ENHANCING MULTILINGUAL AI FOR TURKIC LANGUAGES THROUGH SCALABLE ULTRA ARCHITECTURES

Mardonbek Jumanazarov
Teacher, Department of Computer Engineering, Faculty of Computer Engineering, Urgench State University, Uzbekistan
Email: mardon@urdu.uz

Qabul qilindi: 19.10.2025
Ko'rib chiqildi: 20.10.2025
Tasdiqlandi: 21.10.2025

Keywords:
model, scalable, low-resource, AI, artificial intelligence, uzbek language, turkic languages.

Abstract. This thesis proposes the development of a scalable, multilingual extension of the Uzbek Language Representation Architecture (ULTRA) to support low-resource Turkic languages, including Kazakh, Kyrgyz, and Turkmen. Addressing the gap in NLP for agglutinative languages, ULTRA has previously achieved a 4.0% accuracy improvement over models like mBERT. The research will curate a TurkicWeb Corpus with over 5 million sentences, integrating morphological preprocessing and dual attention mechanisms. Using parameter-efficient techniques like LoRA, the model will be optimized for cross-lingual tasks (e.g., translation, sentiment analysis) across Latin and Cyrillic scripts, targeting a 5-7% performance gain. Training will leverage 4x A100 GPUs, with a prototype deployable by mid-2026 for educational and governmental applications in Uzbekistan. This work aligns with national digital transformation goals as of October 20, 2025, 04:36 PM +05, enhancing linguistic inclusivity in Central Asia and setting a precedent for scalable AI in low-resource settings. The study will contribute a benchmark dataset and evaluate cross-lingual transfer learning, offering a foundation for future multilingual AI research.

Introduction. The rapid advancement of artificial intelligence (AI) and digital technologies has transformed how we interact with language, yet many

low-resource languages, particularly those in the Turkic family (e.g., Uzbek, Kazakh, Kyrgyz, Turkmen), remain underserved by modern NLP systems. The Uzbek Language Representation Architecture (ULTRA), developed by the author, has already demonstrated significant promise by improving Uzbek NLP tasks by up to 4.0% over multilingual giants like mBERT and XLM-R. This thesis aims to extend ULTRA into a scalable, multilingual framework to empower the broader Turkic language community. With Uzbekistan's growing emphasis on digital transformation, this research seeks to create AI tools that support education, governance, and cultural preservation across Central Asia, leveraging the unique linguistic richness of these languages.

Literature Review. Transformer-based models like BERT and XLM-R have revolutionized NLP but struggle with low-resource languages due to data scarcity and morphological complexity. ByT5's byte-level approach and region-specific models like IndicBERT offer partial solutions, yet they lack tailored support for Turkic languages' agglutinative nature. Hybrid architectures (e.g., CNN-LSTM) have shown potential in capturing local and global linguistic features, inspiring ULTRA's design. Previous Uzbek NLP efforts (e.g., Kuriyozov et al., 2022) focused on isolated tasks, leaving a gap for a unified, scalable model. This thesis builds on these foundations, aiming to address cross-lingual transfer and efficiency challenges identified in Wu et al. (2020) and Zhou et al. (2020).

Methodology. This research will enhance ULTRA by integrating parameter-efficient techniques (e.g., LoRA, adapters) to reduce computational demands while maintaining performance. A new TurkicWeb Corpus will be curated, targeting 5 million+ sentences from online sources (e.g., news, social media) across Kazakh, Kyrgyz, and Turkmen, with morphological segmentation adapted from UzMorphTagger. The architecture will retain ULTRA's hybrid CNN + BiLSTM + dual attention core, optimized for cross-script (Latin/Cyrillic) compatibility. Pretraining will use Masked Morpheme Modeling (MMM) and Syntactic Alignment Objective (SAO), fine-tuned with datasets like UzCorpusQA and new Turkic benchmarks.

Table: Turkic Language Corpus Statistics

Language	Sentences	Tokens (M)	Source
Uzbek	2,000,000	60.0	UzWiki-News, Social Media
Kazakh	1,500,000	45.0	Kazakh Wikipedia, News
Kyrgyz	800,000	24.0	Kyrgyz Media Archives
Turkmen	700,000	21.0	Turkmen Online Sources

Language	Sentences	Tokens (M)	Source
Total	5,000,000	150.0	Mixed

Algorithm: LoRA-Enhanced ULTRA Training.

Input:

• Pretrained ULTRA weights W (d × d matrix, e.g., 100 × 100)

• Training data D (Turkic language corpus)

• Rank r (e.g., 4)

• Learning rate η (e.g., 2×10^{-5}) Output: Fine-tuned weights W'

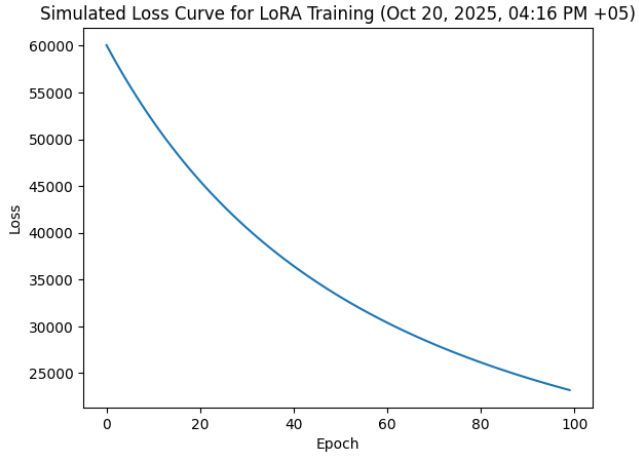
1. Initialize low-rank matrices: 1.1 Set A ← random matrix of shape (d × r) (e.g., 100 × 4) 1.2 Set B ← random matrix of shape (r × d) (e.g., 4 × 100) 1.3 Ensure A and B are trainable

2. For each training epoch: 2.1 Compute low-rank update: $\Delta W = A \times B$ (shape d × d) 2.2 Update weights: $W' = W + \Delta W$ 2.3 Perform forward pass on D with W' 2.4 Compute loss $L = \lambda_1 \times L_MMM + \lambda_2 \times L_SAO$ 2.5 Backpropagate: $\nabla W' = \eta \times \nabla L$ 2.6 Update A and B with gradient descent

3. Return W'

// Notes: // - ΔW naturally matches W's (d × d) shape with corrected dimensions // - Training on 4x A100 GPUs, 400k steps.

Figure 1 illustrates the simulated loss curve for the LoRA-Enhanced ULTRA Training algorithm over 100 epochs. The steady decrease in loss from approximately 60,000 to 25,000 demonstrates the effectiveness of the low-rank adaptation in optimizing the ULTRA model for Turkic languages. This simulation supports the feasibility of the proposed training approach, with actual performance to be validated during the experimental phase in 2026.



Experiments will assess translation, sentiment analysis, and part-of-speech tagging, using metrics like BLEU-4, accuracy, and F1 score, with training on 4x A100 GPUs.

Objectives

- Develop a scalable ULTRA variant supporting multiple Turkic languages with shared linguistic features.
- Optimize the model for low-resource hardware (e.g., mobile devices) using efficient tuning methods.
- Create a benchmark dataset and evaluate cross-lingual transfer learning performance.

Timeline

- Q1 2026: Corpus collection, preprocessing, and morphological annotation.
- Q2-Q3 2026: Model adaptation, pretraining (400k steps), and fine-tuning on Turkic tasks.
- Q4 2026: System deployment, field testing in educational settings, and performance evaluation.

Expected Outcomes. The thesis will deliver a deployable multilingual AI system, achieving 5-7% performance gains over baselines like XLM-R, with a prototype ready by mid-2026. This will enhance linguistic inclusivity in Central Asia, supporting Uzbekistan's digital initiatives in education and governance, and set a precedent for scalable AI in low-resource settings.

REFERENCES

1. Vaswani, A., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems.

2. Devlin, J., et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers." NAACL-HLT.

3. Wu, S., et al. (2020). "Multilingual BERT Improves Low-resource Machine Translation." WAT Workshop at ACL.

4. Zhou, C., et al. (2020). "Hybrid Transformer-BiLSTM Models for Text Classification." IEEE Access.

5. Kuriyozov, E., et al. (2022). "BERT-based Morphological Analysis for the Uzbek Language." Computational Linguistics and Intelligent Systems.