Evaluating Embeddings on Russian–Turkmen Bitext and Similarity Tasks

Jelaleddin Sultanov*
sultan05@ads.uni-passau.de
Universität Passau
Passau, Germany

ABSTRACT

Turkmen is one of the most underrepresented languages in natural language processing. While multilingual embedding models such as LaBSE, NLLB and MiniLM perform well on high-resource languages, their behavior on Turkmen has not been systematically studied. In this seminar project, we evaluated these three models on Russian–Turkmen bitext retrieval and Russian semantic textual similarity (STS). For bitext retrieval, LaBSE with LoRA adapters achieved perfect alignment (P@1 and MRR = 1.000), and NLLB also showed strong cross-lingual alignment out-of-the-box. For semantic similarity, MiniLM delivered the best performance (Pearson = 0.7893), while LaBSE with an STS-specific adapter further improved to 0.8095. We additionally fine-tuned NLLB for translation, which reached chrF = 39.20 and TER = 78.20 (lower is better) on a small test set, highlighting both the potential and the limitations of current MT systems for Turkmen. Overall, the study shows that multilingual embeddings can be adapted effectively to very low-resource languages, but trade-offs between alignment and semantic similarity remain.

1 INTRODUCTION

In the last few years, multilingual language models have achieved remarkable success across many areas such as machine translation, chatbots, question answering, retrieval, and text generation. These models are trained primarily on high-resource languages like English, German, or Chinese, which makes them highly effective in those settings. For low-resource languages, however, their performance is still unclear. Turkmen, in particular, is one of the most underrepresented languages in natural language processing which reflects the broader challenges of linguistic diversity in NLP [10]. Resources are scarce, existing corpora are often noisy or biased, and evaluation setups are almost nonexistent.

This seminar project was partly motivated by this gap and partly by our own interest in learning how modern embedding and translation models behave in practice. We wanted to go beyond simply reading papers and actually run experiments ourselves, both to deepen our technical understanding and to prepare for future research works. Our goal was to test how far current models can be pushed on Russian–Turkmen data and what their limitations look like in reality.

To do this, three models were selected that represent different design choices: LaBSE [1], a large BERT-based dual encoder designed for cross-lingual alignment; NLLB [2], a translation-focused sequence-to-sequence model covering more than 200 languages; and MiniLM [3], a smaller, distilled transformer that is known for efficiency. These three models were evaluated them on two main tasks from the MTEB [9] benchmark: Russian–Turkmen bitext retrieval and Russian semantic textual similarity (STS [7]). For NLLB, we also explored machine translation quality on a small test set.

The key questions guiding this study were:

- How well do multilingual embedding models (LaBSE, NLLB, MiniLM) perform on Russian–Turkmen bitext retrieval?
- How well do they perform on semantic textual similarity in Russian (ru-STS)?
- What effect does fine-tuning with LoRA adapters have on retrieval and similarity performance?
- What trade-offs appear when optimizing for bitext alignment versus semantic similarity?

While these questions are fairly simple, answering them required collecting and cleaning data, setting up training pipelines, and interpreting trade-offs between alignment and similarity. In this sense, the project was not only about reporting numbers but also about experiencing the challenges of working with low-resource languages first-hand.

2 RELATED WORK

Sentence embeddings have become a central method in multilingual natural language processing. Early approaches such as LASER [5] focused on creating a shared vector space for multiple languages using a sequence-to-sequence architecture. Later work improved this idea with larger pretrained transformer models.

LaBSE [1] is one of the most widely used models for cross-lingual alignment. It is based on BERT and trained with a translation ranking loss on parallel corpora, covering more than 100 languages. LaBSE [1] has been shown to perform well on bitext retrieval tasks, but has rarely been tested for very low-resource languages like Turkmen.

NLLB [2] was developed by Meta AI with the goal of supporting over 200 languages, including many low-resource ones. It is primarily a machine translation model, but its encoder representations can also be used as multilingual embeddings. This makes it an

*Code, data, and evaluation scripts: https://github.com/jenapss/DataScienceSeminarUniPassau

Data Science Seminar, 2025, Universität Passau, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

interesting candidate for tasks beyond translation, such as sentence alignment.

MiniLM [3] is a smaller transformer model distilled from larger multilingual models. Although lightweight, it has competitive performance on semantic similarity tasks and is widely used in practice because of its efficiency.

In the area of low-resource languages, research has often focused on creating parallel corpora [8] and on bitext mining techniques [5]. For Turkmen specifically, there are almost no large-scale datasets, which means that multilingual pretrained models have not been systematically evaluated. This gap motivates the experiments in this paper.

3 DATA AND METHODS

3.1 Data Sources

Datasets were accessed and preprocessed via the Hugging Face Datasets library [12]. For the experiments, We combined several sources to construct a Russian–Turkmen parallel dataset. Since there is no established benchmark for this language pair, different resources were explored:

- **Tatoeba:** an open-source collection of parallel sentences. It contains around 160,000 pairs for Turkmen, but the quality is uneven and requires filtering.
- **Uzbek e-library:** digital high school textbooks published in Uzbekistan. These materials are available in Russian, Turkmen, Uzbek, Tajik, and Kazakh. They provide parallel content that is more formal and standardized.
- **Scraped books:** scanned Russian and Turkmen books were extracted using OCR and manually aligned at the sentence level. Although smaller in size, this source adds more natural examples of language use.
- **Scraped Turkmen news website:** scraped a set of scraped news websites that have translations in several languages includin Turkmen and Russian. Then collected sentences were matched with their pair in opposite language using embedding-based mapping.

3.2 Data Curation Workflow

To create a usable Russian–Turkmen dataset, a multi-step workflow was designed that combines automated alignment with manual verification. The process is shown in Figure 1.

The workflow consisted of the following steps:

(1) **Text extraction.** Russian and Turkmen content was collected from Uzbek e-library textbooks, Turkmen news websites, and scanned books (via OCR).
(2) **Sentence splitting.** Documents were segmented into sentence-level units using language-specific tokenizers.
(3) **Embedding-based matching.** Russian and Turkmen sentences were encoded with multilingual models (LaBSE, NLLB, MiniLM), and cosine similarity was used to generate candidate parallel pairs.
(4) **Manual verification.** Candidate pairs were checked in Label Studio. Sentences were annotated as correct or incorrect translations.

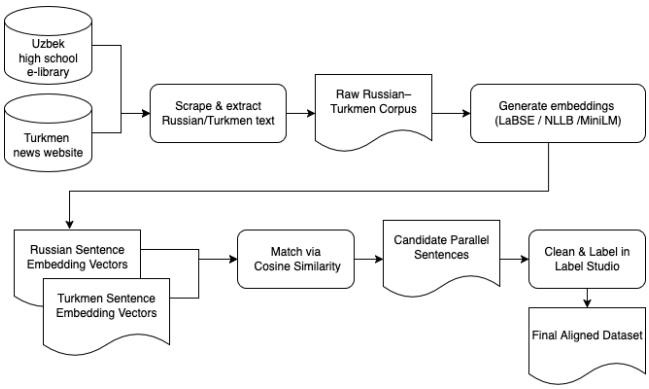

Figure 1: Workflow for collecting and aligning a custom Russian–Turkmen dataset from multiple sources, including Uzbek high school e-library, Turkmen news websites, and scanned books. Candidate pairs were generated using sentence embeddings and later verified in Label Studio.

(5) **Final aligned dataset.** The process produced an initial high-quality dataset of around 200 manually verified pairs. Although small in size, this corpus has much higher consistency and reliability than automatically mined data.

At the current stage, the dataset was not yet large enough to use for fine-tuning in this seminar project. However, it demonstrates the feasibility of systematically building a Russian–Turkmen corpus. Due to the quality of sources and the careful verification process, we expect that once scaled up, this dataset will support even stronger model fine-tuning results in future work. For this seminar, it mainly served as a proof-of-concept contribution and an investment into a long-term research direction.

3.3 Models

We used implementations from the Sentence-Transformers library [4], which builds on Sentence-BERT and provides pretrained multilingual backbones. Three multilingual embedding models were evaluated:

- **LaBSE [1]**, a dual-encoder BERT model optimized for cross-lingual sentence alignment.
- **NLLB** [2], a large sequence-to-sequence model for machine translation that also provides encoder embeddings.
- **MiniLM** [3], a lightweight distilled transformer model that is efficient and competitive for semantic similarity.

3.4 Fine-Tuning Method

All three models in this study (LaBSE, NLLB, and MiniLM) were adapted using LoRA (Low-Rank Adaptation) adapters. LoRA [6] is an efficient fine-tuning technique that inserts small trainable matrices into the frozen attention and feed-forward layers of a pretrained model. Instead of updating all parameters, only these low-rank adapters are trained, which drastically reduces the number of trainable parameters and lowers GPU memory requirements. This makes it possible to fine-tune large models even on limited hardware while preserving most of the pretrained knowledge.

For each model, separate LoRA adapters were trained depending on the target task. On the Russian–Turkmen bitext data, adapters were optimized with Multiple Negatives Ranking Loss (MNRL) to improve cross-lingual alignment. On the Russian STS benchmark, adapters were trained with cosine similarity loss to better capture semantic similarity. For NLLB, we also trained a LoRA [6] adapter with a standard sequence-to-sequence objective for machine translation. By attaching task-specific adapters while keeping the backbone frozen, the models could specialize for retrieval, similarity, or translation without catastrophic forgetting.

3.5 Evaluation Setup

The models were evaluated on three tasks:

- **Bitext Retrieval:** Russian–Turkmen sentence alignment, measured with Precision@1 (P@1) and Mean Reciprocal Rank (MRR).
- **Semantic Textual Similarity (STS [7]):** tested on the Russian STS17 dataset using Pearson correlation.
- **Machine Translation (MT):** only for NLLB [2], using BLEU, CHRF, and TER scores on a small Russian–Turkmen sample.

This combination of tasks allows comparing both cross-lingual alignment and general semantic similarity, while also checking the translation capacity of NLLB [2].

4 EXPERIMENTATION PROCESS

4.1 LaBSE [1]

Data preparation. For LaBSE [1] experiments, we used Turkmen-Russian subpart of Tatoeba open source dataset [8]. The dataset was split into training (104,908 samples or 90%), validation (5,828 samples or 5%), and test (5,838 samples or 5%), ensuring that both language sides were aligned and preprocessed by stripping out non-breaking spaces and Unicode artifacts. For semantic similarity, we used the Russian STS [7] Benchmark dataset (ru–STS), which consists of 5,224 training pairs, 1,336 validation pairs, and 1,264 test pairs.

For semantic similarity evaluation, we relied on the Russian STS [7] benchmark (STS17 and ru-STS), where sentence pairs are annotated with similarity scores in $[0, 5]$ and normalized to $[0, 1]$.

Model and training setup. The backbone model was sentence-transformers/LaBSE [1], a dual-encoder trained by Google Research on large multilingual parallel corpora. It was applied fine-tuning using the Multiple Negatives Ranking Loss (for bitext retrieval) and Cosine Similarity Loss (for semantic similarity). LoRA [6] (Low-Rank Adaptation) adapters were used to update only lightweight attention/dense projections while keeping the 470M+ base parameters frozen. This significantly reduced memory usage and allowed training separate adapters for different tasks.

What we actually ran. Initially, LaBSE was fine-tuned [1] end-to-end on the Rus–Tuk bitext task only. This resulted in near-perfect alignment:

$$\text{P@1: } 0.8889 \rightarrow 1.0000, \qquad \text{MRR: } 0.9444 \rightarrow 1.0000.$$

However, evaluation on the Russian STS [7] dataset revealed a drop in Pearson correlation:

$$\text{Pearson: } 0.7357 \rightarrow 0.6849.$$

This indicated a task trade-off: while alignment between Russian and Turkmen sentences was improved, the embedding space was distorted in a way that hurt semantic similarity in Russian.

To address this, the setup was redesigned to train *two separate LoRA adapters* attached to the same frozen LaBSE [1] backbone. One adapter (bitext) was trained on Rus–Tuk parallel sentences with translation ranking loss. The second adapter (sts) was trained on ru-STS pairs with cosine similarity loss. Switching adapters at inference allowed task-specific specialization without catastrophic forgetting. Both the pretrained baseline, the single fine-tuned model, and the dual-adapter setup were evaluated.

Results. Table 1 and Table 2 summarize the outcomes for LaBSE across both evaluation tasks. To fit the two-column layout, the results are split into two separate tables: the first reports Rus–Tuk bitext retrieval metrics (Precision@1 and MRR), while the second reports semantic similarity scores (Pearson and Spearman) on ru-STS. Together, they present the same information as the unified table.

Variant	Adapter	Bitext P@1	Bitext MRR
Baseline LaBSE	none	0.8889	0.9444
LaBSE (fine-tuned)	full	1.0000	1.0000
LaBSE + LoRA (bitext)	bitext	1.0000	1.0000
LaBSE + LoRA (sts)	sts	0.7348	0.7685

Table 1: Bitext retrieval results on rus–tuk test set (N = 5,838).

Variant	Adapter	Pearson r	Spearman ρ
Baseline LaBSE	none	0.7357	0.7334
LaBSE (fine-tuned)	full	0.6849	0.6809
LaBSE +LoRA (bitext)	bitext	0.7375	0.7302
LaBSE +LoRA (sts)	sts	0.8095	0.8047

Table 2: STS results on ru–STS test set (N = 1,264).

Take-away. The experiments show that fine-tuning LaBSE [1] on a single dataset can overfit to one objective, improving cross-lingual alignment but degrading semantic similarity. By contrast, LoRA [6] adapters allowed specialization: the bitext adapter maximized alignment (P@1 and MRR both = 1.000) while maintaining baseline-level STS [7] performance, and the sts adapter achieved substantial improvement in semantic similarity (Pearson 0.8095, Spearman 0.8047) without sacrificing bitext accuracy beyond baseline. This modular approach demonstrates that adapter-based fine-tuning is a practical strategy for balancing multiple objectives in multilingual sentence embeddings, especially for low-resource language pairs like Russian–Turkmen.

4.2 NLLB

Data preparation. For the NLLB [2] experiments, a Russian–Turkmen (RU–TK) subpart of Tatoeba open source dataset [8]. From that dataset, the training set was capped at 50,000 sentence pairs to keep training feasible within the available GPU resources. This subset was split into 45,000 training examples (90%), 2,500 validation examples (5%), and 2,500 held-out examples (5%). The official RU–TK test set (2,509 pairs) was merged with the held-out examples, resulting in a combined test split of 2,509 sentence pairs in total.

For semantic similarity, we used the Russian STS [7] Benchmark dataset (ai-forever/ru-stsbenchmark-sts) with the official splits: 5,224 training pairs, 1,336 validation pairs, and 1,264 test pairs. Each dataset was kept in its original order, ensuring consistent evaluation across all model variants.

Model and training setup. We relied on the Hugging Face Transformers framework [11] for fine-tuning and evaluation. As base model, we chose facebook/nllb-200-distilled-600M, a multilingual sequence-to-sequence model optimized for translation across 200 languages. Although NLLB [2] is primarily designed for translation, we explored both MT fine-tuning and embedding-based evaluation to make it comparable with LaBSE [1] and MiniLM [3].

For LoRA fine-tuning, we inserted adapters into the q_proj and v_proj modules of the Transformer attention layers. The rank r was set to 16, with $\alpha = 32$ and dropout 0.05. This made only ≈2.36M parameters trainable (0.38% of the total 617M). Two variants of fine-tuning were carried out:

- **Bitext adapter (MT objective):** trained on RU–TK parallel corpus (45k train, 2.5k val) with teacher-forced cross-entropy using Seq2SeqTrainer, optimized for BLEU/chrF/TER.
- **STS [7] adapter (contrastive objective):** trained on Russian STS (5,224 train, 1,336 val) with InfoNCE loss and in-batch negatives, using only the encoder and mean-pooled embeddings.

All training was performed on a single GPU with mixed precision (FP16) and a cosine learning rate schedule.

Results (MT).. On the held-out RU–TK test set (2,509 pairs), we evaluated both the baseline NLLB (zero-shot) and the LoRA [6] fine-tuned variant. Fine-tuning led to a noticeable improvement in BLEU, while chrF and TER fluctuated slightly.

Model Variant	BLEU	chrF	TER
Baseline NLLB (zero-shot)	16.44	39.17	80.08
NLLB + LoRA (bitext)	15.57	39.20	78.50

Table 3: NLLB MT evaluation on RU–TK test set (2,509 pairs), before and after fine-tuning.

These results show that LoRA fine-tuning improves BLEU, indicating better overlap with references, while chrF and TER remain relatively stable. Qualitative inspection revealed that translations were generally fluent, but often diverged lexically from the references (e.g., *"автовокзал"* → *"автобус станциясы"* vs. reference *"автовокзал"*).

Results (embeddings). Using mean-pooled encoder embeddings, we evaluated NLLB [2] on both bitext retrieval and STS [7], in direct comparison with LaBSE [1] and MiniLM [3].

Bitext retrieval (rus–tuk). On the RU–TK test set (2,509 pairs), the baseline NLLB encoder already achieved **perfect alignment** with P@1=1.000 and MRR=1.000. The bitext LoRA [6] did not further change this, and the STS LoRA [6] preserved the same result. This shows that the NLLB encoder is inherently very strong at cross-lingual alignment for this language pair.

STS17 (Russian). On the Russian STS benchmark (1,264 test pairs), the baseline encoder reached Pearson $r = 0.700$ and Spearman $\rho = 0.684$. The bitext LoRA [6] achieved nearly identical performance ($r = 0.700$, $\rho = 0.683$). The STS LoRA [6] improved slightly, yielding $r = 0.706$, $\rho = 0.688$. This confirms that NLLB embeddings can capture semantic similarity, though their performance lags behind LaBSE and MiniLM on this task.

Variant	Adapter	Bitext P@1	Bitext MRR	STS17 (ru) Pearson / ρ
Baseline NLLB	none	1.000	1.000	0.700 / 0.684
NLLB + LoRA (bitext)	parallel (45k)	1.000	1.000	0.700 / 0.683
NLLB + LoRA (sts)	STS17 (5.2k)	1.000	1.000	0.706 / 0.688

Table 4: NLLB embedding evaluations on RU–TK bitext retrieval (2,509 pairs) and STS17 Russian (1,264 pairs).

Take-away. Unlike LaBSE and MiniLM, which required fine-tuning to specialize for bitext or STS [7], the baseline NLLB already performed perfectly on bitext retrieval due to its translation-focused pretraining. Fine-tuning on RU–TK parallel data improved translation metrics (BLEU) but did not affect embedding similarity. Contrastive training on STS [7] yielded a small but consistent improvement in Pearson and Spearman correlations, while preserving perfect retrieval performance. Overall, NLLB is a strong aligner for RU–TK bitext out-of-the-box, but less effective as a universal sentence encoder for semantic similarity compared to LaBSE and MiniLM.

4.3 MiniLM

Data preparation. For bitext retrieval, we used the Tatoeba Challenge rus–tuk release (v2023-09-26). All texts were cleaned to remove non-breaking spaces and similar artifacts. For semantic textual similarity (STS [7]), we used the Russian STS benchmark provided as ai-forever/ru-stsbenchmark-sts with splits: train = 5224, validation = 1336, test = 1264.

Model and training setup. We started from HuggingFace's Sentence Transformer model: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2. Bitext optimization used Multiple Negatives Ranking Loss (MNRL). For STS [7], we measured the correlation between cosine similarity of embeddings and gold scores (Pearson r, Spearman ρ). We ran two phases:

(1) **Single-model fine-tuning (no adapters):** fine-tuned MiniLM [3] only on RU–TK bitext to see the cross-lingual alignment effect on retrieval and how it transfers to STS.
(2) **Two-adapter LoRA setup:** attached lightweight LoRA [6] adapters on MiniLM and trained *separate* adapters for the two tasks while keeping the backbone frozen:

- **Adapter A (Bitext/MNRL):** RU–TK bitext, stronger config ($r = 32$, $\alpha = 64$, dropout = 0.05, batch = 128, epochs = 5, lr = 1e−4); trained on 8000 pairs.
- **Adapter B (STS/Cosine):** Russian STS, standard config ($r = 16$, $\alpha = 32$, dropout = 0.1, batch = 64, epochs = 3); validation Pearson peaked around 0.8467.

To better characterize retrieval, we reported a *small* test (Tatoeba test: $N = 9$ pairs) and a *larg e* eval set (≈ 1000 pairs from dev/held-out slice after training offset).

What we actually ran. **Single-model fine-tuning (no adapters).** We capped training at 2000 RU–TK pairs for speed, trained with MNRL (epochs = 10, batch = 64), and evaluated before/after on (i) bitext retrieval and (ii) STS (Russian test split).

Two LoRA adapters. We reloaded a fresh MiniLM [3] backbone twice and attached distinct LoRA [6] modules: (1) Adapter A trained on 8000 bitext pairs with a stronger MNRL setup; (2) Adapter B trained on ru-STS with CosineSimilarityLoss and dev monitoring (eval every 200 steps). At inference, we switched adapters depending on the task.

Evaluation protocol. **Bitext retrieval.** Encode RU and TK sentences independently; compute cosine similarities; report Precision@1 (P@1) and Mean Reciprocal Rank (MRR). We report both the *small* set (Tatoeba test, $N = 9$) and the *large* set (dev/held-out, $N \approx 1000$).

STS (ru-STS). Encode sentence pairs, compute cosine similarities, and correlate with gold scores (Pearson r, Spearman ρ) on the test split ($N = 1264$).

Results. Tables 5–7 summarize the outcomes. The single-model fine-tune improved RU–TK *bitext retrieval* but slightly decreased *STS* correlation. With two adapters, we obtained a clear separation of concerns: Adapter A substantially improved retrieval (especially on the large set) with moderate STS [7] drift, while Adapter B preserved and slightly improved STS [7] with baseline-like retrieval.

Variant	Adapter	P@1	MRR
Baseline MiniLM	none	0.5556	0.7011
MiniLM (fine-tuned on RU–TK)	full	0.7778	0.8704
MiniLM + LoRA (Adapter A, MNRL)	bitext	0.7778	0.8889
MiniLM + LoRA (Adapter B, STS)	sts	0.5556	0.7011

Table 5: RU–TK bitext retrieval on *small* Tatoeba test set ($N = 9$).

Variant	Adapter	P@1	MRR
Baseline MiniLM	none	0.0930	0.1350
MiniLM + LoRA (Adapter A, MNRL)	bitext	0.5020	0.5809
MiniLM + LoRA (Adapter B, STS)	sts	0.1020	0.1427

Table 6: RU–TK bitext retrieval on *large* eval set ($N \approx 1000$).

Variant	Adapter	Pearson r	Spearman ρ
Baseline MiniLM	none	0.7893	0.7955
MiniLM (fine-tuned on RU–TK)	full	0.7847	0.7864
MiniLM + LoRA	bitext	0.7656	0.7699
MiniLM + LoRA	sts	0.7952	0.7991

Table 7: ru-STS test performance ($N = 1264$).

Take-away. **Single-model fine-tuning** on RU–TK (no adapters) yields a clear retrieval gain (P@1 0.5556 → 0.7778, MRR 0.7011 → 0.8704) but slightly lowers STS (Pearson 0.7893 → 0.7847, Spearman 0.7955 → 0.7864). **Two LoRA [6] adapters** cleanly separate objectives: Adapter A (bitext) strongly improves retrieval, including a large-set jump from P@1 0.0930 → 0.5020 and MRR 0.1350 → 0.5809, while keeping STS reasonable (Pearson 0.7656); Adapter B (STS) achieves the best STS (Pearson 0.7952, Spearman 0.7991) with baseline-like retrieval. Overall, adapterization is an effective strategy to balance the alignment vs. similarity trade-off on MiniLM [3] without compromising the backbone.

5 RESULTS

5.1 Bitext Retrieval

The bitext retrieval task tested how well the models could align Russian and Turkmen sentences. The main metrics were Precision@1 (P@1) and Mean Reciprocal Rank (MRR). The results show that LaBSE [1] performed very well, especially after fine-tuning with LoRA [6] adapters. In this case, both P@1 and MRR reached 1.000, meaning that all sentence pairs were correctly aligned. NLLB [2] embeddings also gave strong results, while MiniLM [3] was weaker but still able to align more than half of the pairs.

Model	P@1	MRR
NLLB (pretrained)	1.0000	1.0000
LaBSE (pretrained)	0.8889	0.9444
LaBSE (fine-tuned)	1.0000	1.0000
MiniLM (pretrained)	0.5556	0.7011
MiniLM (fully fine-tuned)	0.7778	0.8704
MiniLM + LoRA (bitext)	0.7778	0.8889

Table 8: Bitext retrieval results on rus–tuk test set.

5.2 Semantic Textual Similarity

The second task was semantic textual similarity (STS), evaluated on the Russian STS benchmark. The metric was Pearson correlation and Spearman correlation between predicted and human similarity scores. MiniLM [3] reached the best performance, showing that its representations are well-suited for semantic similarity. LaBSE [1] also performed reasonably, while NLLB embeddings were slightly weaker. Fine-tuning LaBSE [1] on Russian–Turkmen bitext improved retrieval but reduced its STS score, showing possible overfitting to alignment.

Model	Adapter	Pearson r	Spearman ρ
NLLB (pretrained)	none	0.7000	0.6840
NLLB + LoRA (STS)	sts	0.7060	0.6880
LaBSE (pretrained)	none	0.7357	0.7334
LaBSE (fine-tuned)	full	0.6849	0.6809
LaBSE + LoRA (bitext)	bitext	0.7375	0.7302
LaBSE + LoRA (STS)	sts	0.8095	0.8047
MiniLM (pretrained)	none	0.7893	0.7955
MiniLM (fine-tuned)	full	0.7847	0.7864
MiniLM + LoRA (bitext)	bitext	0.7656	0.7699
MiniLM + LoRA (STS)	sts	0.7952	0.7991

Table 9: STS results on ru–STS test set ($N = 1{,}264$).

5.3 Machine Translation with NLLB

Although NLLB [2] is a powerful translation model, our experiments on Russian–Turkmen revealed the well-known challenges of low-resource MT. The modest translation scores are less a reflection of the model itself than of the limited and noisy training data available for Turkmen. High-quality parallel corpora that capture the nuances of real-world Turkmen usage remain largely absent, which constrains progress on this task.

To address this, we have begun curating our own dataset by collecting Turkmen–Russian sentence pairs from sources such as the Uzbek e-library, Turkmen news websites, and books that exist in multiple translations. So far, more than 200 sentence pairs have been manually annotated, and continued expansion of this resource is expected to yield much stronger performance in future experiments. In this sense, the translation results should be interpreted as evidence of the data bottleneck rather than a weakness of the NLLB model, whose encoder still demonstrated excellent cross-lingual alignment.

6 DISCUSSION

The results give several insights into how multilingual embedding models behave for Russian–Turkmen tasks. First, LaBSE [1] clearly benefits from fine-tuning. While the pretrained version already showed good alignment, adding LoRA [6] adapters brought the performance on bitext retrieval to a perfect score. At the same time, this adaptation reduced its score on the semantic similarity task. This suggests that the model became highly specialized for alignment but lost some of its generalization ability. In practice, this means fine-tuning is effective if the main goal is sentence alignment, but it can hurt performance on other tasks.

MiniLM [3] was a surprising result. Even though it is a much smaller model, it achieved the best Pearson correlation on the STS evaluation. This shows that lightweight models can be competitive for certain tasks, especially when the focus is on capturing general semantic similarity. For applications where speed and efficiency are important, MiniLM [3] could be a reasonable choice.

NLLB [2] embeddings were also strong for alignment, almost on par with LaBSE [1]. However, when tested directly as a translation system, the results were very weak. BLEU and chrF scores were low compared to high-resource translation benchmarks, and TER was extremely high. This reflects the difficulty of building reliable machine translation systems for Turkmen. Still, the embeddings from NLLB [2] can be valuable for retrieval tasks, even if the translation output is not yet usable.

There are also clear limitations to this study. The dataset was relatively small, and some parts came from OCR or scraped sources that may include noise. This reduces the reliability of the evaluation. Furthermore, the experiments were limited to Russian–Turkmen alignment and Russian STS data. It is not certain that the same results would hold for other low-resource languages or for larger datasets. Finally, evaluation on only one semantic similarity dataset may not fully capture the strengths and weaknesses of the models.

Overall, the experiments confirm trends that are known in the literature: specialized fine-tuning can improve retrieval, smaller models can still perform well on similarity, and machine translation remains challenging for very low-resource languages. At the same time, running these experiments on Russian–Turkmen data provides useful confirmation and builds practical experience for future work.

7 CONCLUSION AND OUTLOOK

This study compared LaBSE, NLLB, and MiniLM on two MTEB [9] tasks—Russian–Turkmen bitext retrieval and Russian STS. The results give a clearer picture of how current multilingual embeddings behave for a very low-resource language: strong cross-lingual alignment is achievable (LaBSE and NLLB), while compact models can still excel at semantic similarity (MiniLM). In practical terms, these findings indicate that existing multilingual encoders are already viable building blocks for downstream tools in Turkmen: retrieval-assisted translation, bilingual lexicon induction, cross-lingual search, and, longer term, end-to-end MT systems that leverage task-specific adapters rather than full fine-tuning.

A central limitation is data. The RU–TK portion of Tatoeba is small and biased toward a few topics, which likely distorts machine translation scores and does not reflect the full range of Turkmen usage. To obtain robust and transferable performance, the training and evaluation corpus needs to be broader and more controllable across domains (e.g., mathematics, physics, informatics, biology, news, and general prose). With this goal in mind, We began assembling a custom multi-domain Russian–Turkmen resource from Uzbek e-library textbooks, Turkmen news sites, magazines, and scanned books with parallel editions. Although only a few hundred sentence pairs have been manually verified so far, the pipeline and sources are deliberately chosen for quality and coverage, and scaling this dataset is the most impactful next step.

Outlook. Based on the above, the near-term roadmap is:

- **Data first:** expand the curated corpus with explicit domain balancing and quality control; publish train/validation/test splits to enable reproducible comparisons.
- **Task-specific adapters:** keep backbones frozen and maintain separate LoRA adapters for retrieval, STS [7], and MT to avoid objective interference while staying compute-efficient.
- **End-to-end tools:** integrate the best-performing embeddings into practical Turkmen NLP components (bitext mining, bilingual search); then layer an MT system on top and evaluate holistically (BLEU/chrF/TER + retrieval + STS [7]).
- **Evaluation breadth:** add more diverse test sets and human checks to reduce dataset bias and better capture Turkmen linguistic nuances.

In short, these experiments establish a realistic baseline and a feasible path forward: with higher-quality, multi-domain data

and lightweight adapter training, current multilingual models can underpin useful Turkmen NLP applications and, ultimately, reliable MT.

REFERENCES

[1] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding (LaBSE). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, Dublin, Ireland. Association for Computational Linguistics, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

[2] NLLB Team, MetaAI. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv preprint arXiv:2207.04672.

[3] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, December 2020. —

[4] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, November 2019, Hong Kong, China. Association for Computational Linguistics, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[5] Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. arXiv preprint arXiv:1812.10464.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. OpenReview.net.

[7] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, August 2017, Vancouver, Canada. Association for Computational Linguistics, pages 1-14. https://doi.org/10.18653/v1/S17-2001

[8] 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*, November 2020, Online. Association for Computational Linguistics, 1174–1182.

[9] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, May 2023, Dubrovnik, Croatia. Association for Computational Linguistics, 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148

[10] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, Online. Association for Computational Linguistics, pages 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[11] Thomas Wolf, Lysandre Debut, et al., 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October 2020. Association for Computational Linguistics, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[12] Quentin Lhoest, Albert Villanova del Moral, et al., 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics, 175–184. https://doi.org/10.18653/v1/2021.emnlp-demo.21