Navigating Data Scarcity in Low-Resource English-Tatar Translation using LLM Fine-Tuning

Ahmed Khaled Khamis
Georgia Institute of Technology
akhamis6@gatech.edu

Abstract

The scarcity of high-quality parallel corpora remains the primary bottleneck for English-Tatar machine translation. While the OPUS project provides various datasets, our tests reveal that datasets like WikiMatrix, GNOME, and NLLB, suffer from significant noise and incorrect labeling, making them unsuitable for training robust encoder-decoder translation models that typically requires larger amount of high quality data. Furthermore, we demonstrate that small-scale multilingual Large Language Models (LLMs), such as Qwen3 (4B-30B), Gemma3 (4B-12B) and others, show severe "Turkish interference", and they frequently hallucinate Turkish vocabulary when prompted for Tatar. In this paper, we navigate this data scarcity by leveraging Llama 3.3 70B Instruct, which is the only model in our zero-shot benchmarks capable of maintaining distinct linguistic boundaries for Tatar. To address the lack of gold-standard data, we curated a synthetic dataset of 7,995 high-quality translation pairs using a frontier model as a teacher. We then performed 4-bit LoRA fine-tuning to train Llama for English-Tatar translation. Our results show a performance leap: while fine-tuning on the limited Tatoeba dataset (1,193 samples) yielded a CHRF++ score of 24.38, while fine-tuning on our synthetic dataset achieved 32.02 on the LoResMT 2026 shared task test set. We release our curated dataset and fine-tuned models to support further research in low-resource Turkic machine translation.

1 Introduction

Neural Machine Translation (NMT) has been improving by the success of massive datasets and high-parameter architectures. However, for low-resource languages such as Tatar, progress remains constrained by a deep and persistent data scarcity gap. While open source projects like the OPUS corpus (Tiedemann and Thottingal, 2020) offer several English-Tatar datasets, their utility is limited due to poor alignment, incorrect labels, and significant linguistic noise. Our investigation into the *WikiMatrix*, *GNOME*, and *NLLB* corpora revealed that a substantial portion of these datasets are practically unusable for translation tasks. In the face of such scarcity, traditional encoder-decoder architectures like *MarianMT* (Junczys-Dowmunt et al., 2018) struggle to generalize, often collapsing into repetitive or nonsensical outputs.

Simultaneously, the rise of multilingual Large Language Models (LLMs) promised a new era of zero-shot translation. Yet, our benchmarking of modern instruction-tuned models—including Qwen3 (4B-30B) (Yang et al., 2025), Gemma3 (4B-12B) (Team et al., 2025), and Llama 3 (3B-8B) (Grattafiori et al., 2024), showed a critical failure mode: "Turkic interference" Despite their multilingual training, these smaller models frequently confuse Tatar with Turkish, outputting Turkish vocabulary and morphological structures when prompted for Tatar. This suggests that smaller parameter counts may be insufficient to maintain the nuanced linguistic boundaries required for low-resource languages.

This paper details our approach[1] for the ***LoResMT*** 2026 English-Tatar Shared Task. We demonstrate that navigating data scarcity requires both: Model Scale and Synthetic Curation. During our zero-shot evaluations, we identified that *Llama 3.3 70B Instruct* was a model that's capable of generating coherent Tatar without significant Turkish bias. However, even at this scale, the 1,193 available samples from the Tatoeba corpus proved insufficient for competitive performance. To bridge this gap, we used DeepSeek-R1 (DeepSeek-AI et al., 2025) to curate a high-quality synthetic dataset of 7,995 translation pairs. By fine-tuning the 70B parameter model on this curated data using the on an

[1] Code: https://github.com/KickItLikeShika/llm-loresmt

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 198–202
March 28, 2026 ©2026 Association for Computational Linguistics

NVIDIA H100, we achieve a CHRF++ score of 32.02, outperforming our Tatoeba-only baseline of 24.38.

2 The Challenge of Data Scarcity

The primary obstacle in English-Tatar translation is not just the quantity of data, but lack of high-fidelity, correctly labeled corpora. This section details our evaluation of available resources and the failure of zero-shot models to address this gap.

2.1 Evaluation of Existing Corpora

We conducted a comprehensive review of the English-Tatar datasets available via the OPUS corpus (Tiedemann and Thottingal, 2020), including NLLB (Team et al., 2022), WikiMatrix (Schwenk et al., 2019), GNOME (Deshpande et al., 2024), XLEnt (El-Kishky et al., 2021), and QED (Lamm et al., 2020). Our qualitative analysis revealed that many of these datasets are unsuitable for training high-quality translation models. The WikiMatrix and NLLB corpora, for instance, exhibited significant noise where Russian or Turkish segments were incorrectly labeled as Tatar. In the GNOME and XLEnt datasets, we observed a high frequency of "hallucinated" labels—where the source and target segments were semantically unrelated. Training on such corpora led to models that to completely fail to generate coherent Tatar syntax.

2.2 The Tatoeba Baseline

Among the open-source datasets, only the Tatoeba (Tiedemann, 2020) corpus provided a degree of reliable alignment. However, after aggressive filtering for duplicates, empty segments, and language identification, we were left with only 1,193 high-quality sentence pairs. While useful as a starting point, a dataset of roughly 1,200 samples is insufficient for training a robust encoder-decoder model from scratch or for effectively adapting a decoder-only LLM.

2.3 Zero-Shot Benchmarking and Turkic Interference

Given the data scarcity, we explored the zero-shot capabilities of several modern, instruction-tuned Large Language Models. We tested models across various scales, including Qwen3 (4B, 8B, 14B, 30B) (Yang et al., 2025), Gemma3 (4B, 12B, 27B) (Team et al., 2025), Llama 3.2 3B, and Llama 3 8B (Grattafiori et al., 2024).

Despite their general multilingual proficiency, all models within this parameter range exhibited a specific failure mode that we called "Turkic Interference" when prompted to translate into Tatar, these models consistently struggled to differentiate between Tatar and its high-resource relative, Turkish. Common errors included:

- Lexical and Morphological failures: Using Turkish words instead of Tatar equivalents.
- Instruction failure: Models frequently generated long explanations in English rather than the requested translation or repeated the same word multiple times.

Llama 3.3 70B Instruct was the only model in our evaluation that demonstrated a foundational ability to maintain Tatar's linguistic identity, serving as the necessary baseline for our fine-tuning experiments.

3 Methodology

To navigate the extreme data scarcity of the English-Tatar pair, we curated a synthetic dataset. Our methodology centers on using a high-parameter frontier model to bootstrap a high-quality corpus, followed by parameter-efficient fine-tuning (Xu et al., 2023) of a 70B parameter decoder-only model.

3.1 Synthetic Data Curation

Given the unreliability of existing web-scraped corpora, we utilized DeepSeek for dataset generation. We curated a dataset consisting of 7,995 high-quality translation pairs. Unlike the noisy OPUS datasets, these pairs were generated through structured prompting designed to ensure grammatical correctness in Tatar.

The structural properties of the resulting dataset (Figure 1) provide insight about the linguistic relationship between the two languages. English sentences have a mean length of 14.99 words, while Tatar translations average 10.78 words, while the character counts remain nearly identical (75.21 for English vs. 74.74 for Tatar).

3.2 Model Selection and Architecture

Based on our zero-shot benchmarks, we selected Llama 3.3 70B Instruct as our base model. While 8B parameter models exhibited significant linguistic "bleeding" from high-resource Turkic languages like Turkish, the 70B parameter scale provided the

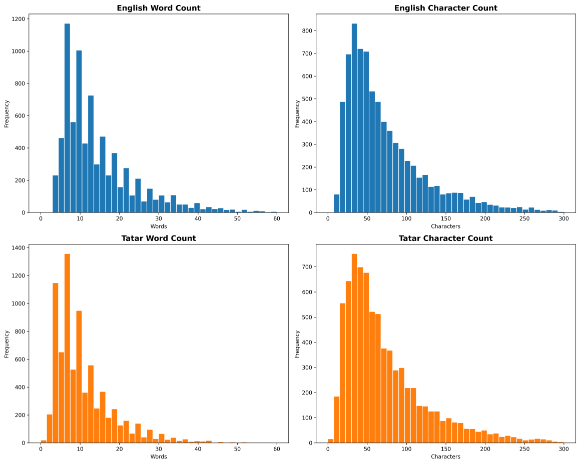

Figure 1: Linguistic profile of the curated synthetic dataset. The top row displays English word and character counts; the bottom row displays the corresponding Tatar counts.

necessary internal representation to maintain Tatar-specific syntax and vocabulary. To make training feasible, we used 4-bit quantization. This allowed us to load the 70B parameter weights into GPU memory with minimal impact on perplexity, providing a foundation for subsequent Parameter-Efficient Fine-Tuning (PEFT) (Xu et al., 2023).

3.3 Training Configuration

We utilized the Unsloth framework (Daniel Han and team, 2023) to perform LoRA (Hu et al., 2021) fine-tuning. This framework provides optimized CUDA kernels that significantly reduce VRAM consumption and increase training speed, enabling us to fine-tune a 70B model on a single node efficiently. Our training was conducted with the following hyperparameters: **LoRA Parameters**: We targeted all linear modules (including *q_proj, k_proj, v_proj, o_proj*, and MLP layers) with a Rank (r) of 64 and an Alpha of 64. This relatively high rank was chosen to maximize the model's capacity to adapt to Tatar's specific morphological requirements.

Optimization Strategy: We employed the *train_on_responses_only* technique. By masking the loss for the instruction and source English text, we ensured the gradient updates were computed exclusively on the Tatar translation output.

Hyperparameters: The model was trained for 2 epochs with a learning rate of 2e-4 and a linear scheduler. We used a global batch size of 16 (8 per device with 2 gradient accumulation steps) and a weight decay of 0.001 to prevent overfitting on the synthetic samples.

4 Experimental Results

4.1 Quantitative Performance

We evaluated our models on the official LoResMT 2026 English-Tatar shared task test set using the CHRF++ metric. Our results (Table 1) demonstrate the clear advantage of synthetic augmentation.

Model	Training Data	CHRF++
Llama 3.3 70B	Tatoeba	24.3842
Llama 3.3 70B	**Synthetic**	**32.0251**

Table 1: Translation performance comparison on the LoResMT 2026 shared task test set.

The transition from the tiny Tatoeba corpus to our curated synthetic dataset resulted in a 7.64 point jump in CHRF++. This improvement suggests that the model effectively internalized the distinct Tatar syntax and vocabulary, successfully overcoming the "Turkic interference" observed in zero-shot baselines.

4.2 Training Environment

All experiments were conducted on a single NVIDIA H100 (94GB) GPU using the `Unsloth` framework for 4-bit LoRA fine-tuning. This setup enabled us to train the 70B parameter model with a total batch size of 16 in approximately 1.5 hours. The efficiency of this pipeline demonstrates that high-parameter models can be adapted to low-resource tasks with limited computational overhead if the data quality is sufficiently high.

5 Discussion and Error Analysis

The significant performance gains observed in our experiments underscore two critical factors in low-resource machine translation for Turkic languages: the necessity of high-parameter model scales and the role of synthetic supervision in decoupling linguistic interference.

5.1 Linguistic Interference and Model Scale

A primary challenge identified in our zero-shot benchmarks was the "Turkic interference" phenomenon, where models frequently defaulted to Turkish (tr) vocabulary and morphology when prompted for Tatar (tt). We hypothesize that smaller models (e.g., 3B to 30B parameters) possess an internal representation that is insufficient to maintain distinct boundaries between closely related languages within the same family. In these models, the high-resource presence of Turkish in the pre-training data caused the model to prioritize Turkish tokens over Tatar counterparts.

Our results suggest that the 70B parameter scale of Llama 3.3 is a critical threshold for English-Tatar translation. The larger parameter count appears to provide a more granular latent space, allowing the model to isolate and preserve Tatar-specific features, even under extreme data scarcity. Fine-tuning on our curated synthetic dataset further reinforced these boundaries, effectively "teaching" the model to resist the Turkish default.

5.2 Qualitative Analysis

A qualitative review of the outputs from the Tatoeba-baseline and the synthetic-augmented model reveals several key improvements. Models fine-tuned only on the limited Tatoeba corpus often struggled with Tatar's complex sturcture, occasionally producing "broken" words.

In contrast, the synthetic-augmented model demonstrated a better command of Tatar morphology. For example, in translating complex temporal or locative phrases, the model correctly utilized Tatar-specific markers rather than the more common Turkish equivalents found in zero-shot outputs. Furthermore, the fine-tuned model strictly adhered to the "no-explanation" instruction, whereas zero-shot models frequently included English commentary text.

5.3 The Value of Synthetic Supervision

Our findings demonstrate that for low-resource languages, the *quality* and *purity* of the training data are more important than data volume. By using a frontier model to generate a curated dataset, we were able to provide the 70B model with a "clean" signal of what constitutes correct Tatar. This approach successfully bypassed the noise inherent in web-scraped corpora like WikiMatrix or GNOME, which our analysis showed multiple mislabeled Turkish or Russian data. The 7.64 CHRF++ jump proves that synthetic data from a superior LLM can serve as a high-fidelity surrogate for native-speaker data in extreme scarcity scenarios.

6 Conclusion

In this work, we addressed the data scarcity in English-Tatar machine translation by transitioning from traditional web-scraped corpora to a high-quality synthetic curation strategy. Our investigation revealed that existing large-scale datasets for Tatar often contain significant linguistic noise, while small-scale multilingual LLMs frequently suffer from Turkic interference, failing to distinguish Tatar from higher-resource relatives like Turkish.

By leveraging the Llama 3.3 70B parameter model and a curated synthetic dataset of 7,995 translation pairs, we achieved a robust performance benchmark, with CHRF++ score of 32.02 on the LoResMT 2026 shared task. Our results demonstrate that model scale is critical for preserving the linguistic integrity of low-resource languages within dense language families. Furthermore, we provide evidence that synthetic data generated by frontier models can serve as a high-fidelity training signal, successfully bypassing the limitations of noisy, web-scraped corpora.

References

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Darshan Deshpande, Shambhavi Sinha, Anirudh Ravi Kumar, Debaditya Pal, and Jonathan May. 2024. Gnome: Generating negotiations through open-domain mapping of exchanges. *Preprint*, arXiv:2406.10764.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. *Preprint*, arXiv:2104.08597.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. *Preprint*, arXiv:1804.00344.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering. *Preprint*, arXiv:2009.06354.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *Preprint*, arXiv:1907.05791.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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 and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. *Preprint*, arXiv:2010.06354.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.