Script Correction and Synthetic Pivoting: Adapting Tencent HY-MT for Low-Resource Turkic Translation

Bolgov Maxim
Independent Researcher
Moscow, Russia
bolgov1458@yandex.ru

Abstract

This paper describes a submission to the LoResMT 2026 Shared Task for the Russian-Kazakh, Russian-Bashkir, and English-Chuvash tracks. The primary approach involves parameter-efficient fine-tuning (LoRA) of the Tencent HY-MT1.5-7B multilingual model. For the Russian-Kazakh and Russian-Bashkir pairs, LoRA adaptation was employed to correct the model's default Arabic script output to Cyrillic. For the extremely low-resource English-Chuvash pair, two strategies were compared: mixed training on authentic English-Chuvash and Russian-Chuvash data versus training exclusively on a synthetic English-Chuvash corpus created via pivoting through Russian. Baseline systems included NLLB 1.3B (distilled) for Russian-Kazakh and Russian-Bashkir, and Gemma 2 3B for English-Chuvash. Results demonstrate that adapting a strong multilingual backbone with LoRA yields significant improvements over baselines while successfully addressing script mismatch challenges. Code for training and inference is released at: https://github.com/defdet/low-resource-langs-mt-adapt

1 Introduction

Low-resource machine translation remains a critical challenge, particularly for agglutinative languages with complex morphology such as those in the Turkic family (Mirzakhalov et al., 2021). The LoResMT 2026 Shared Task included five translation tracks for low-resource Turkic languages; this work focuses on three of them: Russian-Kazakh (Ru-Kk), Russian-Bashkir (Ru-Ba), and English-Chuvash (En-Cv).

While large multilingual models have demonstrated strong performance on major languages, they often exhibit systematic issues for low-resource Turkic varieties (Zoph et al., 2016; Team et al., 2022). Notably, the Tencent HY-MT1.5-7B model (Zheng et al., 2025), despite its strong multilingual capabilities, outputs translations for Turkic languages exclusively in Arabic script rather than the Cyrillic orthographies used in contemporary Central Asian contexts. This work explores whether Low-Rank Adaptation (LoRA) (Hu et al., 2021) is sufficient to both transfer knowledge to previously unseen language pairs and correct script mismatches without external normalizers.

For the English-Chuvash pair, where direct parallel data is scarce and partially synthetic in origin, the efficacy of synthetic pivoting versus mixed training was evaluated. Synthetic pivoting involves translating the source side of a high-resource pivot language corpus (Russian-Chuvash) into the desired source language (English) using a strong pivot model, then training directly on the resulting synthetic parallel data. This offline data generation approach avoids the compounding errors inherent in runtime cascade translation (En→Ru→Cv), where translation errors from the first stage (En→Ru) propagate and amplify in the second stage (Ru→Cv). By generating synthetic parallel data offline, the model learns a direct En→Cv mapping on a consistent, albeit synthetic, training distribution (Caswell and Bapna, 2022; Elmadani and Buys, 2024).

2 System Description

2.1 Base Model and Adaptation Strategy

The `tencent/HY-MT1.5-7B` model (Zheng et al., 2025) served as the backbone for all experiments. This is a decoder-only transformer model with 7 billion parameters, pretrained on large-scale multilingual parallel data covering over 33 languages. The model underwent supervised fine-tuning on translation tasks followed by Group Relative Policy Optimization (GRPO) reinforcement learning to improve translation quality.

Despite its strong multilingual capabilities, the model outputs translations for Central Asian Turkic

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 217–221

languages exclusively in Arabic script rather than the Cyrillic orthographies used in contemporary contexts. To address this script mismatch and adapt to previously unseen language pairs, we employed Low-Rank Adaptation (LoRA) rather than full fine-tuning to maintain computational efficiency and avoid catastrophic forgetting.

LoRA projections targeted the attention mechanism (q, k, v, o layers) with rank $r = 16$. This introduces approximately 13.6 million trainable parameters, allowing the model to adjust its internal representations for Chuvash, Bashkir, and Kazakh syntax while retaining its broad multilingual knowledge base (Hu et al., 2021).

2.2 Script Adaptation via LoRA

Rather than developing an external rule-based normalizer or character mapper to convert Arabic output to Cyrillic, the script adaptation was handled entirely through LoRA fine-tuning on Cyrillic training data. This approach allows the model to learn the script preference through gradient updates rather than requiring explicit post-processing rules.

After two epochs of adaptation on Cyrillic-script parallel data, the model successfully suppressed Arabic token generation, producing exclusively Cyrillic output for all Turkic target languages. Manual inspection of generated outputs confirmed zero instances of Arabic characters in the adapted model's translations.

2.3 Synthetic Pivoting for Chuvash

For the English-Chuvash track, two data strategies were explored:

- **Mixed Training:** Combining authentic En-Cv data (upsampled) with Russian-Chuvash data to leverage multilingual transfer (Nguyen and Chiang, 2017).
- **Synthetic Pivoting:** Generating a purely synthetic En-Cv dataset by translating the Russian source side of the Ru-Cv corpus into English using `facebook/wmt19-ru-en` (Ng et al., 2019), then training exclusively on the resulting synthetic En-Cv pairs.

The synthetic pivot approach was motivated by the desire to avoid runtime error compounding that occurs in cascade systems (En→Ru→Cv), where translation errors accumulate at each stage. By generating synthetic parallel data offline, the model learns a direct En→Cv mapping on a consistent, albeit synthetic, training distribution (Elmadani and Buys, 2024).

3 Experimental Setup

3.1 Data

Openly available datasets were utilized for all tracks:

- **Ru-Kazakh:** ISSAI KazParc corpus (Yeshpanov et al., 2024), with additional English-Kazakh data included for cross-lingual transfer.
- **Ru-Bashkir:** AigizK Bashkir-Russian parallel corpus (Shakirov and Kunafin, 2023).
- **English-Chuvash:** We utilized two datasets provided by `alexantonov` on Hugging Face: the `chuvash_english_parallel` corpus (200k sentence pairs sourced from books with assistance of MT) and the `chuvash_russian_parallel` corpus (1.4M manually collected samples) (Plotnikov and Antonov, 2024).

3.2 Training Configuration

Training was conducted using the Hugging Face Transformers `Trainer` with manual adaptations for Supervised Fine-Tuning (SFT), excluding prompts from label loss calculations. Hardware consisted of 4 NVIDIA A100 (80GB) GPUs. Key hyperparameters are specified in Table 1.

Hyperparameter	Value
Epochs	2
Batch size (train)	4 per device
Batch size (eval)	8 per device
Learning rate	1×10^{-4}
LR scheduler	Cosine decay
Warmup ratio	0.05
Optimizer	AdamW
Weight decay	0.01
Max gradient norm	1.0
LoRA rank	16
LoRA targets	q, k, v, o

Table 1: Training hyperparameters.

3.3 Evaluation and Decoding

Beam search decoding was employed for all inference tasks. For pivot translations (Ru→En)

used to generate the synthetic English-Chuvash corpus, `facebook/wmt19-ru-en` (Ng et al., 2019) was used with beam size 3 to balance translation quality and computational cost, as the Russian-Chuvash dataset is substantial (1.4M sentence pairs). Final submission generation used beam size 5 to maximize translation quality on the test set. The evaluation metric was chrF++, as chosen by the organizers.

4 Results

Adapted models were compared against strong baselines to assess the effectiveness of LoRA-based adaptation for low-resource Turkic translation. For Russian-Kazakh and Russian-Bashkir, we used NLLB 1.3B (distilled) (Team et al., 2022) as the baseline. For English-Chuvash, we compared against Gemma 2 3B (Team et al., 2024), fine-tuned on the authentic English-Chuvash dataset on the same hardware. The results in Table 2 demonstrate substantial improvements across all three language pairs. For Russian-Kazakh, the LoRA-adapted model achieved a 32-point chrF++ gain over NLLB, suggesting that the base model's multilingual representations transfer effectively to this pair despite the script mismatch. The Russian-Bashkir track showed a 24-point improvement, with the adapted model successfully learning Bashkir morphology and Cyrillic orthographic conventions from the parallel data. The English-Chuvash results, while showing a smaller absolute gain of 12.8 chrF++ points over the Gemma baseline, are notable given the extreme scarcity and partially synthetic nature of the training data.

Pair	Baseline	Base	Our score
Ru-Kk	NLLB 1.3B (dist.)	16.0	**48.0**
Ru-Ba	NLLB 1.3B (dist.)	28.0	**52.0**
En-Ch	Gemma 2 3B	23.0	**35.8**

Table 2: chrF++ scores on the public test set. "Our" refers to the LoRA-adapted Tencent model.

4.1 Chuvash Data Strategy Comparison

For English-Chuvash, two training strategies were compared against the baseline. Results are shown in Table 3.

We conducted a manual qualitative analysis of approximately 50 randomly sampled translations. Translations were assessed with the assistance of Gemini 3 Pro, Google Translate, and Yandex Translate for semantic verification.

Model	Training Data	chrF++
Gemma 2 3B	En-Cv + Ru-Cv (upsampled)	23.0
HY-MT1.5-7B	En-Cv + Ru-Cv (upsampled)	34.4
HY-MT1.5-7B	Synthetic En-Cv only	**35.8**

Table 3: Comparison of data strategies for English-Chuvash translation.

The analysis revealed that the mixed-training model exhibited systematic factual errors stemming from domain mismatch between the literary book-sourced English-Chuvash data and the more diverse Russian-Chuvash corpus. The synthetic pivot model, trained on general-domain Russian data translated into English, demonstrated more consistent handling of everyday and technical terminology. Table 4 presents representative examples.

These examples illustrate that while both models struggle with rare medical and legal terminology, the synthetic pivot system consistently produces semantically closer approximations. The mixed model's errors often stem from overfitting to the literary register of the book corpus, where metaphorical language (e.g., "red" for ginger) is common but inappropriate for factual translation tasks.

5 Conclusion

Parameter-efficient adaptation of a large multilingual model successfully addressed low-resource Turkic translation tasks, achieving substantial gains over NLLB and Gemma baselines. LoRA fine-tuning proved sufficient for both knowledge transfer and script correction, with the adapted models producing exclusively Cyrillic output. For extremely rare pairs like English-Chuvash, synthetic pivoting outperformed mixed training by providing a consistent direct mapping while avoiding the error compounding typical for runtime cascade systems (Elmadani and Buys, 2024).

6 Limitations

Dependence on Pivot Quality The synthetic Chuvash approach relies heavily on the quality of the Ru→En pivot translation. The `wmt19-ru-en` model may not be optimal for the chosen Chuvash dataset. Commercial MT systems designed for literal translation may yield better synthetic corpora.

English	Mixed Training	Synthetic Pivot	Analysis
Ginger root	Хĕрлĕ ӳсен-тăран (red plant)	Имбирь тымарĕ (ginger root)	Correct translation. Mixed model associates "ginger" with color.
Lettuce	Çĕрулми (potato)	Салат (lettuce)	Correct translation. Mixed model confuses vegetables.
Invalid ideas	Инвалидла шухăшсене (disabled thoughts)	Ниме юрăхсăр шухăшсене (worthless thoughts)	Contextually correct vs. 'false friend' error.
Bladder stones	Шăпăр шăтăкĕнчи (in broom hole)	Сечечкăсенче чулсем (stones in sections)	Partial hallucination in both, but synthetic captures core medical term.
Cure nausea	Ытлашши çиекен çынсене (for people who eat too much)	Çăмăллăха тӳрлетме пултарать (can restore ease)	Synthetic attempts semantic equivalent; mixed produces unrelated phrase.
Liability claims	Тĕрĕслев (checking)	Ответ тытасси (accountability)	Synthetic preserves legal concept; mixed oversimplifies.

Table 4: Qualitative comparison of translation outputs.

Suboptimal Data Mixing Training exclusively on synthetic En-Cv data, while effective, represented a missed opportunity. Since the Tencent backbone is fluent in both Russian and English, including the high-quality manually curated Ru-Cv corpus alongside synthetic En-Cv data could have provided a "correctness anchor" and regularization, potentially reducing factual errors.

Script Adaptation Data Requirements The LoRA-based script adaptation approach requires sufficient high-quality parallel data in the target script and a strong multilingual foundation model. For languages with extremely limited digital presence, a dedicated normalization layer would be more data-efficient. In scenarios with fewer than several thousand parallel sentences, external normalizers and character mapping systems may be the only viable option for script conversion.

References

Isaac Caswell and Ankur Bapna. 2022. Unlocking zero-resource machine translation to support new languages in google translate. Google Research Blog.

Khalid N. Elmadani and Jan Buys. 2024. Neural machine translation between low-resource languages with synthetic pivoting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12144–12158, Torino, Italia. ELRA and ICCL.

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

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*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. 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 (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nikolay Plotnikov and Alexander Antonov. 2024. Open the data! chuvash datasets. *Preprint*, arXiv:2407.11982.

Iskander Shakirov and Aigiz Kunafin. 2023. Bashkir-russian parallel corpora. Hugging Face dataset.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving

open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. KazParC: Kazakh parallel corpus for machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9633–9644, Torino, Italia. ELRA and ICCL.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025. Hy-mt1.5 technical report. *Preprint*, arXiv:2512.24092.

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*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.