Small but Mighty: Building State-of-the-Art Machine Translation Systems for Finno-Ugric and Turkic Low-Resource Languages with 55 Million Parameters

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) and massively multilingual systems show impressive capabilities, they systematically underperform on morphologically complex, extreme low-resource languages. In this paper, we demonstrate that massive parameter scale is not a prerequisite for State-of-the-Art (SOTA) translation. We introduce `MultiMT`, a suite of highly optimized Neural Machine Translation (NMT) models utilizing only 55 million parameters, tailored for 15 underrepresented Finno-Ugric and Turkic languages. Through exhaustive empirical analysis, we uncover a fundamental directional asymmetry in cross-lingual transfer: high-resource anchor languages significantly improve source comprehension ($xx \rightarrow ru$) but induce severe capacity bottlenecks during target generation ($ru \rightarrow xx$). By restricting target-side multilinguality and iteratively injecting synthetic data, our compact models achieve new SOTA results in $ru \rightarrow xx$ translation, drastically outperforming massive proprietary LLMs (e.g., DeepSeek-V4, Gemini-3). We open-source our tokenizers and models to establish a rigorous low-resource baseline.

1 Introduction

Massively multilingual machine translation (MMT) systems (e.g., NLLB-200 (Team et al., 2022), MADLAD-400 (Kudugunta et al., 2024), OmniMT (Team et al., 2026)) and General-Purpose Large Language Models (LLMs) such as Gemini-3 and DeepSeek-V4 rely on cross-lingual transfer and billions of parameters to overcome data scarcity. However, a significant translation quality gap remains for extreme low-resource, typologically distinct languages. For the highly agglutinative Finno-Ugric and Turkic languages of Russia, massive general-purpose models systematically fail due to sub-optimal English-centric tokenizers, pre-training data starvation, and the "Curse of Multilinguality" - where limited capacity causes the catastrophic forgetting of underrepresented linguistic structures. This leaves researchers and regional language activists with a critical hardware and modeling dilemma.

In this paper, we challenge the assumption that massive scale is necessary for extreme low-resource translation. We hypothesize that carefully constructed, linguistically-driven NMT systems with fractional parameter counts can outperform SOTA foundation models by avoiding parameter dilution and structurally aligning with target morphology. To prove this, we develop `MultiMT`: a suite of specialized, 55-million-parameter Transformer models covering 15 Finno-Ugric and Turkic languages paired with Russian.

Through systematic evaluation across linguistic grouping, scaling laws, and synthetic data generation, our main contributions are three-fold: **(1) Directional Asymmetry of Cross-Lingual Transfer:** We empirically prove that cross-lingual transfer behaves asymmetrically. While including high-resource related languages acts as a powerful structural anchor for translating *to* Russian, it causes severe negative interference and capacity bottlenecks when generating low-resource languages *from* Russian. **(2) Typological Scaling and Asymmetric Architectures:** We document how data scaling applies uniquely to agglutinative models. Preliminary tests on asymmetric architectures suggest that structurally adjusting encoder depth can boost language understanding for heterogeneous language families. **(3) The "David vs. Goliath" Benchmark:** We conduct the first comprehensive benchmark comparing tailored NMT models against modern LLMs for 15 regional languages of Russia. Our 55M-parameter models achieve SOTA performance for generative $ru \rightarrow xx$ translation, significantly outperforming models thousands of times their size.

To accelerate future NLP research and reproducibility, we publish all `MultiMT` model configurations, weights, and tokenizer to HuggingFace.

2 Previous Research

The paradigm of machine translation has heavily shifted towards massively multilingual neural machine translation models and Large Language Models. Foundational works such as NLLB, MADLAD-400, and OmniMT demonstrated that scaling parameters and parallel data across hundreds of languages simultaneously yields strong baseline translations. Currently, general-purpose LLMs are also showing highly competitive zero-shot and few-shot translation capabilities on global benchmarks (Kocmi et al., 2023). However, despite vast parameter counts, these massively multilingual systems frequently underperform on extreme low-resource and morphologically complex indigenous languages due to tokenizer misalignment and training data imbalance (Ahia et al., 2023).

To combat the data scarcity of underrepresented languages, researchers heavily rely on cross-lingual transfer. It has been proven that leveraging related high-resource languages—and careful proxy language selection—can act as an anchor for low-resource translation (Team et al., 2022; Kudugunta et al., 2024; Team et al., 2026), even for language isolates as was shown on Nivkh by Izmailova et al. (2026). However, aggressive multilinguality often introduces a capacity bottleneck, causing negative interference (Pfeiffer et al., 2022).

Recent literature has narrowed the focus toward building specialized, family-centric NLP systems. For Finno-Ugric languages of Russia, Yankovskaya et al. (2023) provided a foundational evaluation of back-translation (Sennrich et al., 2016) and multilingual grouping. In the Turkic domain, Syzdykov et al. (2026) introduced TurkicNLP, standardizing toolkits and transliteration for several related languages. In line with our work, Karpov et al. (2026) investigated MT pipelines adapted for Bashkir, Kazakh, Kyrgyz, Tatar, and Chuvash, demonstrating that custom-made synthetic data generation drastically outperforms baseline LLM zero-shots.

Building upon these works, our paper provides the first comprehensive, side-by-side empirical investigation of typological scaling laws, influence of synthetically-generated data on low-resource translation quality, and SOTA LLM comparisons across Finno-Ugric and Turkic languages simultaneously.

3 Languages and Data

3.1 Training Data

We utilize publicly available parallel and monolingual texts sourced from HuggingFace (e.g. Fineweb-2 (Penedo et al., 2025)), community-driven Kod-odin[1] repository, digitized dictionaries [2], RuWiki [3], and regional websites for 15 low-resource languages: Finno-Ugric (Meadow Mari, Hill Mari, Udmurt, Komi, Erzya, Moksha, Mansi) and Turkic (Tatar, Bashkir, Chuvash, Yakut, Tyvan, Karachay-Balkar, Nogai, Khakass), supported by five related high-resource languages (Uzbek, Azerbaijani, Kyrgyz, Estonian, Hungarian). All texts are paired with Russian as the pivot language. Data statistics are summarized in Table 1.

To resolve script inconsistencies prevalent in underrepresented languages (e.g., Mansi utilizing Private Use Area unicode, Chuvash using diacritic marks from Latin alphabets), we strictly normalized all corpora prior to training and evaluation by resolving Cyrillic/Latin homoglyph collisions.

3.2 Evaluation Corpora

In order to evaluate translation quality across the selected 15 languages, we examined existing benchmarks. One of the most common evaluation datasets is Flores+ (Team et al., 2022), currently supporting Tatar, Bashkir, Meadow Mari, Khakass, Tyvan (Kuzhuget et al., 2024), and Erzya (Gordeev et al., 2024) in both dev and devtest splits, and Udmurt in the dev split only. We also note the Smugri-Flores evaluation testset. Yankovskaya et al. (2023) took the first 250 sentences from the Flores+ devtest and commissioned professional translators to create a Flores+ extension with Finno-Ugric languages. This benchmark contains Meadow Mari, Hill Mari, Udmurt, Komi, Mansi, Erzya, and Moksha languages.

For languages lacking standard benchmarks, we commissioned native speakers to translate the Flores+ *dev* split for Karachay-Balkar, Moksha, Mansi, Yakut, and Komi. However, translators frequently struggled due to non-existent terminology in the target languages and the cultural detachment of the source topics. To mitigate the English-centric domain bias of Flores+ and address these terminological disconnects, we curated and commissioned translations for *Diversity500*, a

[1] https://github.com/kod-odin/lang-tasks/wiki/
[2] https://dict.fu-lab.ru/
[3] https://ru.ruwiki.ru

515-sentence multi-domain corpus. Domains include news texts (WMT-21), travel guides (Flores+), RuWiki excerpts, conversational dialogues, classic Russian literature, forum texts, and folklore lines. Dataset samples were manually selected by native speakers based on text length, named entity constraint, grammatical correctness, and cultural relevance to Russian regions. We commissioned professional translations for this corpus for the Khakass and Nogai languages.

The specific benchmark utilized for scoring each language is detailed in Table 1. We report the chrF++ metric [4] (Popović, 2015, 2017), as standard word-level BLEU (Post, 2018) heavily penalizes partial morphological matches in highly agglutinative languages. Neural-based metrics (e.g., COMET) are omitted due to a fundamental lack of support for these extreme low-resource languages.

Lang	Sentences	Dicts	Mono	Eval Set
tt	30,442		1,992	Flores devtest
cv	1,461	143	564	Flores devtest
ba	1,418		3,008	Flores devtest
tyv	296	20	935	Flores dev
krc	286		98	Flores dev
kjh	148	4	9	Diversity500
sah	58	156	1,811	Flores dev
nog	3	58	43	Diversity500
az	30,618			
uz	25,173			
ky	1,541			
kv	1,258	227	514	Flores dev
mhr	413		1,992	Flores dev
mns	115	36	314	Flores dev
udm	97	166	1,471	Smugri-Flores
myv	91	5	293	Flores devtest
mdf	86		156	Flores dev
mrj	15	216	163	Smugri-Flores
et	48,364			
hu	52,285			

Table 1: Data distributions per language (in thousands) and assigned evaluation benchmarks.

4 MT System Modeling

For our experimental setup, we decided to use a standard encoder-decoder Transformer with positional encoding, compatible with Transformers Seq2SeqTrainingArguments and Trainer. Throughout all the experiments, we used constant ‘8‘ attention heads, ‘0.1‘ dropout rate, ‘0.1‘ positional dropout rate, ‘200‘ max sequence length, and 512-dimensional embeddings and 2048-dimensional feed-forward networks. Models are trained until convergence on a train holdout with early stopping patience set to ‘4‘. We also use label smoothing of ‘0.1‘, scheduler warm-up for 16,000 iterations, a learning rate of ‘0.8e-3‘, and no weight decay.

In addition, we trained separate source and target Byte-Pair Encoding (BPE) tokenizers using the `tokenizers` library. To ensure balanced lexical representation and prevent data starvation, we applied a simple sampling strategy during both tokenizer and model training: any language corpus containing fewer than 1 million examples (including dictionary data) was duplicated to reach a minimum threshold of 1 million samples. Finally, exclusively for tokenizer training, we imposed a hard cap on high-resource languages, randomly subsampling them to a maximum of 20 million examples per language. This limits resource-heavy languages from dominating the subword vocabulary space.

All models were trained on a single NVIDIA Tesla A100 80GB GPU, and inferenced on an NVIDIA GeForce GTX 1080 Ti. For evaluation, we inferenced our models with a beam search approach where we set the beam size to ‘4‘.

4.1 Multilingual Setup

Our first set of experiments is focused towards understanding the cross-lingual transfer capabilities of Finno-Ugric and Turkic low-resource languages. We used a 6 encoder 6 decoder (6e6d) layer architecture, in addition to separate 10k src and 10k dst tokenizers trained for each experiment according to the data used.

We trained direct models (xx→ru, ru→xx) on parallel data, multilingual models based on language family group (finno→ru, ru→finno, turk→ru, ru→turk), multilingual models without additional high-resource languages (finnoLow→ru, ru→finnoLow, turkLow→ru, ru→turkLow), and highly multilingual models on combined Finno and Turk setups (all→ru, ru→all).

Table 2 indicates that direct models exhibit poorer performance on average across the family groups compared to multilingual setups in both xx→ru and ru→xx directions (direct vs multi: ru-finno 34.42 vs 46.276, ru-turk 33.338 vs 44.289, finno-ru 33.86 vs 48.359, turk-ru 31.442 vs 46.368).

[4] sacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1

Direction	Direct	MultiLow	Multi	All
tt-ru	**47.82**	47.18	47.62	46.5
cv-ru	47.12	**47.29**	46.89	43.78
ba-ru	43.98	**46.98**	46.97	45.15
tyv-ru	36.08	**52.52**	52.39	46.2
krc-ru	23.55	41.51	**41.58**	39.31
kjh-ru	21.75	46.24	**46.31**	42.27
sah-ru	18.44	**43.0**	41.72	34.26
nog-ru	12.8	44.82	**47.46**	41.41
kv-ru	50.27	50.83	**51.75**	50.28
mhr-ru	41.81	50.14	51.1	**51.57**
mns-ru	23.18	**46.0**	45.75	45.5
udm-ru	32.18	46.46	**46.9**	46.73
myv-ru	31.58	41.73	42.87	**43.21**
mdf-ru	41.34	54.74	**56.77**	56.29
mrj-ru	22.13	**45.17**	43.37	43.65
ru-tt	48.59	**48.75**	48.67	48.25
ru-cv	47.76	**48.69**	46.71	43.8
ru-ba	45.57	**47.33**	46.81	45.67
ru-tyv	39.1	**52.25**	50.18	46.65
ru-krc	26.11	**43.32**	42.27	39.23
ru-kjh	24.86	**46.73**	44.62	43.35
ru-sah	20.38	**44.42**	40.26	33.88
ru-nog	14.33	**38.94**	34.79	32.86
ru-kv	48.83	**50.51**	49.41	48.3
ru-mhr	42.83	**54.8**	52.63	51.48
ru-mns	21.95	39.73	**41.87**	40.29
ru-udm	33.67	**46.34**	45.77	44.72
ru-myv	35.87	**44.81**	44.35	44.25
ru-mdf	43.6	55.68	**56.08**	55.01
ru-mrj	19.05	**34.7**	33.82	33.98

Table 2: chrF++ automatic metric results of trained direct, multilingual low-resource, multilingual family-based, and highly multilingual models.

Only the tt-ru direct model performed best compared to other setups, with quality being comparable to the Multi setup.

Interestingly, we note that when translating from Russian to low-resource languages, multilingual setups without high-resource languages perform similarly or better across all 15 selected languages (ru-finno 46.653 vs 46.276, ru-turk 44.585 vs 44.289). Worth noting that the low-resource languages we work with use the Cyrillic writing script, while Estonian, Hungarian, Azerbaijani, and Uzbek use Latin. We hypothesize that this script mismatch may be one of the contributing factors towards such results. Test this hypothesis, we transliterated Uzbek sentences from Latin to Cyrillic using TurkicNLP toolkit (Hakimov, 2026), removed the Azerbaijani language from training, and re-trained tokenizers to fit the new language set (UzCyrl) accordingly. Additionally, we considered that better representation of low-resource languages in ru→turkLow tokenizer may also contribute to such findings, therefore we also re-trained ru→turkLow model with tokenizer from ru-turk setup (LowTok).

As shown in Table 3, reducing the amount of high-resource training samples and transliterating Uzbek (UzCyrl) performed similarly to ru→turk setup (Multi). However, model trained on ru→turkLow data with ru→turk tokenizer (LowTok) outperformed ru→turk setup on Bashkir, Tyvan, Khakass, Yakut, Nogai languages. This suggests that when translating to low-resource languages, it is important to simultaneously tune the language set in multilingual modeling and focus on language representation in tokenizers.

Direction	Low	LowTok	UzCyrl	Multi
ru-tt	**48.75**	48.58	48.3	48.67
ru-cv	**48.69**	46.74	46.99	46.71
ru-ba	**47.33**	47.03	46.59	46.81
ru-tyv	**52.25**	50.75	50.57	50.18
ru-krc	**43.32**	41.96	41.44	42.27
ru-kjh	**46.73**	45.06	44.56	44.62
ru-sah	**44.42**	40.99	40.69	40.26
ru-nog	**38.94**	35.57	36.43	34.79

Table 3: chrF++ automatic metric results on Turkic languages with different language sets and tokenizers.

When examining translation from low-resource languages to Russian ($xx \rightarrow ru$), we observe a directional asymmetry in cross-lingual transfer compared to the $ru \rightarrow xx$ direction. Multilingual models containing high-resource languages (Multi: finno-ru 47.867, turk-ru 46.192) performed similarly or better than setups strictly limited to low-resource languages (MultiLow: finno-ru 48.359, turk-ru 46.368). Interestingly, the highly multilingual `all-ru` model performed significantly better on Finno-Ugric languages compared to Turkic (All vs Multi: turk-ru 42.36 vs 46.368, finno-ru 48.176 vs 48.359), achieving the highest performance on Meadow Mari, Erzya, similarly on Udmurt, Hill Mari, Moksha and Mansi. We attribute this split to the intersection of model capacity, areal linguistics and macro-typology.

Both language families share deep agglutina-

tive profiles, but the Finno-Ugric languages that succeeded (Meadow Mari, Erzya) and the Turkic languages providing the bulk of the data (Tatar, Bashkir, Chuvash) belong to the Volga-Kama *Sprachbund* (Helimski, 2003). Due to centuries of geographic contact, languages of the Volga-Kama linguistic area exhibit structural convergence - such as shared properties in causativization and grammar (Bradley et al., 2022). This suggests that in the joint `all-ru` model, the Volga-Permic (Finno-Ugric) languages gained positive transfer from the 33-million-sentence Tatar, Bashkir and Chuvash (Turkic) corpora through shared linguistic characteristics. However, Turkic languages already possessed high internal genealogical homogeneity (Kuriyozov et al., 2020), making the baseline `turk-ru` model highly optimal. So for Turkic languages transitioning to the `all-ru` setup forced the 10k vocabulary and 55M model parameters to accommodate over 100 million sentences of distinct Hungarian and Estonian languages, providing zero synergy. This suggests that in low-resource setups, cross-lingual transfer is most effective when it exploits the structural and lexical convergences of a shared *Sprachbund*, whereas generalized multilinguality outside of established contact zones and language groups hurts performance.

4.2 Model Scaling

In low-resource machine translation, the relationship between available parallel data and model capacity dictates the efficiency of training. We investigated whether neural scaling laws apply uniformly across different agglutinative families by establishing a grid search over two parameters: percentage of available training data (20%, 40%, 60%, 80%, 100%) and model depth (1e1d, 2e2d, 4e4d, 6e6d, 8e8d, 10e10d). We trained this 5×6 grid for the `ru-finno` and `ru-turk` baseline setups, maintaining the previously fixed tokenizers.

We report the averaged chrF++ gains across each language family relative to the 20% data baseline. As illustrated in Figures 1 and 2, models with shallow capacities (1e1d through 4e4d) exhibit strict representational starvation: they show almost zero gain from additional training data. These models lack the parameter threshold necessary to decouple complex agglutinative morphology from pure lexical memorization, causing them to plateau prematurely regardless of data volume. On the other side, models with higher capacity (6e6d and above) cross this representational threshold and effectively leverage additional data, with 10e10d models benefiting the most from 100% of data.

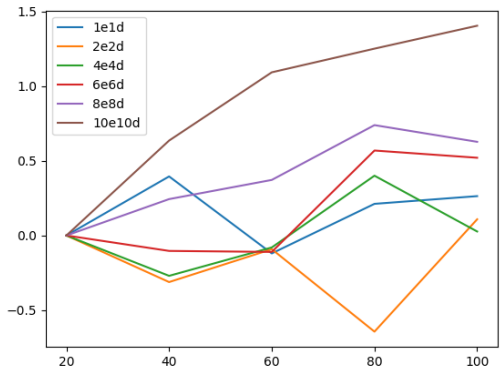

Figure 1: chrF++ gains on ru-turk languages compared to model trained on 20% of available training data

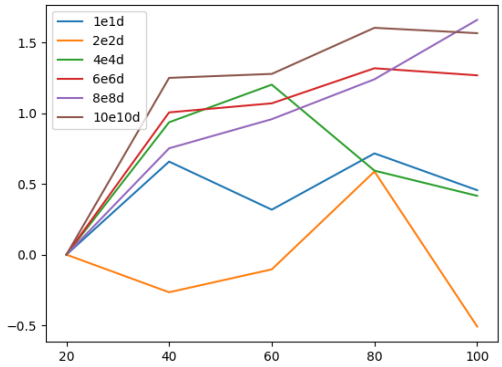

Figure 2: chrF++ gains on ru-finno languages compared to model trained on 20% of available training data

We note similar trends visible in both ru-turk and ru-finno models, suggesting that model scaling laws apply similarly for Finno-Ugric and Turkic language family groups, despite differences in language set, proportion of dictionary data and amount of parallel sentences present.

4.3 Encoder-Decoder Symmetry

Standard NMT systems conventionally employ symmetric architectures (e.g., 6e6d). However, having established that cross-lingual transfer exhibits strong directional asymmetry, we investigated whether asymmetric architectures provides advantage for these specific language families. For the generation-heavy $ru \rightarrow xx$ task, we shifted capacity towards the decoder (2e10d, 4e8d). For the comprehension-heavy $xx \rightarrow ru$ direction, we shifted capacity towards the encoder (8e4d, 10e2d).

Size	ru-finno	ru-turk	finno-ru	turk-ru
2e10d	45.06	43.362	-	-
4e8d	45.73	43.868	-	-
6e6d	**46.28**	**44.29**	48.86	**46.37**
8e4d	-	-	49.69	46.03
10e2d	-	-	**49.92**	45.39

Table 4: chrF++ automatic metric results on Turkic and Finno multilingual models, according to model encoder-decoder architecture.

As shown in Table 4, the symmetric 6e6d baseline generally remained the most robust architecture for $ru \rightarrow xx$ translation. However, the drop in performance when transitioning to the asymmetric 4e8d setup was remarkably marginal. This suggests a delicate capacity trade-off: while the decoder requires massive parameter volume to generate complex, agglutinative morphology across multiple languages, the encoder cannot be starved entirely, as Russian's fusional morphology still requires encoder depth to build a robust latent representation.

Differently, we observed asymmetrical architectural preferences during $xx \rightarrow ru$ translation. The `finno-ru` model achieved its absolute peak performance with a deeper encoder and shallower decoder (10e2d), outperforming the symmetric baseline. In contrast, the `turk-ru` model's performance monotonically degraded with increased encoder depth, peaking at the symmetric 6e6d setup.

Initially, we hypothesized that this disparity between the `finno-ru` and `turk-ru` models was caused by target-side data composition. As shown in Table 1, the Finno-Ugric languages contain a higher percentage of dictionary corpora, therefore we theorized that a shallow decoder excelled simply because it was more frequently tasked with producing shorter Russian outputs, rendering a 6-layer decoder redundant. To test this hypothesis, we conducted an ablation study where we filtered all dictionary corpora from the `finno-ru` training set and retrained the architectures. The results yielded 47.779 chrF++ for the symmetric 6e6d baseline, 48.816 chrF++ for 8e4d, and an absolute peak of 49.014 chrF++ for the highly asymmetric 10e2d model. Such metric differences are still marginal, however they suggest that Finno-Ugric languages, in contrast to the homogeneous Turkic family, may benefit from increased encoder capacity in language understanding tasks.

4.4 Synthetic Data

Having established the boundaries of architectural capacity and linguistic grouping, our final challenge is overcoming absolute data scarcity. Unsupervised data augmentation techniques, particularly Back-Translation (BT) and Forward-Translation (Zhang and Zong, 2016) (FT, or self-training), are foundational for low-resource scaling. To rigorously isolate the impact of synthetic data without introducing confounding variables, we maintained the controlled, symmetric 6e6d baseline configurations established in Section 4.1 (namely `finno-ru`, `turk-ru`, `ru-finnoLow`, and `ru-turkLow`). While integrating the areal *Sprachbund* findings or asymmetric architectures may increase absolute performance further, we leave their massive-scale synthetic integration for future work to preserve a clean ablation environment.

For our generation setup, we utilized the entirety of the collected low-resource monolingual corpora. For the Russian side, we uniformly sampled 2 million sentences from RuWiki and Newscrawl, constrained to a 200-token maximum length. We inferenced the baseline models using greedy decoding to generate the synthetic translations. We then trained models iteratively: first by adding only Back-Translated data (+bt1), followed by a mix of Back- and Forward-Translated corpora (+bft1). We finally used the best performing round-1 models to generate a second iteration of synthetic corpora (+bt2, +bft2).

Data	ru-finnoL	ru-turkL	finno-ru	turk-ru
no	46.65	46.30	48.36	46.37
+bt1	50.47	47.82	52.07	46.25
+bft1	47.33	46.87	**53.42**	47.27
+bt2	**51.44**	**48.16**	51.05	47.60
+bft2	49.13	47.24	52.10	**48.10**

Table 5: chrF++ automatic metric results when training with synthetic data. +bt represents back-translated data, bft represents back-translated and forward-translated data.

As shown in Table 5, synthetic data injection yields significant absolute chrF++ gains in our study. Interestingly, we note a distinct directional asymmetry regarding the effect of Forward-Translated (FT) data.

In the comprehension-heavy $xx \rightarrow ru$ direction, adding FT actively improved performance across both language families (`+bft1` outperform-

ing +bt1). Because the model is trained on a higher volume of highly fluent Russian targets (coming from both BT and parallel corpora) compared to the synthetic FT generations, the FT data is seamlessly absorbed. Therefore, FT acts as a highly effective form of Sequence-Level Knowledge Distillation, reinforcing the model's mapping confidence without harming its generative capabilities.

On the other hand, in the generation-heavy $ru \rightarrow xx$ direction, adding Forward-Translation actively degraded performance across the ru-finnoL and ru-turkL setups. In this case, low-quality FT data saturated with hallucinations and broken grammar dominated the ru-multiLow setups, yielding a negative impact on the models' translation capabilities.

Finally, we note that iterative generation (+bt2) provides further gains for almost all setups, safely cementing a new SOTA baseline for $ru \rightarrow xx$ directions without requiring billion-parameter scaling. A notable outlier to this iterative improvement is the finno-ru model. As noted in Table 1, the original Finno-Ugric training data is saturated with dictionary alignments. This resulted in a massive domain shift after the first BT iteration for both finno-ru and ru-finno models. Consequently, the second-stage BT yielded significantly different, more complex structural distributions that ultimately confused the encoder in an untagged setup, leading to representational collapse (Caswell et al., 2019).

5 Comparison with Existing Systems

To validate the real-world utility of our efficiently scaled NMT architectures, we benchmarked our best MultiMT models against a highly competitive suite of General-Purpose Large Language Models (LLMs) and commercial MT APIs.

For the LLM benchmarks, we utilized a 3-shot prompting strategy to minimize cross-lingual interference and to better align models to the low-resource translation task. Additionally, all instructions were formulated in Russian. Examples were strictly fixed across translation directions to ensure reproducibility, outputs were post-processed to resolve Cyrillic/Latin encoding collisions prior to chrF++ evaluation, and parsed to remove conversational words unrelated to translation. The full instruction templates can be seen in Figures 3, 4. Commercial MT systems (Majbyr.com, Neurotolge) were queried via public APIs during April 2026.

```
Ты эксперт-лингвист и профессиональный переводчик. Твоя
задача - перевести текст с {src_lang_name} на {dst_lang_name}
язык. Важно - в ответе должен быть только перевод без
дополнительных комментариев и пояснений.
Вот несколько примеров перевода:

{src_lang_name}: {example source 1}
{dst_lang_name}: {example translation 1}

{src_lang_name}: {example source 2}
{dst_lang_name}: {example translation 2}

{src_lang_name}: {example source 3}
{dst_lang_name}: {example translation 4}

Текст для перевода:
{text to translate}
```

Figure 3: Russian-language instruction template used for Few-Shot LLM inference.

```
You are an expert linguist and a professional translator.
Your task is to translate the text from {src_lang_name} to
{dst_lang_name} language. It is important that the response
should contain only the translation without additional
comments and explanations.
Here are some examples of translation:

{src_lang_name}: {example source 1}
{dst_lang_name}: {example translation 1}

{src_lang_name}: {example source 2}
{dst_lang_name}: {example translation 2}

{src_lang_name}: {example source 3}
{dst_lang_name}: {example translation 4}

Text to translate:
{text to translate}
```

Figure 4: English translation of the Few-Shot instruction template.

As demonstrated in Table 6, MultiMT drastically outperforms both specialized MT systems and massive LLMs in the $ru \rightarrow xx$ generative direction. Despite utilizing thousands of times fewer parameters than DeepSeek-V4 (1.6T) or Mistral (673B), our 55M-parameter models achieve absolute SOTA performance across both language families. This confirms that massive pre-training scale cannot independently overcome the morphological bottlenecks inherent to agglutinative low-resource generation. In the $xx \rightarrow ru$ comprehension tasks, massive LLMs (particularly Gemini 3 Flash and DeepSeek) achieve superior performance. This aligns with target-side data dominance: because these LLMs have undergone trillion-token pre-training on Russian corpora, they possess extraordinary target-side language fluency, allowing them to effectively "guess" semantic translations via few-shot cross-lingual projection. However, MultiMT remains highly competitive, demonstrating that dedicated MT systems can rival commercial giants without incurring massive compute costs.

System	Parameters	ru-finno	ru-turk	finno-ru	turk-ru
MultiMT	55M	**51.44**	**48.16**	51.05	47.60
DeepSeek-V4 pro	1.6T	43.059	44.816	56.551	52.921
DeepSeek-V4 flash	1.6T	38.141	41.499	55.01	51.579
DeepSeek-V3.1 terminus	685B	31.711	34.566	44.836	45.681
Qwen 3.5	397B	39.936	40.92	54.186	49.415
Kimi k2.6	1T	27.469	31.272	40.649	42.551
Mistral Large 3	673B	31.204	33.49	39.934	39.472
GPT OSS	120B	22.289	26.54	30.327	33.631
MADLAD-400	3B	17.458	22.996	40.703	35.049
Majbyr	Closed	46.083	-	57.743	-
Neurotolge	Closed	43.623	-	52.936	-
Gemini-3 flash	Closed	42.697	45.84	**60.204**	**56.356**
Grok 4.3	Closed	30.49	35.321	48.321	45.716
GPT 5.4 mini	Closed	28.374	33.758	48.86	47.705

Table 6: chrF++ automatic metric results of publicly available Large Language Models and specialized systems.

6 Conclusion

In this paper, we introduced `MultiMT`, a suite of 55-million-parameter machine translation systems specifically tailored for 15 low-resource Finno-Ugric and Turkic languages. Moving beyond standard model training, we conducted a rigorous evaluation of multilinguality, model scaling, asymmetric architectures, and synthetic data injection.

Our findings yield three crucial insights for low-resource NLP. First, we demonstrated a fundamental directional asymmetry in cross-lingual transfer: while translating *from* low-resource languages benefits heavily from the inclusion of high-resource anchor languages, translating *into* low-resource languages suffers from negative interference, dictating that target-side capacity must remain strictly reserved for the languages in question. Second, preliminary results suggest that asymmetric architectures with deeper encoders may help in agglutinative translation. Finally, our benchmark evaluation proves that massively scaled models remain vulnerable to agglutinative low-resource languages. Despite possessing up to thousands of times fewer parameters, our specialized `MultiMT` models significantly outperform foundational models in Russian-to-Low-Resource generation tasks.

By releasing our tokenizers, model configurations, and checkpoints, we aim to establish a robust baseline and encourage the development of specialized, compute-efficient translation systems for underrepresented languages.

Limitations

While our findings present robust solutions for low-resource machine translation, we acknowledge several limitations in our methodology. First, due to the immense computationally demanding nature of our 5×6 scaling grid and exhaustive multilingual ablations, we maintained static hyperparameters across all model sizes and language groups. Optimal performance for extreme low-resource MT likely requires dynamic hyperparameter tuning. Second, to constrain the experimental search space, we did not exhaustively ablate tokenization strategies. Across our experiments, we universally utilized separate source and target BPE tokenizers fixed at a strict 10,000-subword limit. Third, our synthetic data generation pipeline utilized untagged Back-Translation, which induced synthetic source drift during iterative rounds. Future iterations should apply strict sequence-level filtering and tagged BT to prevent representational collapse.

Finally, our typological conclusions are bound to the Finno-Ugric and Turkic language families. While our findings regarding asymmetric transfer and encoder-decoder capacity thresholds are highly applicable to agglutinative morphology, they may not generalize uniformly to deeply isolating languages or languages outside of the Volga-Kama or Eurasian linguistic areas. We rely on future research to validate these scaling laws on structurally divergent language families.

Ethics Statement

While our models offer compute-efficient alternatives to massive LLMs, we acknowledge several potential societal risks inherent to low-resource machine translation. Because the indigenous languages discussed are currently poorly supported by commercial LLMs and MT systems, it is relatively easy for native speakers to distinguish AI-generated text from authentic human writing. However, as our models significantly raise the generative quality bar, this distinction fades. Native speakers possess a high baseline of trust when interacting with text uniquely written in their minority language, implicitly assuming human authorship. Consequently, if our models are misused to generate malicious, phishing, or scam content, they could disproportionately harm vulnerable populations and severely damage trust in native-language digital communications.

Additionally, if deployed at scale, these models carry the risk of flooding the web with machine-translated content. This data pollution poses a long-term risk to the linguistic ecosystem, forcing future researchers to rely on complex preprocessing and filtering pipelines to distinguish and remove synthetic texts from authentic cultural data.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *Preprint*, arXiv:2305.13707.

Jeremy Bradley, Nikolett F. Gulyás, and András Czentnár. 2022. Causatives in the languages of the volga-kama region. *STUF - Language Typology and Universals*, 75(1):99–128.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *Preprint*, arXiv:1906.06442.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. Flores+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Sherzod Hakimov. 2026. Turkicnlp: An nlp toolkit for turkic languages. *Preprint*, arXiv:2602.19174.

Eugene Helimski. 2003. Areal groupings (sprachbünde) within and across the borders of the uralic language family: A survey. *Nyelvtudományi közlemények*, 100:156–167.

Eleonora Izmailova, Alexey Sorokin, and Pavel Grashchenkov. 2026. Adapting multilingual NMT to language isolates: The role of proxy language selection and dialect handling for nivkh. In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 127–137, Rabat, Morocco. Association for Computational Linguistics.

D. Karpov and 1 others. 2026. No one-size-fits-all: Building systems for translation to bashkir, kazakh, kyrgyz, tatar and chuvash using synthetic and original data. In *Proceedings of the LoResMT Workshop at ACL*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Sneha Kudugunta and 1 others. 2024. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint*.

Elmurod Kuriyozov, Yerai Doval, and Carlos Gómez-Rodríguez. 2020. Cross-lingual word embeddings for turkic languages. *Preprint*, arXiv:2005.08340.

Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. Enhancing Tuvan language resources through the FLORES dataset. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Jonas Pfeiffer, Naman Goyal, Xi Victoria Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. *Preprint*, arXiv:2205.06266.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copen-

hagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. *Preprint*, arXiv:1511.06709.

A. Syzdykov and 1 others. 2026. Turkicnlp: An nlp toolkit for turkic languages. *arXiv preprint*.

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.

Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Xiang "Tony" Cao, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, and 13 others. 2026. Omnilingual mt: Machine translation for 1,600 languages. *Preprint*, arXiv:2603.16309.

E. Yankovskaya and 1 others. 2023. Machine translation for low-resource finno-ugric languages. In *Proceedings of the NoDaLiDa / ACL Anthology*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1535–1545.

Appendix

6.1 Details on Diversity500

In Section 3.2 we introduced the Diversity500 evaluation corpus. Our motivation for creating this benchmark was the issue of excessively difficult examples in Flores+ and English-centric domains irrelevant for people speaking selected low-resource languages. Therefore, we decided to manually filter an already-existing benchmarks. We filtered 50 sentences from Flores+ devtest travel-guides domain, main criteria in this section were text length and presence of named-entities. For news domain we selected WMT-21 benchmark, from which we selected 100 texts mainly based on length. However, we made exceptions for longer texts, if they contained topics closely related to regions (e.g. street names, local celebrities or local news agencies). For wiki domain we randomly selected 100 articles present in RuWiki [5], and selected texts based on length, excessive presence of named entities, the ability to understand the example out of context, and structural divergence (most texts had typical wiki-format of: Title - Short description). For other domains, we selected texts from monolingual ruscorpora[6]. For books, we filtered 100 texts from "Russian classic literature" section based on fluency, length, and possibility to understand them without additional context. For 100 conversational examples we filtered "Speech" subset, mainly looking at ability to understand text out of context. For forum texts we utilized Social-Media subset, and selected 50 examples resembling blog posts or comments. While for folklore we filtered 15 examples of 4-lined sections.

6.2 Evaluation corpora collections

In order to provide a truthful comparison between experiments and other publicly available systems, we commissioned native speakers of Karachay-Balkar, Moksha, Mansi, Yakut, Komi, Khakass, and Nogai to create our benchmarks. We sourced translators through contacts in open-source datasets, regional linguistic institutes, and specialized language-preservation organizations. Table 7 shows what corpora was commissioned for what language, with amount of examples present.

Prior to the translation phase, we established strict annotation guidelines. We explicitly prohibited the use of existing LLMs or commercial MT systems to prevent machine-translation artifacts and system-biases. If specific source texts were impossible to translate due to a lack of existing native terminology, translators were permitted to skip them.

After receiving translations, quality assurance was managed in partnership with the same organizations and people providing the translations. Because datasets were translated by multiple native speakers, we asked them to cross-evaluate subsamples of their colleagues translations, by verifying translation appropriateness, naturalness, and grammatical accuracy. No issues were reported.

We also checked that evaluation data did not leak into train. We normalized the texts by stripping all

[5] https://ru.ruwiki.ru

[6] https://ruscorpora.ru

punctuation, after which we computed the Levenshtein distance between the entirety of our evaluation benchmarks (Flores+ and Diversity500) and both our training splits and the RuWiki/NewsCrawl monolingual corpora. We then sorted the results and manually evaluated the sentence pairs exhibiting the highest similarities score. This manual inspection verified a strict absence of exact matches or subset string overlaps, explicitly confirming that no training data contamination occurred.

Language	Dataset	Size
krc	Flores dev	997
mdf	Flores dev	997
mns	Flores dev	988
sah	Flores dev	997
kv	Flores dev	997
kjh	Diversity500	515
nog	Diversity500	515

Table 7: Details on commissioned evaluation corpora.

6.3 Non-aggregated Comparison Table

To be transparent with our comparison table results, we provide full tables with no language aggregation for language family groups. Additionally, we add to comparison systems, that partially support some of selected 15 languages - TranslateGemma 27b, NLLB-200, Google Translate. Results are present in Tables 8, 9, 10, 11, where we can clearly see that on $ru \rightarrow xx$ directions `MultiMT` achieved SOTA performance on 12 out of 15 languages, with exceptions on Komi, Yakut, Bashkir.

System	ru-mhr	ru-mrj	ru-kv	ru-myv	ru-mdf	ru-mns	ru-udm
MultiMT	**58.71**	**45.75**	54.99	**47.08**	**59.28**	**43.61**	**50.64**
deepseek-v3-1-terminus	34.51	31.51	33.48	33.42	33.3	23.13	32.63
deepseek-v4-flash	44.84	35.95	43.03	37.82	40.59	24.44	40.32
deepseek-v4-pro	50.47	41.97	48.94	42.25	45.76	26.67	45.35
gemini-3-flash-preview	48.48	40.16	45.75	42.13	47.63	28.63	46.1
google-translate	48.8	-	46.24	-	-	-	45.96
gpt-5.4-mini	27.88	25.1	32.01	26.43	31.32	26.44	29.44
gpt-oss-120b	21.19	23.39	22.03	23.05	25.56	17.68	23.12
grok-4.3	34.78	29.35	31.66	31.47	34.61	20.24	31.32
kimi-k2.6	29.03	27.43	25.79	28.02	33.39	20.85	27.77
madlad400	18.6	11.59	13.06	16.32	20.07	-	25.11
majbyr	55.68	42.77	**59.46**	41.37	47.58	29.19	46.53
mistral-large-2512	33.12	27.62	34.72	31.77	36.57	22.94	31.69
neurotolge	53.02	37.49	54.61	42.63	47.86	24.51	45.24
qwen3.5-397b-a17b	46.61	37.99	41.64	38.6	41.78	28.58	44.35
translate-gemma-27b	-	-	27.42	-	-	-	25.17

Table 8: chrF++ automatic metric for Finno-Ugric languages when translating from Russian.

System	mhr-ru	mrj-ru	kv-ru	myv-ru	mdf-ru	mns-ru	udm-ru
MultiMT	57.31	49.13	54.09	47.48	63.45	50.96	51.52
deepseek-v3-1-terminus	47.37	42.41	49.81	44.1	50.26	37.35	42.55
deepseek-v4-flash	58.02	53.03	58.55	52.4	63.19	45.17	54.71
deepseek-v4-pro	60.85	53.47	59.12	53.74	65.25	46.68	56.75
gemini-3-flash-preview	64.58	56.52	62.26	**57.03**	**68.91**	**52.65**	**59.48**
google-translate	58.11	-	56.78	-	-	-	55.07
gpt-5.4-mini	50.46	43.5	54.08	45.35	53.85	45.24	49.54
gpt-oss-120b	29.5	32.1	27.88	26.95	34.49	32.42	28.95
grok-4.3	51.41	45.79	50.67	46.2	56.51	39.32	48.35
kimi-k2.6	42.56	38.87	41.58	39.93	51.1	35.89	34.61
madlad400	42.19	44.43	43.05	40.07	51.02	29.9	34.26
majbyr	**65.58**	**56.93**	**63.46**	52.96	65.04	43.24	56.99
mistral-large-2512	37.69	35.39	44.54	38.52	46.35	36.86	40.19
neurotolge	57.66	50.58	58.54	42.64	60.93	46.82	53.38
qwen3.5-397b-a17b	58.61	53.94	57.06	52.82	62.28	38.16	56.43
translate-gemma-27b	-	-	33.82	-	-	-	32.75

Table 9: chrF++ automatic metric for Finno-Ugric languages when translating to Russian.

System	ru-tt	ru-ba	ru-cv	ru-sah	ru-tyv	ru-kjh	ru-krc	ru-nog
MultiMT	**48.43**	47.43	**48.8**	47.01	**52.41**	**48.84**	**43.99**	**48.37**
deepseek-v3-1-terminus	43.45	39.08	37.2	36.53	32.82	23.0	34.36	30.09
deepseek-v4-flash	47.52	47.29	46.14	43.83	40.51	30.76	40.42	35.52
deepseek-v4-pro	48.11	**47.46**	48.52	48.73	45.4	37.9	43.24	39.17
gemini-3-flash-preview	47.62	46.85	48.66	**49.13**	47.67	37.95	43.95	44.89
google-translate	46.15	-	48.42	47.47	46.86	-	-	-
gpt-5.4-mini	47.3	46.25	32.31	26.6	27.77	24.99	33.83	31.01
gpt-oss-120b	42.47	36.84	21.55	21.53	21.8	18.13	25.03	24.97
grok-4.3	45.11	43.77	37.24	36.5	29.41	24.7	36.18	29.66
kimi-k2.6	43.24	41.84	31.48	27.81	25.54	20.99	31.44	27.84
madlad400	34.94	38.43	21.82	23.54	15.8	8.75	22.71	17.98
mistral-large-2512	45.6	37.99	34.47	26.33	28.31	25.34	38.15	31.73
nllb200	45.88	45.77	-	-	-	-	-	-
qwen3.5-397b-a17b	46.94	47.02	46.22	47.05	43.03	23.79	39.1	34.21
translate-gemma-27b	-	38.97	29.32	28.86	30.01	-	-	-

Table 10: chrF++ automatic metric for Turkic languages when translating from Russian.

System	tt-ru	ba-ru	cv-ru	sah-ru	tyv-ru	kjh-ru	krc-ru	nog-ru
MultiMT	47.36	47.11	48.53	45.0	54.43	47.26	42.71	52.43
deepseek-v3-1-terminus	49.28	47.57	47.83	49.25	47.67	30.17	44.63	49.05
deepseek-v4-flash	50.34	50.38	53.34	54.34	55.08	45.81	49.07	54.27
deepseek-v4-pro	50.55	49.89	53.44	55.2	56.76	50.53	50.14	56.86
gemini-3-flash-preview	51.37	**51.07**	**56.0**	**57.56**	**61.76**	**55.71**	**53.07**	**64.31**
google-translate	**51.98**	-	53.54	53.27	55.77	-	-	-
gpt-5.4-mini	50.5	49.84	47.44	46.28	46.04	42.28	46.31	52.95
gpt-oss-120b	47.27	44.61	25.7	24.12	28.13	25.59	33.25	40.38
grok-4.3	47.6	47.58	47.0	47.72	46.47	36.46	44.61	48.29
kimi-k2.6	49.16	48.07	45.42	40.97	38.68	27.59	42.32	48.2
madlad400	29.38	32.89	41.11	33.06	35.77	26.68	41.99	39.51
mistral-large-2512	48.51	44.88	40.07	29.8	35.74	29.5	40.95	46.33
nllb200	46.18	46.97	-	-	-	-	-	-
qwen3.5-397b-a17b	49.86	49.72	53.07	53.34	54.53	32.77	47.83	54.2
translate-gemma-27b	-	45.46	41.2	39.09	37.12	-	-	-

Table 11: chrF++ automatic metric for Turkic languages when translating to Russian.