CHALLENGES AND OPPORTUNITIES FOR LOW-RESOURCE TURKIC LANGUAGES IN THE ERA OF LARGE LANGUAGE MODELS

Author: Ulanbek Soliyev, Student, Japan Digital University

Academic Supervisor: Lochinbek Boboev, PhD, Head of the IT Department, Japan Digital University

Abstract

Large Language Models (LLMs) have become the foundation of modern Artificial Intelligence applications, enabling significant progress in text generation, machine translation, question answering, and conversational systems. However, their performance remains uneven across languages due to substantial differences in the availability of digital resources. While high-resource languages benefit from extensive web content and large-scale linguistic datasets, many Turkic languages continue to experience limited representation in multilingual AI systems. This imbalance reduces the quality of AI-generated content and restricts access to advanced language technologies for millions of native speakers.

This paper examines the major challenges affecting LLM performance in low-resource Turkic languages, including digital content scarcity, insufficient training datasets, agglutinative linguistic structures, and limited evaluation benchmarks. Recent multilingual AI initiatives are reviewed to identify current progress and remaining limitations. Based on existing research, the study proposes a conceptual framework emphasizing community-driven content creation, open linguistic resources, multilingual benchmarks, and collaborative dataset development. The findings suggest that sustainable improvement in LLM performance depends not only on larger models but also on strengthening the digital ecosystem of low-resource languages through continuous content and resource development.

Keywords: Language Models, Low-Resource Languages, Turkic Languages, Natural Language Processing, Multilingual Artificial Intelligence, Digital Resources

Introduction

Large Language Models (LLMs) have become a cornerstone of modern Artificial Intelligence, enabling significant advances in text generation, machine translation, information retrieval, and conversational systems. Their widespread adoption across education, research, and industry has increased the importance of multilingual language support. [1][5]

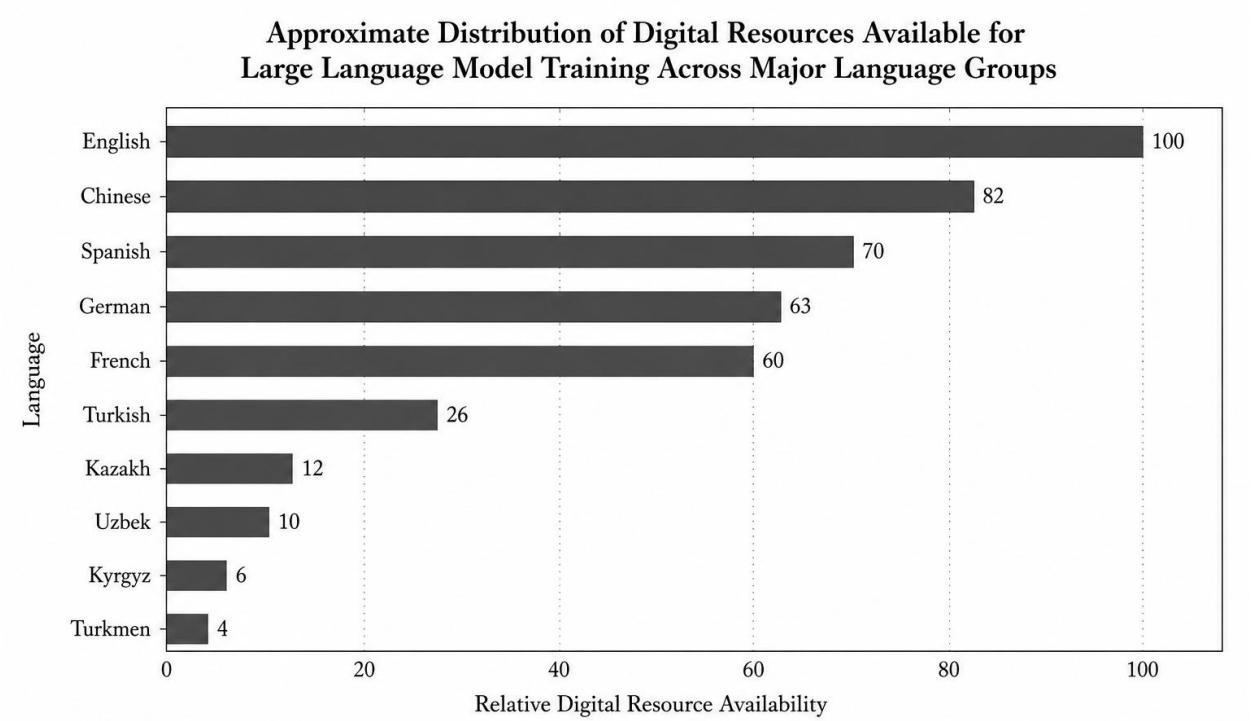
Despite these advances, LLM performance remains highly uneven across languages. High-resource languages such as English, Chinese, Spanish, and German benefit from extensive digital content, large linguistic corpora, and publicly available datasets, whereas many low-resource languages remain underrepresented in multilingual AI systems. This imbalance directly affects the quality of language understanding and generation. [2][6]

The challenge is particularly significant for Turkic languages. Although more than 200 million people speak Turkic languages worldwide, many—including Uzbek, Kazakh, Kyrgyz, Turkmen, and Karakalpak—still lack sufficient digital content, annotated corpora, benchmark datasets, and language technologies. Consequently, their representation in modern LLMs remains limited, resulting in lower performance across various NLP tasks. [7][8]

In addition to limited data availability, the agglutinative structure of Turkic languages presents further computational challenges for tokenization and language modeling. Recent studies suggest that improving AI performance for these languages requires not only larger models but also richer digital resources and language-specific datasets. [3][10]

This study investigates the key factors limiting LLM performance in low-resource Turkic languages and reviews current research addressing these challenges. It further discusses practical directions for strengthening digital resources and improving multilingual AI support for underrepresented languages. [4][6]

Approximate Distribution of Digital Resources Available for Large Language Model Training Across Major Language Groups.



The values illustrate the approximate relative availability of publicly accessible digital resources commonly used for multilingual Large Language Model training. They are intended for conceptual comparison rather than exact measurement.

Understanding Low-Resource Languages

The term *low-resource language* refers to a language that lacks sufficient digital and linguistic resources for developing modern Natural Language Processing (NLP) systems. Unlike high-resource languages, which possess extensive digital corpora, annotated datasets, and computational tools, low-resource languages have limited publicly available resources, reducing their representation in multilingual AI models. [6][11]

The distinction between high- and low-resource languages depends primarily on the availability of digital resources rather than the number of native speakers. Languages such as English, Chinese, German, and Spanish benefit from vast amounts of web content, scientific publications, and open linguistic datasets that support AI training and evaluation. [2][13]

Although the Turkic language family includes more than 200 million speakers, many of its languages—including Uzbek, Kazakh, Kyrgyz, Turkmen, and Karakalpak—remain digitally underrepresented. Limited web content, annotated corpora, and benchmark datasets reduce their contribution to multilingual language model training and negatively affect AI performance. [7][14]

Another challenge is the agglutinative morphology of Turkic languages. Words are formed by attaching multiple suffixes to a common root, producing numerous valid word forms. This increases vocabulary complexity and makes tokenization and subword segmentation more difficult for multilingual language models. [10][15]

Overall, improving AI support for low-resource Turkic languages requires expanding digital content, developing open linguistic resources, and creating standardized evaluation benchmarks. These efforts are essential for ensuring broader representation of Turkic languages in future multilingual AI systems. [4][11]

Comparison Between High-Resource and Low-Resource Languages

Characteristic	**High-Resource Languages**	**Low-Resource Turkic Languages**
Digital Content Availability	Very High	Low
Public Training Datasets	Extensive	Limited
Wikipedia Coverage	Large	Small
NLP Benchmarks	Comprehensive	Limited
LLM Performance	High	Moderate to Low
Community Contributions	Extensive	Developing

Why Do Large Language Models Perform Poorly for Turkic Languages?

The performance of Large Language Models (LLMs) depends heavily on the quantity and quality of the data used during pre-training. Since modern transformer-based models learn statistical patterns from large multilingual corpora, languages with abundant digital resources achieve better representation than low-resource languages. [2][15]

One of the primary factors limiting LLM performance in Turkic languages is the scarcity of high-quality digital content. Major multilingual datasets, including Common Crawl, Wikipedia, OSCAR, and mC4, contain significantly fewer documents in Uzbek, Kazakh, Kyrgyz, Turkmen, and other Turkic languages than in English or other high-resource languages. This imbalance reduces the quality of language representations learned during training. [3][16]

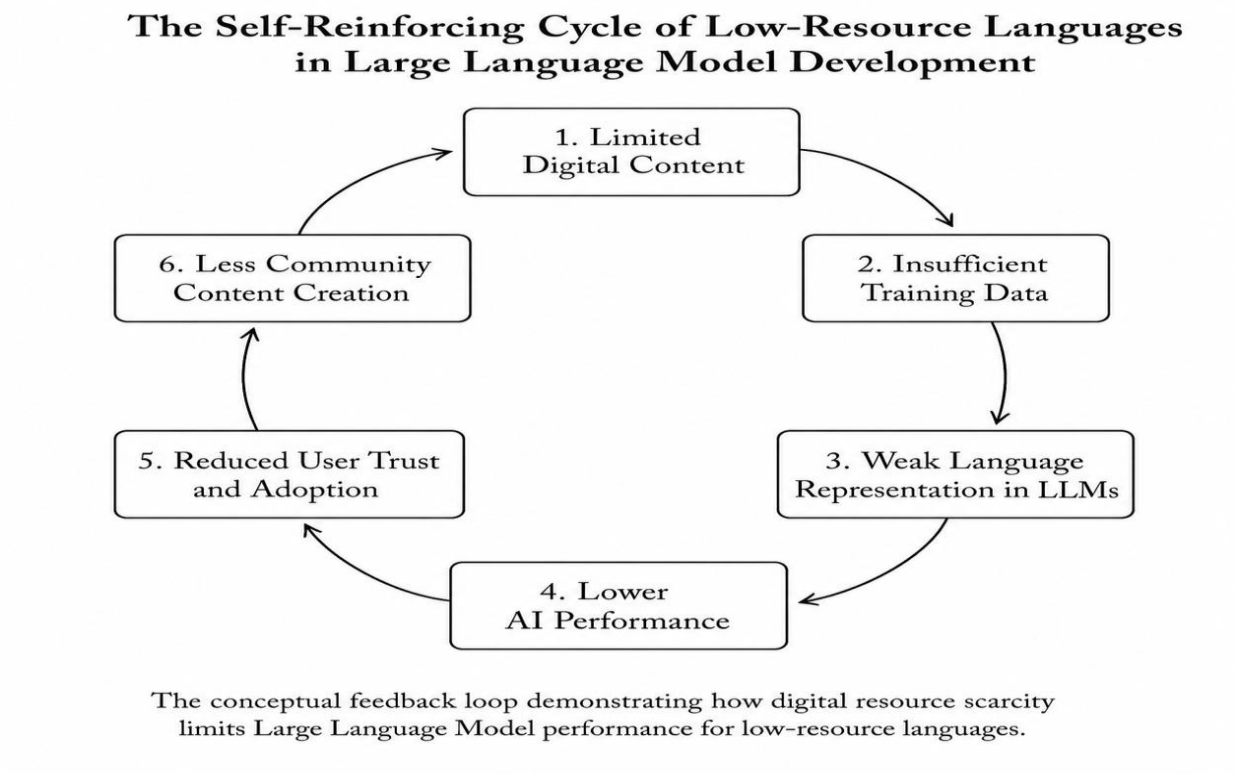
Another limitation is the shortage of annotated linguistic resources required for supervised NLP tasks such as machine translation, question answering, and text classification. Compared with high-resource languages, most Turkic languages lack sufficient benchmark datasets for both model training and evaluation. [6][13]

The agglutinative structure of Turkic languages presents an additional challenge. A single root word can generate numerous grammatical forms through suffixes, increasing vocabulary complexity and making tokenization less efficient for multilingual language models. This often results in weaker semantic representations and lower model accuracy. [10][15]

Limited digital resources also reduce the coverage of scientific knowledge, technical terminology, and domain-specific information available during model training. Consequently, LLMs tend to generate less accurate and less comprehensive responses in many Turkic languages, particularly in specialized fields such as medicine, engineering, and computer science. [4][16]

Overall, the relatively weak performance of LLMs in Turkic languages is primarily caused by limited digital content, insufficient linguistic datasets, morphological complexity, and inadequate evaluation resources. Addressing these challenges requires expanding open digital resources and strengthening community-driven language resource development rather than relying solely on larger model architectures. [11][12]

Factors Limiting Large Language Model Performance in Low-Resource Turkic Languages



The conceptual feedback loop demonstrating how digital resource scarcity limits Large Language Model performance for low-resource languages.

Current Research and Existing Solutions

In recent years, considerable efforts have been made to improve Natural Language Processing (NLP) for low-resource languages through multilingual language models, open datasets, and collaborative data collection initiatives. These projects have significantly expanded language coverage and improved multilingual AI performance. However, recent studies indicate that many low-resource languages, including several Turkic languages, continue to receive substantially lower support than high-resource languages. [14][6]

One of the most influential initiatives is **No Language Left Behind (NLLB)**, developed by Meta AI. The project introduced a multilingual translation model covering more than 200 languages and demonstrated that large-scale multilingual training can substantially improve translation quality for underrepresented languages.

Nevertheless, translation quality remains closely tied to the availability of parallel training data, which is still limited for many Turkic languages. [11]

Another important contribution is **BLOOM**, an open multilingual Large Language Model trained through international collaboration. Unlike proprietary systems, BLOOM provides researchers with an open platform for studying multilingual language modeling and developing language-specific applications. Despite supporting numerous languages, its performance varies considerably according to the quantity and quality of available training data. Consequently, languages with limited digital resources continue to achieve lower accuracy than high-resource languages. [12]

Large multilingual datasets have also contributed to recent progress. Resources such as **FLORES-200**, **Common Voice**, **OSCAR**, and **mC4** provide multilingual text and speech corpora that are widely used for training and evaluating AI systems. These datasets have expanded language coverage considerably; however, the representation of many Turkic languages remains relatively small compared with English and other widely spoken languages. [13][14]

Beyond model development, researchers increasingly emphasize the importance of community participation in building sustainable language resources. Crowdsourced text collection, collaborative corpus creation, and open-source linguistic resources have become essential strategies for improving language technologies in underrepresented languages. Several studies conclude that long-term progress depends not only on larger AI models but also on continuous expansion of publicly available, high-quality digital content. [6][16]

Although these international initiatives represent significant progress, they do not fully eliminate the resource gap affecting Turkic languages. The limited availability of digital content, benchmark datasets, and language-specific NLP resources continues to restrict the performance of Large Language Models. Therefore, future research should focus not only on developing more powerful multilingual models but also on strengthening the digital ecosystems of low-resource languages through collaborative resource creation and open knowledge-sharing initiatives. This perspective forms the basis of the conceptual framework proposed in the following section. [15][11]

Comparison of Major Multilingual AI Initiatives Supporting Low-Resource

Languages

Initiative	Primary Purpose	Supports Turkic Languages	Main Limitation
NLLB	Machine Translation	Partial	Limited parallel data
BLOOM	Multilingual LLM	Partial	Data imbalance
FLORES-200	Evaluation Benchmark	Yes	Limited benchmark tasks
Common Voice	Speech Corpus	Partial	Uneven language coverage
OSCAR / mC4	Training Corpus	Partial	Limited Turkic content

Conclusion

Large Language Models have significantly advanced multilingual Natural Language Processing; however, their performance remains highly dependent on the availability of digital and linguistic resources. This study has shown that many Turkic languages continue to face substantial challenges due to limited digital content, insufficient annotated datasets, agglutinative linguistic structures, and the lack of standardized evaluation benchmarks. These limitations reduce the quality of AI-generated outputs and restrict access to advanced language technologies for millions of native speakers. [2][10][15]

The findings indicate that improving AI support for Turkic languages requires more than increasing model size or computational capacity. Sustainable progress depends on expanding high-quality digital content, developing open linguistic resources, creating benchmark datasets, and encouraging collaborative community participation in language resource development. Strengthening these components will improve the

representation of Turkic languages in future multilingual AI systems and contribute to more inclusive language technologies. [3][14][16]

Future research should focus on building larger open corpora, improving language-specific evaluation benchmarks, and developing multilingual datasets that better represent underrepresented Turkic languages. Such efforts will help reduce the resource gap between high- and low-resource languages while promoting broader accessibility and long-term preservation of linguistic diversity in the era of Artificial Intelligence. [4][11][13]

References:

[1] OpenAI. *GPT-4 Technical Report*. arXiv:2303.08774, 2023.

[2] S. Tukenov et al. *SozKZ: Training Efficient Small Language Models for Kazakh from Scratch*. arXiv:2603.20854, 2026.

[3] X. Xue et al. *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*. Proceedings of NAACL, 2021.

[4] S. Bender et al. *On the Opportunities and Risks of Foundation Models*. arXiv:2108.07258, 2021.

[5] OpenAI. *GPT-4o System Card*. OpenAI, 2024.

[6] M. Joshi et al. *State and Challenges of NLP for Low-Resource Languages*. Communications of the ACM, 2024.

[7] Wikipedia contributors. *Turkic Languages*. Available: https://en.wikipedia.org/wiki/Turkic_languages

[8] Ethnologue. *Languages of the World*. SIL International, 2025.

[9] W3Techs. *Usage Statistics and Market Share of Content Languages for Websites*. Available: https://w3techs.com/technologies/overview/content_language

[10] X. Song et al. *Tokenization Matters: Challenges for Morphologically Rich Languages*. ACL Findings, 2023.

[11] Meta AI. *No Language Left Behind: Scaling Human-Centered Machine Translation*. Nature, 2022.

[12] Hugging Face. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 2022.

[13] Meta AI. *FLORES-200: Benchmarking Machine Translation Across 200 Languages*. 2022.

[14] Mozilla Foundation. *Common Voice Dataset*. Available: https://commonvoice.mozilla.org

[15] A. Conneau et al. *Unsupervised Cross-lingual Representation Learning at Scale*. ACL, 2020.

[16] H. Nigatu, J. Canny, and S. Chasins. *Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study*. arXiv:2405.16669, 2024.