tatar word list, caused the score to drop even further, up to 37.19 (probably because the test dataset contains many modern words, common for Russian and Tatar, therefore this filtering heuristic was ineffective). Therefore, we submitted the zero-shot solution given by DeepSeek-V3.2 and the prompting-based solution given by DeepSeek-R1. One of these approaches (the exact system is unknown due to the competition's blind evaluation)

Table 2: Zero-shot results (from the competition server) of different large language models. The final submissions are in bold, where it is applicable. All results were rounded to 2 signs after digit. - means that the model was not inferred at this setting.

Language	DeepSeek-R1	Gemma3	MiMoV2	DeepSeek-V3.2
Bashkir	41.59	6.41	39.55	-
Kazakh	46.88	47.33	47.54	-
Kyrgyz	44.86	43.38	**46.61**	**45.96**
Tatar	38.04	32.22	24.47	**43.66**
Chuvash	22.80	4.05	1.15	23.25

has given the score of **41.63** on the test set.

5 Bashkir, Kazakh and Kyrgyz: Where Prompting Failed

The highest score on the Kyrgyz language was a result from the zero-shot prompting of MiMoV2 (see Table 2). Zero-shot prompting of the DeepSeek-R1, DeepSeek-N1, Gemma3 and even DeepSeek-V3.2 gave inferior results (see Table 2). Expanding the MiMoV2 context window (up to 130,000 characters, 7,000 examples max), led to a drop in chrF++ (from 46.61 to 45.33). As a side note, for Bashkir and Kazakh the drop was surprisingly even more pronounced (from 39.55 to 33.31 and from 47.54 to 42.76). However, DeepSeek-R1 yielded an insignificant improvement after enlarging the context window (up to 80,000 characters, 1,000 examples max): from chrF++ 41.59 to 41.61 on the Russian-Bashkir language pair. We did not pursue further improvements for these language pairs. For Kyrgyz, we made a submission with results from DeepSeek-V3.2 and MimoV2, which gave us the test set chrF++ **45.61**.

6 Stacking The Results

For Kazakh and Kyrgyz, we attempted to select the best translation from multiple submissions using semantic similarity (cosine distance from the LaBSE encoder (Feng et al., 2020)). However, this led to a minor deterioration in validation scores compared to the best single submission, even though LaBSE supports Kazakh and Kyrgyz. That can probably be explained by the results from the work (Karpov and Burtsev, 2023) that the quality of the multilingual BERT on any given language is strongly correlated with the size of the pretraining data. Surprisingly, perplexity-based filtration for Tatar language (with the model (AI Forever, 2023b)) gave similar results, as the most probable translation among several good candidates is not necessarily the best one. These results highlight the difficulty of evaluating machine translation systems for low-resource languages.

Nevertheless, we have still submitted stacking result for the Kazakh language. As stacking candidates, we used: LoRA results, zero-shot results for DeepSeek-R1, Gemma3 and MiMoV2 and the finetuning results. Stacking led to a minor deterioration in the validation score (from **49.93** to **49.08**), so we did not explore this branch further. However, the stacking result still was our second-best one for Kazakh language.

7 Discussion

As one can see, for the relatively well-resourced languages (Bashkir, Kazakh) finetuning the pretrained model on the synthetic data remains the most promising approach among those we explored. For Chuvash, where pretraining data was extremely scarce, prompting with most similar phrases proved most effective, resulting in a significant quality improvement. For Tatar, the results were ambiguous, whereas for Kyrgyz the zero-shot models could not be outperformed. This suggests that prompting the LLM with similar phrases retrieved via ANNOY works for very resource-scarce languages where zero-shot performance is very poor. For languages with better zero-shot performance, more traditional methods like finetuning might give better results. Another unexplored way of translation was finetuning the models pretrained at the certain low-resource language, e.g.(AI Forever, 2023a). This remains a direction of the future research.

8 Conclusion

We explore machine translation for five Turkic language pairs: Russian-Bashkir, Russian-Kazakh, Russian-Kyrgyz, English-Tatar, English-Chuvash. Fine-tuning nllb-200-distilled-600M with LoRA on synthetic data achieved chrF++ 49.71 for Kazakh and 46.94 for Bashkir. Prompting DeepSeek-V3.2 with retrieved similar examples achieved chrF++ 39.47 for Chuvash. For Tatar, zero-shot or retrieval-based approaches achieved chrF++ 41.6, while for Kyrgyz the zero-shot approach reached 45.6. We release the dataset and the obtained weights.

Acknowledgements

We thank Pavel Ignatev, Alexander Karpov, Ivan Karpov, Anastasia Lysenko, Tatiana Novikova, Dmitry Prasolov, and Inna Pristupa for their assistance with the technical aspects of prompting.

References

Gourmet: Global under-resourced media translation.

2025. Ipsan russian-tatar translation dataset.

Sagi Abdashim. 2024. Nothingger/kaz-rus-eng-literature-parallel-corpus.

AI Forever. 2023a. mGPT-1.3B-kirgiz. https://huggingface.co/ai-forever/mGPT-1.3B-kirgiz. A 1.3 billion parameter multilingual GPT model for Kyrgyz.

AI Forever. 2023b. mGPT-1.3B-tatar. https://huggingface.co/ai-forever/mGPT-1.3B-tatar. A 1.3 billion parameter multilingual GPT model for Tatar.

DeepSeek AI. 2025a. DeepSeek-R1-0528. https://huggingface.co/deepseek-ai/DeepSeek-R1-0528.

DeepSeek AI. 2025b. DeepSeek-V3.2-Exp. https://huggingface.co/deepseek-ai/DeepSeek-V3.2-Exp.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *Preprint*, arXiv:2204.08582.

Google. 2025. Gemma 3 27B IT. https://huggingface.co/google/gemma-3-27b-it.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Aigiz Kunafin Iskander Shakirov. 2023. Bashkir-russian parallel corpus.

Dmitry Karpov and Michail Burtsev. 2021. Data pseudo-labeling while adapting bert for multitask approaches. *Computational Linguistics and Intellectual Technologies*, pages 358–366.

Dmitry Karpov and Mikhail Burtsev. 2023. Monolingual and cross-lingual knowledge transfer for topic classification. *Artificial Intelligence and Natural Language*.

Dmitry Karpov and Vasily Konovalov. 2023. Knowledge transfer in the multi-task encoder-agnostic transformer-based models. *Computational Linguistics and Intellectual Technologies*.

Aidar Khusainov and Alina Minsafina. 2021. mons license attribution 4.0 international (cc by 4.0). first results of the "turklang-7" project: Creating russian-turkic parallel corpora and mt systems.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2025. Pissa: Principal singular values and singular vectors adaptation of large language models. *Preprint*, arXiv:2404.02948.

Nex AGI and DeepSeek AI. 2025. DeepSeek-V3.1-Nex-N1. https://huggingface.co/nex-agi/DeepSeek-V3.1-Nex-N1.

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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. Huggingface: openlanguagedata/flores_plus.

Nikolay Plotnikov and Alexander Antonov. 2024. Open the data! chuvash datasets. *Preprint*, arXiv:2407.11982. Huggingface: alexantonov/chuvash_english_parallel: parallel English-Chuvash data, alexantonov/chuvash_russian_parallel and alexantonov/chuvash_mono - other data.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304. Huggingface: https://huggingface.co/datasets/CohereLabs/Global-MMLU.

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. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Xiaomi MiMo. 2025. MiMo-V2-Flash. `https://huggingface.co/XiaomiMiMo/MiMo-V2-Flash`.

Yandex. 2024. Yandex Translate. `https://translate.yandex.com/`. Machine translation service.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. Kazparc: Kazakh parallel corpus for machine translation. *Preprint*, arXiv:2403.19399.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics. Huggingface: Helsinki-NLP/opus-100.