American Journal of Applied Science and Technology

VOLUME Vol.05 Issue 12 2025
PAGE NO. 89-94
DOI 10.37547/ajast/Volume05Issue12-14

Developing A Question-Answering System In The Uzbek Language Based On The Xlm-Roberta Model

Khujayarov I.Sh.

Department of Information Technology, Samarkand branch of Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, Uzbekistan

Ochilov M.M.

Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, Uzbekistan

Kholmatov O.A.

Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi, Uzbekistan

Received: 16 October 2025; **Accepted:** 08 November 2025; **Published:** 12 December 2025

Abstract: This article presents the issue of testing the XML-RoBERTa model for generating questions and answers in the Uzbek language. In the study, the XML-RoBERTa model was adapted to the Uzbek language from a dataset consisting of context, question and answer pairs in the Uzbek language and, as a result, a model was developed to generate a fragment of the answer to the user's question from the context. ROUGE, EM (Exact Match) and F1 control metrics were used to determine the performance of the model.

Keywords: XLM-RoBERTa, question-answer system, natural language processing, ROUGE, F1, EM, Uzbek language.

INTRODUCTION:

Nowadays, artificial intelligence has penetrated almost all spheres of human life. Natural language processing (NLP) technologies are one of the most important areas of artificial intelligence. In particular, models such as BERT, RoBERTa, and XML-RoBERTa, which are based on the rapidly developing transformer architecture, show high results in semantic analysis of text. Through this semantic analysis, it is possible to bring question-answering (Q&A) systems to high efficiency. Q&A systems are widely used to find the information a person needs and find answers to their questions. These Q&A systems can be used in call centers, public services, banking, and the education system to automatically respond to user questions in a manner appropriate to their needs [1].

Methods for creating question-answering systems are one of the areas of NLP, and they aim to build systems that automatically answer questions asked by people [2].

We can divide question-answering systems into two main types: open-domain and closed-domain Q&A systems [3].

Open domain question-answering systems are systems that cover a wide range of topics and are aimed at answering user questions from unlimited data sources [4]. These systems extract the necessary answer from websites, databases, or other sources that have answers to the user's question. Open domain question-answering systems mainly use modern natural language processing techniques, including Transformer models (such as BERT, LlaMA, GPT), web indexing algorithms, and data sorting methods. As an advantage of open domain systems, we can say that they can provide comprehensive answers on a topic. However, since they use many sources to find information, they require the use of high-level search and filtering algorithms to ensure

the reliability of the answer [1].

Closed domain question-answering systems are systems that are specific to a specific field or topic and specialize in answering questions related to that field. These systems use a database designed for a specific field and are mainly used in fields such as banking, medicine, law, and education. For example, closed domain systems in the banking sector answer customers' questions about that particular bank, which can be questions about loans, deposits, and payments [4].

Currently, there are many different models that answer user questions, such as GPT, GROK, deepseek, but research on developing systems that answer questions in the Uzbek language that are specific to a particular field is limited. The article describes the results and background of research conducted on developing a system that answers questions based on a specific context.

METHODOLOGY

This section describes the process of creating a question-and-answer (Q&A) system in the Uzbek language based on the XML-RoBERTa model, the architecture of the XML-RoBERTa model, the methods of preparing the dataset and training the model. ROUGE, EM (Exact Match) and F1 evaluation metrics were used to determine the performance of the model.

2.1. Dataset preparation. Due to the lack of existing question-and-answer sets in the Uzbek language, a data set in the Uzbek language was prepared. The question-and-answer set included more than 10,000 context (text containing information related to the question), question (question containing the answer in the context), answer (short text providing the most correct answer to the question) pairs. A sample of this data set is given in Table 1 below.

Table 1. Example of a dataset with context, question, and answer columns

Context	Question	Answer
Milliy toifali sport hakami toifasini berish xizmatini ko‘rsatuvchi vakolatli organ yoshlar siyosati va sport vazirligining viloyat boshqarmalari. (The authorized body providing services for awarding the national category of sports referee is the regional departments of the Ministry of Youth Policy and Sports.)	Viloyatdagi qaysi tashkilot milliy sport hakami toifasini beradi? (Which organization in the region awards the national sports referee category?)	Yoshlar siyosati va sport vazirligining viloyat boshqarmalari (Regional departments of the Ministry of Youth Policy and Sports)
Oliy, o‘rta maxsus va professional ta’lim hujjat dublikatini berish xizmatini Pedagogik innovatsiyalar, kasb-hunar ta’limi boshqaruv va malaka oshirish institutida ko‘rsatadi (The Institute of Pedagogical Innovations, Vocational Education Management and Advanced Training provides services for issuing duplicate documents for higher, secondary specialized and professional education)	Dublikatlar berish vakolati qaysi tashkilotda? (Which organization has the authority to issue duplicates?)	Pedagogik innovatsiyalar, kasb-hunar ta’limi boshqaruv va malaka oshirish institutida (At the Institute of Pedagogical Innovations, Vocational Education Management and Advanced Training)

Since the contexts and questions are intended to answer questions about the Uzbek language in public

services, frequently asked questions about public services were collected from the lex.uz platform, a resource that may provide answers to questions.

2.2. Types of models used. The BERT model was originally proposed by Google in 2018, and it brought a step forward in deep natural language learning tasks. Models in the BERT family include the following features:

In bidirectional context learning, the model understands the meaning of words in the text by taking into account the words before and after that word.

The Transformer Encoder architecture analyzes the text semantically layer by layer. In the pre-training and fine-tuning tasks, the model is based on general language knowledge and is adapted for the selected domain. The BERT family includes several popular models (Figure 1).

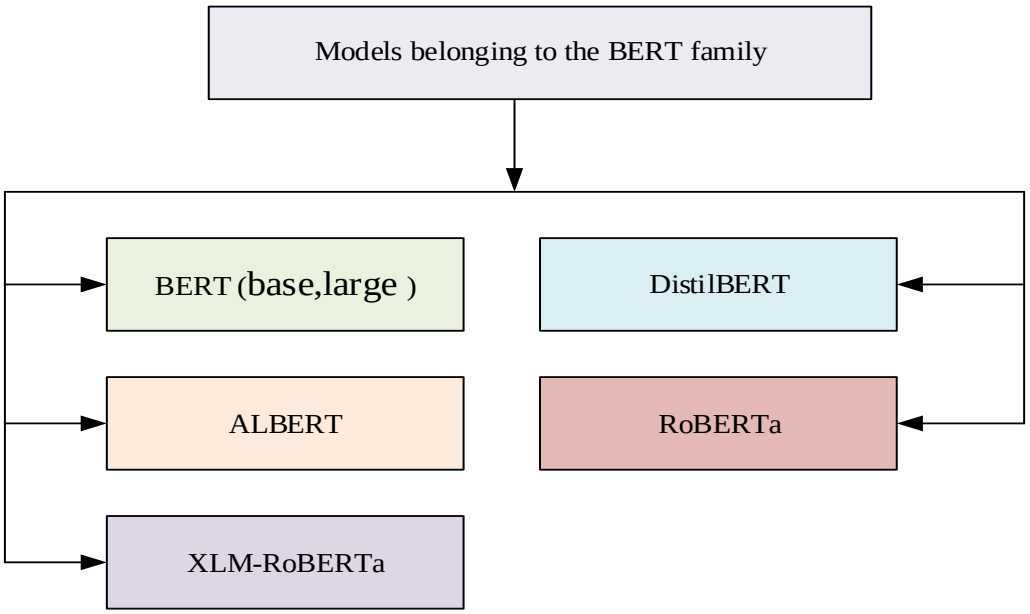

Figure 1. Popular models from the BERT family

Bidirectional Encoder Representations from Transformers (BERT) is a model introduced in October 2018 [5]. This model learns text as a sequence of vectors through self-supervision. The BERT model is trained by predicting a hidden token and the next sentence. With this training, BERT learns the contextual, hidden information of tokens, like ELMo and GPT-2 [6]. BERT was initially implemented in English in two model sizes, BERT (Base -110 million parameters) and BERT (Large 340 million parameters). Both were trained on the Toronto BookCorpus (800M sentences) and the English Wikipedia (2500M sentences).

2.3. Model architecture used. The models in the BERT family are of the "encoder-only" transformer architecture. This model consists of 4 modules: tokenizer (this module converts a portion of English text into a sequence of integers, i.e. tokens), embedding (this module converts a sequence of tokens into a series of real-valued vectors representing tokens. This represents the transformation of discrete token types into a low-dimensional Euclidean space). Encoder (encoder) consists of a set of self-focusing transformer blocks. The task head module converts the final representation vectors into one-hot encoded tokens by generating an approximate probability distribution over the token types. It can be viewed as a simple decoder, decoding the latent image into token types, or as an "unlinking layer". This module is not often used for tasks such as answering questions or classifying emotions. [Figure 2] [7].

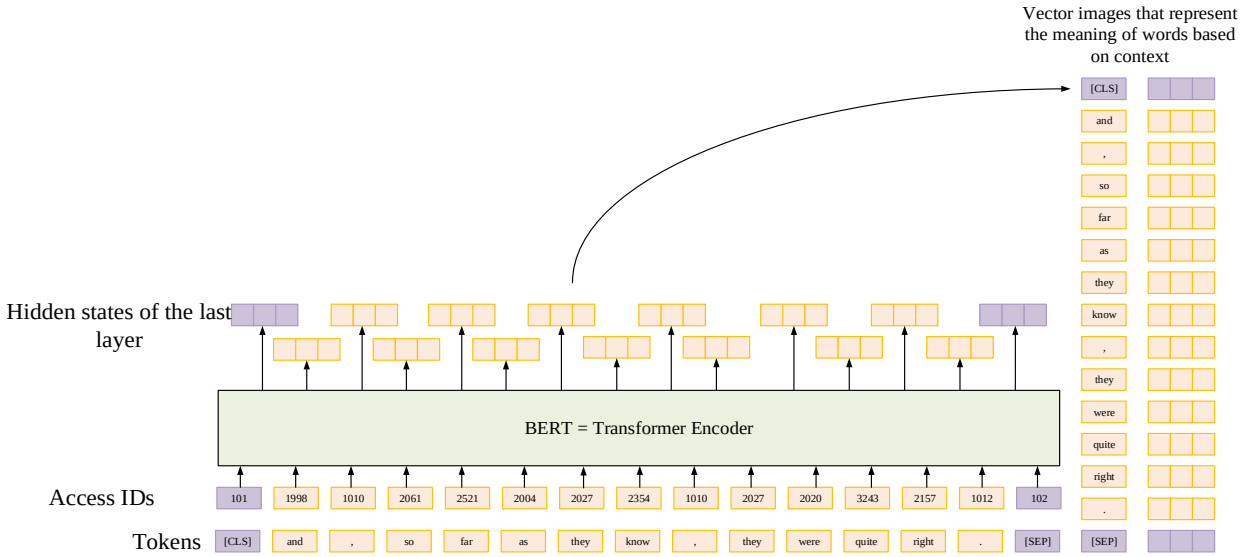

Figure 2. High-level schematic diagram of BERT.

It takes text, tokenizes it into a sequence of tokens, adds optional special tokens, and uses a Transformer encoder. The hidden states of the last layer can be used as contextual word input. As an advantage, we can say that it has very good general language understanding of the English language, including high results in many downstream tasks (classification, NER, QA). The base size is also suitable for use, and the large size is recommended for tasks where maximum accuracy is required. As a disadvantage, we can say that it does not support the Uzbek language well, and it requires a lot of resources for fine tuning tasks.

DistilBERT is a simplified version of the BERT model, but its performance is based on the BERT model. This model has about ~66M parameters, making it ~40% smaller and faster than BERT-base [8]. The advantages are that it is small and fast; it is very suitable for real-time applications and resource-constrained environments (mobile, CPU), and it retains most of the accuracy of BERT-base in many downstream tasks. The disadvantages are that the BERT model performs slightly worse than other larger models in high-quality tasks.

The main idea of the RoBERTA model is to optimize the training mode of the BERT model while preserving its architecture — more data, larger batches, longer training, dynamic masking changes. As a result, the RoBERTA model performed better than the BERT model on a number of tasks [9]. The RoBERTa model, like the BERT model, has base and large types. The RoBERTa base model has 12 layers and ~125M parameters. The RoBERTa large model consists of 24 layers and 355M parameters. As an advantage, we can say that it gives better results than BERT because it is trained with more data. As a disadvantage, we can say that it requires large computational resources, and it requires special training for certain tasks for a specific language.

The main idea of the ALBERT model is to reduce the parameters of larger models — this is done by connecting weights between layers [10]. The advantages are that it significantly reduces the memory and training resource requirements, meaning that a large model can be trained and deployed with fewer resources. The disadvantages are that it can limit the performance of the model in some cases.

The XLM-RoBERTa model is an extension of the RoBERTa model with multilingual data. The XLM-R (XLM-RoBERTa) model supports 100 languages and was trained on ~2.5 TB of cleaned data [11]. This model also has XLM-R-base and XLM-R-large sizes, with XML-R-base typically having 12 layers and approximately ~270M parameters, while the XLM-R-large model has 24 layers and approximately ~550M parameters [11]. As an advantage, it can be applied to Uzbek language problems with less data due to its strong multilingual transfer capabilities. As a disadvantage, we can say that there is a risk of “overfitting” during fine-tuning for small language datasets.

Based on the above information, among these models of the BERT family, the XLM-RoBERTa model was selected to develop a QA system in the Uzbek language.

2.4. Evaluation metrics. Since models such as XLM-RoBERTa, BERT, and RoBERTa extract the answer to the question from the context, we use the Exact Match (EM) and F1 Score evaluation metrics to evaluate the effectiveness of the question-answering system based on the XLM-RoBERTa model.

Exact Match calculates the percentage of cases in

which the answer given by the system is 100% the same as the real answer. This metric strictly checks and considers the answer to be incorrect even if a word or character is left out in the answer.

$$EM = \frac{Number\ of\ fully\ matched\ answers}{Total\ number\ of\ questions} \times 100 \quad (1)$$

This evaluation metric is recommended for evaluating question answering systems developed using the XML-RoBERTa model, as it distinguishes between model responses and actual responses. It consists of the harmonic mean of the Precision and Recall indicators.

The imbalance of these classes is the data of the newly created set of Uzbek punctuation marks. Therefore, the evaluation indicators of the work should be selected accordingly. For this, three indicators were used: precision - precion (shown in equation 1), recall - recall (shown in equation 2) and F1 scores for each label (shown in equation 3). These indicators are defined as follows:

$$Precision\ (P) = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1\text{–}Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2TP + FP + NP} \quad (3)$$

In Equations 1 and 2, TP represents the number of true positives, F1 represents the number of false positives, and FN represents the number of false negatives. Precision (P), recall (R), and F-scores (F1) are calculated taking into account the unbalanced class distribution. Therefore, using these metrics is considered a suitable assessment for the problem [12].

RESULTS

The XML-RoBERTa model separates the question from the context without generating the answer, and since there was an overfitting problem when fine-tuning the QA system in this model, we used the early stopping function to stop training when the validation loss value approached the training loss value or gave a large value, as a result, the number of training epochs (epoch) stopped at 5 with a training loss value of 0.686 and a validation loss value of 0.687. In addition, the adam optimizer was used as the optimization algorithm for training, the learning rate in training was set to 2e-1, and the batch size was set to 16. During the fine-tuning process, the data (dataset) was divided into 80% for training, 10% for validation, and 10% for testing. Exact match and F1-score evaluation metrics were used to determine the effectiveness of the QA system.

The experiment yielded the following results for the exact match and F1-score evaluation metrics (Table 2).

Table 2. Results determined by the Exact match and F1-score evaluation methods.

Baholash metodi	EM	F1-score
XML-RoBERTa	52.23	**78.54**
XML-RoBERTa (fine tuned)	56.4	56.2

According to the results obtained in this study, the EM evaluation metric showed a value of 56.4, and the F1-score evaluation metric showed a value of 56.2. Considering that the highest score of these results is 100, it can be said that the result obtained from the study showed a good result.

CONCLUSION

In the study, various models of the BERT family were studied and, based on the studied data, a question-answering system was developed based on the XML-RoBERTa model. Exact match and F1-score evaluation methods were used to evaluate the developed

system, and it gave accuracy indicators of 56.4 in the exact match evaluation metric and 56.2 in the F1-score evaluation metric. Considering that the best result in these evaluation metrics is 100, it can be concluded that a question-answering system in the Uzbek language based on the RoBERTa model can be developed and used in call centers. However, in some call centers, it may not give sufficient results in cases where the answer to the question needs to be obtained from several contexts.

REFERENCES

1. Muhammadjon Mahmudovich, Ochilov Mannon Musinovich, Xolmatov Orzimurod Abjalolovich, Narzullayev Oybek Otabek o‘g‘li. Vektor fazo modeli hamda jumlalar o‘xshashligi o‘lchovlariga asoslangan savol - javob tizimi ishlab chiqish. (2025). digital transformation and artificial intelligence, 3(1),23-30. https://dtai.tsue.uz/index.php/dtai/article/view/v3i14
2. Liu Y etcs. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019. Minneapolis, MN, USA. June 2–7, 2019.
3. Raffel, C. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Proceedings of the 37th International Conference on Machine Learning (ICML 2020).
4. Raffel, C., Shinn, E., Roberts, A., Lee, K., & Narang, S. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Proceedings of the 37th International Conference on Machine Learning (ICML), Long Beach, CA, USA, June 9–15, 2019.
5. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". October 11, 2018. arXiv:1810.04805v2
6. Ethayarajh, Kawin, How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. September 1, 2019. arXiv:1909.00512
7. Zhang Tianyi, Wu Felix, Katiyar Arzoo, Weinberger Kilian Q, Artzi Yoav. Revisiting Few-sample BERT Fine-tuning, March 11, 2021. arXiv:2006.05987
8. Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 1 Mar 2020. arXiv:1910.01108v4
9. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, OmerLevy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: ARobustly Optimized BERT Pretraining Approach. 26 Jul 2019. arXiv:1907.11692v1.
10. Zhenzhong Lan, Mingda Chen, Piyush Sharma, Google Research Sebastian Goodman, Radu Soricut. ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS. 9 Feb 2020. arXiv:1909.11942v6 [cs.CL]
11. Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., & Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116v2 [cs.CL], 8 April 2020.
12. Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. 2020. arXiv:1911.02116v2.