IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 4 July 2024, accepted 22 July 2024, date of publication 25 July 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3433426

RESEARCH ARTICLE

Development of a Geographical Question-Answering System in the Kazakh Language

ASSEL MUKANOVA[1], ALIBEK BARLYBAYEV[1,2], AIZHAN NAZYROVA[1,2], LYAZZAT KUSSEPOVA[1], BAKHYT MATKARIMOV[2], AND GULNAZYM ABDIKALYK[1]
[1]Higher School of Information Technology and Engineering, Astana International University, 010000 Astana, Kazakhstan
[2]Faculty of Information Technologies, L. N. Gumilyov Eurasian National University, 010008 Astana, Kazakhstan

Corresponding author: Aizhan Nazyrova (aizhan_nazyrova@aiu.edu.kz)

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19577922 The technology of creating an intelligent question answering system in the Kazakh language).

ABSTRACT The study presents a detailed framework designed to develop a Question-Answering System (QA System) for the Kazakh language, highlighting its importance in the field of Low Resource Languages (LRL) Text Processing. This effort aims to fill the gap in resources for languages that lack substantial digital tools. Specifically, the project focuses on geographical questions about Kazakhstan, aiming to enhance accessibility and understanding of the nation's geography. The challenges associated with LRL text processing are addressed through the creation of a question-answer corpus, training a Bidirectional Encoder Representations from Transformers (BERT)-based model, and evaluating the system using Bilingual Evaluation Understudy (BLEU) metrics. The endeavor begins with the careful compilation of a corpus containing 50,000 questions, which supports the subsequent development phases and ensures the creation of a robust QA System. In the second phase, a BERT model equipped with 91,821,056 parameters is trained, enhancing the model's ability to understand the complex linguistic nuances of the Kazakh language. The final phase involves a rigorous evaluation using BLEU metrics, where the system achieves an impressive average score of 0.9576. This score indicates a high level of agreement between the system-generated answers and the reference answers, demonstrating the system's effectiveness at interpreting and responding to queries about Kazakh geography. This study significantly contributes to the field by providing a systematic and nuanced approach to QA System development and underscores the model's effectiveness through thorough evaluation and comparative analysis.

INDEX TERMS Question answering system, Turkic languages, Kazakh language, transformers, BERT, BLEU score.

I. INTRODUCTION

The Question Answering System (QAS) constitutes a fundamental task within Natural Language Processing (NLP), designed to furnish precise responses in natural language to user queries [1]. Diverging from traditional search engines, a QAS directly formulates conclusive answers instead of presenting a compilation of hyperlinks [2]. This design enhances user-friendliness and operational efficiency within QAS. Leading search engines, such as Google or Bing [3], are progressively incorporating quality management methodologies into their search functionalities, aiming to elevate their overall intelligence. In the course of seeking information, web search engines merely guide users to potential answer locations, necessitating subsequent manual examination of search results for the actual answer. The prospect of an automated system capable of extracting or generating answers from retrieved documents, as opposed to merely presenting them

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

to the user, is both intriguing and promising. Consequently, QAS exemplifies the pursuit of natural language answers for user-generated inquiries.

The field of Question Answering (QA) resides at the convergence of Natural Language Processing (NLP), Information Retrieval (IR), Logical Reasoning, Knowledge Representation, Machine Learning, Semantic Search, and serves as a means to quantitatively assess the comprehension and reasoning prowess of any Artificial Intelligence (AI) system [4], [5]. Recent advancements, fueled by robust computational capabilities and the advent of cutting-edge deep learning algorithms, have propelled this interdisciplinary domain forward. Notably, contemporary deep learning models have exhibited superior performance compared to human counterparts in addressing single-paragraph question answering benchmarks, exemplified by achievements in widely recognized assessments such as the Stanford Question Answering Dataset (SQuAD) [5], [6], [7].

QA systems may be categorized according to the nature of the questions they endeavor to address. Factoid questions are those that typically seek straightforward, specific pieces of information [8]. QAS with factoid questions involves a system that is specialized in answering questions that have clear, factual answers, often derived from structured databases, documents, or knowledge sources. For example, a factoid question could be: "What is the capital of France?" The expected answer would be a single piece of factual information, in this case, "Paris."

QAS with Simple Reasoning questions refers to Question-Answering Systems that are designed to handle questions that involve basic logic or reasoning to arrive at an answer [9]. While factoid questions typically have straightforward and direct answers, simple reasoning questions may require the system to make basic inferences or deductions based on the information available. For instance, a simple reasoning question could be: "If it is raining outside, what should you carry with you?" The answer involves a logical inference, where the system needs to understand that carrying an umbrella is advisable when it's raining. QAS with simple reasoning questions often involves incorporating basic logical rules or patterns into the system to allow it to analyze and process information in a more nuanced way. These systems may use techniques from natural language processing and machine learning to understand the context, identify relationships, and provide reasoned responses to questions that go beyond straightforward factual queries.

QAS with Complex Reasoning questions refers to Question-Answering Systems that are designed to handle questions that involve more intricate logic, multi-step reasoning, and possibly deeper understanding of contextual information [10]. Unlike simple reasoning questions, which often require basic inferences, complex reasoning questions demand a more sophisticated level of cognitive processing. For example, a complex reasoning question could be: "If a train leaves City A at a certain time and travels at a certain speed, while another train leaves City B at a different time and speed, when and where will the two trains meet?" This question involves multiple steps of computation and reasoning, including calculating distances, speeds, and determining the point of intersection. To handle complex reasoning questions, QAS may utilize advanced techniques in artificial intelligence, machine learning, and natural language processing. These systems may need to understand and manipulate more abstract concepts, engage in logical deduction, and navigate through complex scenarios to provide accurate and meaningful answers. Such systems are valuable for applications that require a deeper level of comprehension and problem-solving ability, such as advanced decision support systems or expert systems.

QAS with Fusion List questions refer to combining or integrating information from multiple sources or types of data to generate a comprehensive answer [11]. In this context, fusion list questions might involve queries that require the QAS to synthesize information from diverse sets of data or knowledge bases. For example, a fusion list question could be: "List the top five cities in Europe with the highest GDP, considering both historical and current data." This question might require the QAS to integrate economic data from various time periods and cross-reference it with information about European cities. Handling fusion list questions would likely involve a QAS that can navigate and extract relevant information from different datasets, fuse them together, and present a consolidated and coherent response. This could involve techniques like data integration, cross-referencing, and contextual understanding.

QAS with Interactive Context questions refers to Question-Answering Systems that are designed to handle queries where the context is dynamic, evolving, or interactive [12]. In traditional question-answering scenarios, the system receives a single question and provides an answer based on the available information. However, with interactive context questions, the system may need to engage in a back-and-forth exchange with the user, taking into account the evolving context of the conversation. For example, in a traditional QAS scenario, a question might be: "What is the population of New York City?" The system provides a static answer based on the current information.

In contrast, an interactive context question might involve a series of inquiries and responses, with each exchange influencing the context. For instance:

User: "What is the population of New York City?"

System: "The population of New York City is approximately 8.4 million."

User: "And how has it changed in the last decade?"

System: "The population has increased by around 5% in the last decade."

Handling interactive context questions requires the QAS to maintain and update a dynamic context model, understanding the flow of the conversation and adapting responses based on previous interactions. This may involve memory management, context tracking, and real-time processing to ensure that the system can provide accurate and relevant information in an ongoing dialogue. Interactive context QAS is particularly relevant in conversational AI applications, virtual assistants, and chatbots where users may engage in multi-turn conversations, asking follow-up questions or seeking clarification.

QAS with Speculative questions refers to Question-Answering Systems that are designed to handle queries that involve speculation, hypothesis, or conjecture [13]. These questions typically inquire about potential outcomes, possibilities, or hypothetical scenarios rather than seeking factual information. For example, a speculative question could be: "What would happen if humans could live on Mars?" This question prompts the QAS to consider hypothetical situations, potential challenges, and theoretical implications rather than providing a concrete, factual answer. Handling speculative questions requires the QAS to understand the context, infer potential outcomes, and engage in speculative reasoning. It may involve analyzing known facts, extrapolating from existing data, and considering various hypothetical scenarios to provide informed responses. Speculative question answering is important in contexts where users are exploring possibilities, brainstorming ideas, or considering future scenarios. QAS with speculative capabilities can contribute to decision-making processes, scenario planning, and creative problem-solving by exploring alternative futures and potential consequences.

Nevertheless, the Question Answering System (QAS) process can be categorically delineated into three distinct parts. Question analysis is the First part in the functioning of a QAS [14]. During this phase, the system interprets and processes the user's query to understand its meaning, intent, and context. The goal is to transform the natural language input into a format that the system can use to retrieve or generate an appropriate answer. The system analyzes the syntactic structure of the question, identifying the grammatical components such as nouns, verbs, and adjectives. Parsing techniques help break down the sentence into a structured representation, which aids in understanding the relationships between different elements. Understanding the meaning of words and phrases in the given context is crucial. This involves semantic analysis to capture the intended meaning of the question. Techniques like word sense disambiguation help the system choose the correct interpretation of ambiguous words based on context. Determining the user's intent is vital for providing relevant answers. Intent recognition involves identifying the primary goal or purpose of the user's query. Machine learning models or rule-based systems may be used to categorize the query into specific intents. Identifying entities (specific objects, locations, people, etc.) mentioned in the question is essential for extracting relevant information. Named Entity Recognition (NER) techniques are commonly used to identify and categorize entities within the text [15]. Understanding the context of the question may involve considering previous interactions in a conversation or referring to external knowledge bases. Contextual analysis helps the system provide more accurate and relevant answers based on the ongoing conversation or broader context. Different types of questions (e.g., factoid, reasoning, opinion-based) may require different approaches for answering. Identifying the question type helps the system apply the appropriate methods. Dealing with ambiguity is a crucial aspect of question analysis. Some questions may have multiple interpretations, and the system needs to handle such cases intelligently.

Document retrieval [16] is the Second part in QAS where the system searches for and retrieves relevant documents or passages that may contain information necessary to answer a user's query. This phase aims to locate potential sources of information based on the content of the user's question. Before retrieval, a corpus of documents needs to be indexed. Indexing involves creating a structured representation of the documents, which allows for efficient and fast retrieval. Techniques such as inverted indexing are commonly used, where terms in the documents are mapped to the documents containing them. Document and query representations are often transformed into numerical vectors for efficient computation. Techniques such as word embeddings or more advanced contextual embeddings like Bidirectional Encoder Representations from Transformers (BERT) may be employed [17]. Vectorization enables the comparison of the similarity between the query vector and document vectors. Various models are used to rank and retrieve documents based on their relevance to the user's query. Common models include Term Frequency-Inverse Document Frequency (TF-IDF) [18], Best Match 25 (BM25) [19], and neural network-based models. Machine learning models may be trained on labeled data to learn the relevance of documents to specific queries. Query expansion techniques may be applied to enhance retrieval by adding synonyms or related terms to the original query. This helps to capture a broader range of relevant documents. Thesauri or external knowledge bases may be used for expanding queries. Depending on the nature of the QAS and the user query, different retrieval strategies may be employed. These could include passage retrieval, document retrieval, or even more granular entity-level retrieval. Strategies may also consider factors like recency, popularity, or other relevance signals. Documents are scored based on their relevance to the user's query. Scoring mechanisms vary depending on the retrieval model used. Documents are ranked by their scores, and the top-ranked documents are considered as potential sources for generating an answer. Given the potentially large size of document corpora, retrieval systems need to be efficient. Techniques like caching, pruning, and parallel processing may be employed to speed up the retrieval process. Depending on the application, the document retrieval system may need to handle real-time updates to the document corpus. This is particularly relevant in dynamic environments where new information becomes available.

Extraction of Answer [20] is the Third part in a Question-Answering System (QAS) involves retrieving relevant information from the retrieved documents or passages to form a concise and accurate response to the user's query. This phase builds upon the document retrieval and question analysis stages. Once relevant documents or passages are retrieved, the system extracts the specific portions of text that are likely to contain the answer to the user's question. Techniques may include selecting sentences, paragraphs, or snippets from the documents. Coreference resolution [21] is the process of determining when different words or expressions in the text refer to the same entity. Resolving coreferences helps ensure the correct interpretation of pronouns or repeated references. For example, resolving that "he" refers to a specific person mentioned earlier in the text. Considering the context of the extracted text is crucial for generating a coherent and contextually relevant answer. The system may need to understand the meaning of words and phrases in the context of the entire document or conversation. If the system extracts multiple potential answers, a ranking mechanism may be applied to determine the most likely correct answer. The ranking may be based on factors like relevance, confidence scores, or the frequency of occurrence in the retrieved documents. Assigning confidence scores to extracted answers provides a measure of the system's certainty about the correctness of the response. Confidence scores help in presenting answers in a more nuanced manner, especially in cases where the system is uncertain. Contemporary search engines exhibit precise responsiveness to specific question types through advanced techniques. Hand-designed functional models, exemplified by end-to-end neural architectures, have demonstrated notable progress in acquiring nuanced linguistic features, yielding substantial performance improvements in established reading comprehension benchmarks.

Deep Learning (DL) methodologies, capable of assimilating vast amounts of information, excel in constructing intricate operational representations. DL has made considerable strides in handling various data types, including voice, text, and sequential data. SQuAD [22] was meticulously curated by formulating questions based on data sourced from Wikipedia. SQuAD 2.0 [17] amalgamates the 100,000 questions from SQuAD 1.1 with over fifty thousand deliberately unanswerable questions generated competitively by crowd workers, designed to pose challenges in appearing answerable [23]. Evaluating the performance of DL models on SQuAD 2.0 is imperative to discern the handling of questions without answers (out of scope). Additionally, there exist alternative datasets such as Freebase [24], Microsoft Machine Reading Comprehension (MS MARCO) [25], DBpedia [26], and CNN & Daily Mail [27]. These datasets not only enhance the efficiency of addressing Question Answering (QA) tasks on neural architectures but also serve as a robust test bed for evaluating the performance of these models.

The prevailing architectural paradigm of Convolutional Neural Networks (CNN) with an encoder and a decoder [28] is widely employed for constructing efficient NLP models. In contrast, BERT models exclusively leverage attention mechanisms, completely eschewing recurrence and convolutions. BERT, a pre-trained language model, generates profound bidirectional representations by employing bidirectional Transformers. This entails that each word in every layer of the network considers contextual information from both preceding and subsequent positions. The pre-trained BERT representations can be fine-tuned to achieve state-of-the-art performance across a diverse array of tasks. BERT utilizes masked language modeling during pre-training to establish deep bidirectional representations. Additionally, a binarized next sentence prediction is employed in the pre-training process to comprehend relationships between two sentences [29]. The supervised paradigm for training Machine Reading Comprehension (MRC) models [30] represents a promising stride towards comprehensive NLP systems. According to this analysis, attentive and discerning readers can disperse and integrate semantic information over extended distances. Transformers exhibit a capacity to capture relatively long dependencies, surpassing the limitations imposed by fixed-length perspectives in language modeling. Language modeling, a pivotal challenge, necessitates previous input values, especially in applications such as pre-training (unsupervised). While Long Short-Term Memory (LSTM) serves as a conventional solution in NLP [31], yielding commendable outcomes across diverse applications, the introduction of control mechanisms in LSTMs and the utilization of gradient clipping techniques may prove insufficient in fully addressing this challenge.

The authors of this paper [32] explore how leveraging pre-trained models can enhance the performance of QA systems, which are crucial in various applications such as search engines, virtual assistants, and customer support systems. In their research implemented a QA model using transfer learning principles. In this study [33] authors explore innovative methods for embedding features that can enhance the performance of question classification systems, which are crucial for various applications such as natural language processing (NLP), information retrieval, and automated question-answering systems. This research [34] provides valuable insights into the multifaceted challenges encountered in Turkish Natural Language Processing studies. By addressing linguistic complexities, improving data availability, overcoming technological limitations, and fostering collaboration within the research community, significant advancements can be made in this important area of study. This study [35] concluded that integrating sentiment analysis into Turkish question-answering systems represents a significant advancement in human-robot interaction technology. It opens up new possibilities for creating more empathetic and responsive robotic assistants capable of engaging with users on a deeper emotional level.

II. MATERIALS AND METHODS

A. DATASET

SQuAD is a widely used format for training and evaluating language models, particularly for question-answering tasks.

SQuAD provides a large collection of reading comprehension datasets that consist of questions posed by human annotators on a set of Wikipedia articles. The goal is for models to read the passage and provide accurate and relevant answers to the questions. SQuAD has become a benchmark in the NLP community, and researchers often use it to assess the performance of various language models, including Large Language Models (LLMs) like GPT-3, BERT, and others. The dataset covers a diverse range of topics and requires models to understand context, reason, and generate meaningful responses. Training on SQuAD helps language models improve their ability to comprehend and generate human-like responses in a question-answering context, making them more effective in a wide range of natural language understanding tasks.

SQuAD 1.0: The original version of the dataset, released in 2016. It contains over 100,000 question-answer pairs based on articles from English Wikipedia.

SQuAD 2.0: Released in 2018, SQuAD 2.0 introduced a new twist to the task. In addition to the questions that have answers in the provided passage, it includes unanswerable questions. This version encourages models to not only answer questions but also determine when a question does not have a clear answer in the given context.

SQuAD 2.1: This is a modified version of SQuAD 2.0. In SQuAD 2.1, the unanswerable questions from SQuAD 2.0 were removed. It was created to address concerns about the difficulty of the unanswerable questions.

It's worth noting that these versions were designed to challenge and evaluate models on different aspects of question-answering and reading comprehension. The introduction of unanswerable questions in SQuAD 2.0 aimed to make the task more realistic and closer to real-world scenarios.

Until today, a corpus of questions and answers in the Kazakh language in the SQuAD notation has not been created. But there are similar works that describe attempts to create a similar corpus in the Kazakh language to create question-answer systems. It's interesting to learn about the efforts to create a Kazakh language question-answering dataset, especially through machine translation and adaptation of existing datasets. These initiatives demonstrate the adaptability and creativity of researchers in addressing the challenges of limited language resources. Using the babI dataset translated into Kazakh [36] and machine translating the SQuAD 1.1 dataset to create the Kazakh QA dataset (KazQA) [37] are pragmatic approaches to overcome the absence of a dedicated Kazakh SQuAD-like dataset. The babI dataset, comprising 11,000 instances, underwent translation into the Kazakh language and was subsequently stored in a CSV file [36]. It's important to acknowledge the potential challenges and considerations that come with machine translation, such as ensuring the accuracy of the translated content, preserving semantic nuances, and addressing potential biases introduced during the translation process.

The linguistic corpus designed for LLM exploration focuses on the comprehensive examination of the Geography of Kazakhstan. This corpus boasts a substantial size, comprising a total of 1,451,581 words, thereby providing a rich and extensive dataset for linguistic analysis. Within this corpus, a remarkable component is the inclusion of 50,000 questions and answers, all meticulously crafted in the Kazakh language.The chosen case style for this corpus aligns with the standards of scientific and educational discourse, ensuring a meticulous and scholarly representation of geographical information pertaining to Kazakhstan. In addition to its substantive content, the corpus is meticulously annotated using the SQUAD notation. This markup framework enhances the accessibility and interpretability of the corpus, enabling efficient information retrieval and facilitating a structured format for posing questions and obtaining detailed responses within the context of Kazakhstan's geography. The adoption of such notation enhances the overall utility and versatility of the corpus for both linguistic modeling and scholarly inquiry.

B. MACHINE LEARNING MODEL

BERT is a state-of-the-art NLP model that has significantly advanced the field of machine learning and language understanding [38]. Developed by Google, BERT is a transformer-based neural network architecture designed for a wide array of language understanding tasks. The distinctive feature of BERT lies in its bidirectional context-awareness, as it considers both preceding and succeeding words in a given text during the training process. This bidirectional approach allows BERT to capture intricate contextual nuances and dependencies in language, contributing to its superior performance in tasks such as question answering, sentiment analysis, and named entity recognition. In scientific terms, BERT's effectiveness can be attributed to its pre-training strategy, wherein it is initially trained on vast corpora to learn general language representations. Subsequently, fine-tuning on specific tasks tailors its capabilities for more targeted applications. The attention mechanism within BERT facilitates the modeling of relationships among words, enabling a nuanced understanding of contextual semantics. BERT has emerged as a benchmark model in NLP, achieving remarkable results on various benchmark datasets. Its impact extends beyond the research community, influencing the development of subsequent transformer-based models.

The primary aim delineated in this referenced scholarly work [36] was the development and implementation of a deep learning model specifically designed for the Kazakh language within the specialized domain of Question Answering systems. This objective was realized through the utilization of a translated iteration of the babI dataset procured from Facebook, employed as a foundational resource for training the End-to-end Memory Neural Network model. The empirical findings yielded notable results, highlighting the model's commendable performance, as evidenced by an accuracy rate of 92% and an F1 score of 93%. These quantitative metrics unequivocally attest to the effectiveness of the employed methodologies, including the extensive preprocessing procedures systematically applied to the dataset. The latter encompassed a meticulous consideration of the nuanced intricacies intrinsic to the Kazakh language. However, it is imperative to underscore a pivotal limitation within the scope of this investigation. The Question Answering System featured in this study exhibits a binary response capability, restricted to the options of "yes" or "no." This constraint represents a critical delineation of the study's parameters and necessitates careful consideration in the interpretation and assessment of the model's applicability and limitations within the defined context.

In this research [37], the pre-training regimen for ALBERT encompasses two primary tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The foundational goal of this phase is to instill in the model an adept understanding of the distinction between contextual and linguistic elements. To actualize this objective, the ALBERT model concurrently engages in training on two unsupervised tasks: MLM and Sentence Order Prediction (SOP). Within the MLM task, the model confronts masked sentences with the objective of predicting the output of the concealed tokens. This undertaking facilitates ALBERT in acquiring a nuanced comprehension of bidirectional context within sentences. Simultaneously, the SOP task involves the model processing two consecutive sentences from the same class, initially labeling them as positive, and subsequently swapping them and assigning a negative label. This procedural mechanism serves to cultivate the model's proficiency in discerning inter-sentence coherence. Transitioning to the second phase, the model undergoes fine-tuning tailored for a specific task, such as question answering. This involves the replacement of fully connected layers in the model with new layers expressly designed for answering the targeted questions. Subsequent supervised training ensues using a question answering dataset. During this training regimen, the output parameters are acquired de novo, while the remaining components of the model undergo fine-tuning to expedite the training process. Among the pre-trained models, the KazBERT model, trained on cleansed CommonCrawl data in the Kazakh language, achieved the most favorable outcomes. For the test dataset, the KazBERT model demonstrated an Exact Match (EM) accuracy of 62.87 and an F1 score of 77.89. These quantitative metrics serve as a testament to the efficacy and proficiency of the KazBERT model in the context of the specified language and task.

This paper [39] delineates the research and development endeavors directed towards a question-and-answer system rooted in the BERT model tailored for the Kazakh language. This work introduces an innovative amalgamation of normative and statistical methodologies for analyzing questions inherent to closed subject areas within agglutinative languages, specifically catering to the linguistic nuances of Kazakh. The process of question analysis involves two key components: focusing and classification. In the focusing phase, the authors enlisted the expertise of several specialists guided by the principles outlined in the Kazakh language Frequently Asked Questions (FAQ). The subsequent classification of questions was executed through the utilization of a rule-based classifier, incorporating phrases deemed unsuitable for each class. Fundamental models were employed for both focusing and classification, and a comparative analysis of their outcomes is presented. Throughout the course of this research, a corpus comprising 60,000 sentences was meticulously compiled and translated into Kazakh. The study example who achieved the highest F1 score of 88.0% and a matching accuracy score of 71.2% by responding to 257 questions in the test database were selected. The BERT-based methodology demonstrated a noteworthy performance, yielding an F1 score of 78.1% and a matching accuracy of 63.0%. This outcome starkly outperformed the baseline method, which recorded an F1 score of 38.0% and a matching accuracy of 22.2%. Although the F1 score was 9.9% lower than the measured examples and 8.2% lower in matching accuracy, the BERT-based method nonetheless established its superiority in both metrics.

C. MODEL EVALUATION METRIC

Bilingual Evaluation Understudy (BLEU) is a metric used for evaluating the quality of machine-generated translations in the field of natural language processing and machine translation [17]. BLEU is particularly valuable for assessing the performance of automated translation systems by comparing their output to one or more reference translations. The BLEU score is based on the comparison of n-grams (contiguous sequences of n items, typically words) between the candidate translation and the reference translations. The metric computes a precision-based score, considering how many n-grams in the candidate translation overlap with those in the reference translations. BLEU also includes a brevity penalty to address the issue of shorter translations being favored.

The BLEU score is calculated using the formula:

$$BLEU_{score} = BP * exp(\sum\nolimits_{n=1}^{N} w_n \log p_n) \tag{1}$$

where:

BP	–is the brevity penalty.
N	–is the maximum n-gram order considered.
wn	–is the weight assigned to the precision of n-grams.
pn	–is the modified precision for n-grams.

The BLEU score ranges from 0 to 1, with higher scores indicating better translation quality. It is widely used in research and development to assess the performance of machine translation systems and to compare different approaches in the field.

III. EXPERIMENTS AND DISCUSSIONS

In Figure 1, the depicted pipeline delineates the sequential progression originating from the input dataset, traversing

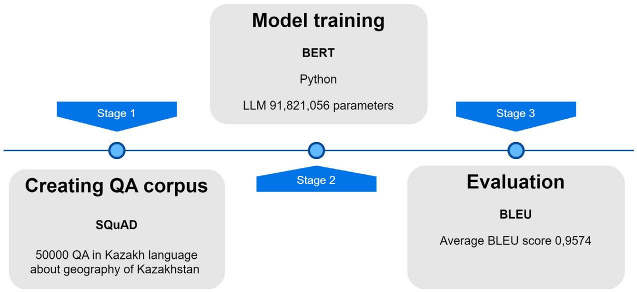


FIGURE 1. Pipeline of question answering system in Kazakh language.

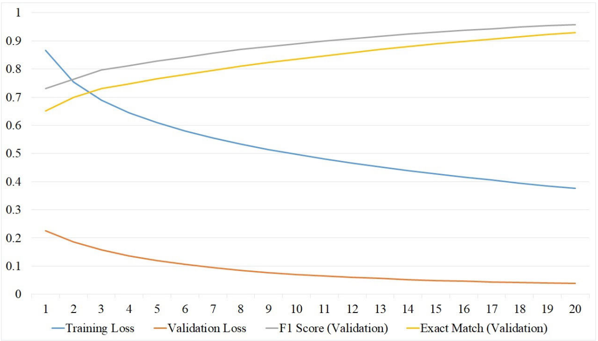


FIGURE 2. Model accuracy.

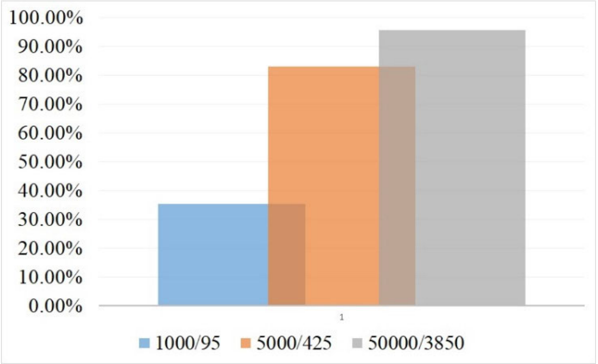


FIGURE 3. Evolving pattern of BLEU score augmentation in response to a significant expansion of the dataset.

through the BERT-based question-answering model, and culminating at the evaluation stage, where the BLEU metric is applied for quantitative assessment.

In the initial phase of the pipeline, the primary objective involves the generation of a question-answer corpus formatted in the SQuAD notation. This corpus specifically comprises 50,000 question-answer pairs in the Kazakh language, with a thematic focus on the geographical domain of Kazakhstan. The meticulous curation of this corpus is paramount for subsequent stages in the development of a robust Question-Answering system.

The second stage of the pipeline centers around the training of a BERT model. Following the training process, the Language Model derived from BERT encompasses an extensive parameter count, boasting a total of 91,821,056 parameters. The magnitude of parameters signifies the model's capacity to capture intricate linguistic nuances and contextual relationships within the Kazakh language, thereby enhancing its effectiveness in generating accurate and contextually appropriate responses.

The third and final stage of the pipeline involves the evaluation of the QA system through the application of BLEU metrics. BLEU serves as a fundamental measure to assess the quality of machine-generated text by comparing it to reference translations. In this context, the QA system is evaluated using BLEU metrics, with an average BLEU score of 0.9576. This high BLEU score indicates a commendable level of alignment between the system-generated answers and the reference answers, affirming the system's proficiency in comprehending and responding to queries within the domain of Kazakh geography.

Pipeline encompasses the meticulous creation of a domain-specific QA corpus, the training of a BERT model with a substantial parameter count, and a robust evaluation process utilizing BLEU metrics. This systematic approach ensures the development of an advanced and accurate QA system tailored to the nuances of the Kazakh language, particularly within the geographical context of Kazakhstan.

Figure 2 shows the performance of a transformer model during training and validation phases for the development of a question and answer system in the Kazakh language.

Figure 2 shows the performance of a transformer model during training and validation phases for the development of a question and answer system in the Kazakh language. The figure includes information for each epoch, tracking key metrics such as Training Loss, Validation Loss, F1 Score on the Validation set, and Exact Match on the Validation set. Figure 2 encapsulates the progressive evolution of a transformer model across multiple training epochs, exhibiting metrics indicative of its learning and generalization capabilities. 0.376098443 training loss value signify improved model convergence. 0.037803103 validation loss indicate its ability to generalize to unseen data. 0.958 F1 scores denote superior model efficacy. 0.93 Exact Match denote the proportion of model-generated predictions that precisely match the ground truth on the validation set, this metric serves as a binary indicator of accuracy. Throughout the training process, a discernible trend emerges, characterized by a consistent decrease in both training and validation losses. Simultaneously, metrics such as F1 Score and Exact Match on the validation set exhibit a persistent upward trajectory. This confluence of trends suggests that the model progressively refines its understanding and prediction capabilities with each successive epoch.

The BLEU metric is commonly used to evaluate the quality of machine-generated text, including QAS. It assesses the similarity between the generated text and a reference text by considering n-gram precision. Table 1 shows the QAS estimates. Figure 3 presents an illustration depicting the dynamics of the increment in BLEU score concerning a substantial augmentation in the dataset comprising both questions and corresponding answers.

TABLE 1. Evaluating the results of generating QAS using the BLEU metric.

Question Answer	Current BLEU score	General BLEU Score	Average BLEU Score
Question (803) Зортөбе тауының батысы мен шығысынан қандай өзеннің салалары ағып өтеді? {'Reference Answer: ': [{'text': 'Қарақатын өзенінің салалары ағып өтеді.', 'answer_start': 274}], 'Predicted_Answer': 'Қарақатын өзенінің салалары ағып өтеді.', 'BLEUScore': [{'google_bleu': 1.0}]}	1.0	768.981	0.95763549 25338474
Question (804) Зортөбе тауының қай бөлігінен Қарақатын өзенінің салалары ағып өтеді? {'Reference Answer: ': [{'text': 'батысы мен шығысынан', 'answer_start': 256}], 'Predicted_Answer': 'батысы мен шығысынан', 'BLEUScore': [{'google_bleu': 1.0}]}	1.0	769.981	0.95768818 47073128
Question (807) Зортөбе тауының қандай топырағында бұта аралас бетеге, қылқанбоз, т.б. астық тұқымдасты шөптесіндер өседі? {'Reference Answer: ': [{'text': 'Бозғылт, қызғылт қоңыр топырағында', 'answer_start': 308}], 'Predicted_Answer': 'Бозғылт, қызғылт қоңыр топырағында', 'BLEUScore': [{'google_bleu': 1.0}]}	1.0	772.633	0.95741446 64407968
Question (808) Зортөбе тауының етегі қандай мақсатта пайдаланылады? {'Reference Answer: ': [{'text': 'мал жайылымына', 'answer_start': 404}], 'Predicted_Answer': 'мал жайылымына', 'BLEUScore': [{'google_bleu': 1.0}]}	1.0	773.633	0.95746717 1309063

The generated response achieved a commendable Current BLEU score of 1.0, signifying a precise match with the reference answer. This high level of correspondence underscores the accuracy of the model in conveying the intended information. The general BLEU score for this response sequence is 773.63, attesting to the model's consistent performance in providing accurate and contextually relevant answers across multiple questions. The average BLEU score for this set of responses is 0.957, indicating a high level of linguistic coherence and alignment with the reference answers. This reflects the model's proficiency in generating scientifically stylized content in response to diverse inquiries.

To solidify our findings, a comparative analysis was undertaken by juxtaposing our results with existing works in the domain of Question-Answering Systems (QAS). The comparative assessment encompassed counterparts of QAS in the Kazakh, Uzbek, and Turkish languages. The outcomes of this comparison are delineated in Tables 2 and 3. Table 2 elucidates the comparison of QAS models based on the F1 score, whereas Table 3 provides a comparative evaluation of QAS models utilizing BLEU metric estimates. Including QAS models in multiple languages allows for a more comprehensive evaluation of the performance of these systems. It enables researchers to explore how well the models generalize across different linguistic contexts and structures. The Kazakh, Uzbek, and Turkish languages are members of the Turkic linguistic group, sharing a common linguistic ancestry. These languages exhibit a significant degree of lexical overlap, featuring numerous identical words. Moreover, they demonstrate substantial similarity in their morphological rules governing the construction of words, phrases, and sentences. The linguistic affinity observed among these Turkic languages contributes to a shared linguistic framework, facilitating similarities in vocabulary and grammatical structures.

TABLE 2. Evaluation of diverse models based on the F1 score parameter.

Model	Authors	Language	F1 score
ALBERT	Shymbayev, M., & Alimzhanov, Y. (2023, May) [37].	Kazakh	84.6
BERT	Rakhimova, D. R., Kasymova, D. T., & Isabaeva, D. N. (2021) [39].	Kazakh	78,1
End-to-end Memory Neural Network	Bilakhanova, A., Ydyrys, A., & Sultanova, N. (2023) [36].	Kazakh	93
BERT	Gemirter, C. B., & Goularas, D. (2021) [40].	Turkish	68.34
BERTurk	Schweter, S. (2020) [41].	Turkish	80.87
ELECTRA	Soygazi, F., Çiftçi, O., Kök, U., & Cengiz, S. (2021, September) [42].	Turkish	91.365
UzBERT	Mansurov, B., & Mansurov, A. (2021) [43].	Uzbek	85.32
UZRoBERTa	Adilova, F., Davronov, R., & Safarov, R. (2023) [44].	Uzbek	96
Our model		Kazakh	95.8

TABLE 3. Comparison of QAS models using the BLEU metric.

Model	Language	BLEU
GPT-3.5-turbo [45]	Kazakh	12.94
mT5 [46]	Uzbek	2.75
uzT5 [46]	Uzbek	2.95
Our model	Kazakh	95.76

Table 2 delineates a meticulous evaluation of diverse models, employing the F1 score metric across multiple linguistic contexts. Each row corresponds to a specific model, furnishing details on the respective authors, language focus, and the attained F1 score. The linguistic domains covered include Kazakh, Turkish, and Uzbek. The F1 scores function as quantitative indicators of precision and recall, with elevated scores signifying superior overall performance. Our Model developed for the Kazakh language, the model attains a notable F1 score of 95.8. This exhaustive comparative analysis provides a nuanced insight into the performance metrics of each model, elucidating their effectiveness within distinct linguistic domains.

IV. CONCLUSION

This study introduces a methodical and comprehensive pipeline for the creation of a sophisticated Question-Answering System, specifically designed to accommodate the intricacies of the Kazakh language, with a focused application within the geographical context of Kazakhstan. The foundational stage involves the meticulous curation of a thematic question-answer corpus, comprising 50,000 pairs in Kazakh and centered around the geography of Kazakhstan. During the second phase, the BERT model undergoes training with a substantial parameter count of 91,821,056, showcasing its capability to capture nuanced linguistic intricacies and contextual relationships inherent to the Kazakh language. The performance evaluation of the transformer model during training and validation reveals a consistent decrease in both training and validation losses. Key metrics, including F1 Score and Exact Match on the validation set, exhibit an upward trajectory, denoting improved model convergence (0.376098443 training loss), the ability to generalize to unseen data (0.037803103 validation loss), superior model efficacy (0.958 F1 scores), and a high degree of accuracy (0.93 Exact Match). These trends illustrate the model's progressive refinement in understanding and predictive capabilities with each successive epoch. The subsequent comparative analysis, contrasting our QAS with existing models in Kazakh, Uzbek, and Turkish languages, provides nuanced insights into the performance metrics of each model. Our Kazakh-centric model achieves a notable F1 score of 95.8 and a BLEU score of 95.76, affirming its versatility and effectiveness across diverse linguistic domains. In conclusion, this research contributes a holistic approach to QAS development, emphasizing the significance of linguistic nuances, meticulous evaluation procedures, and comparative analyses. The demonstrated effectiveness of our model underscores its potential applicability in real-world scenarios, particularly within the unique geographical and linguistic context of Kazakhstan.

ACKNOWLEDGMENT

Special thanks to their colleagues and all participants of the project for their contribution to the development of technology for creating an intelligent system of questions and answers in the Kazakh language.

REFERENCES

[1] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 453–466, Nov. 2019.

[2] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023.

[3] Y. Zhao, J. Zhang, X. Xia, and T. Le, "Evaluation of Google question-answering quality," *Library Hi Tech.*, vol. 37, no. 2, pp. 312–328, Jun. 2019.

[4] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[5] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," 2018, *arXiv:1806.03822*.

[6] W. Wang, M. Yan, and C. Wu, "Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering," 2018, *arXiv:1811.11934*.

[7] A. H. Mohammed and A. H. Ali, "Survey of BERT (bidirectional encoder representation transformer) types," *J. Phys., Conf.*, vol. 1963, no. 1, Jul. 2021, Art. no. 012173.

[8] J. Lin, "An exploration of the principles underlying redundancy-based factoid question answering," *ACM Trans. Inf. Syst.*, vol. 25, no. 2, p. 6, Apr. 2007.

[9] J. W. F. da Silva, A. D. P. Venceslau, J. E. Sales, J. G. R. Maia, V. C. M. Pinheiro, and V. M. P. Vidal, "A short survey on end-to-end simple question answering systems," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5429–5453, Oct. 2020.

[10] R. Etezadi and M. Shamsfard, "The state of the art in open domain complex question answering: A survey," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 4124–4144, Feb. 2023.

[11] M. Yang, L. Chen, Z. Lyu, J. Liu, Y. Shen, and Q. Wu, "Hierarchical fusion of common sense knowledge and classifier decisions for answer selection in community question answering," *Neural Netw.*, vol. 132, pp. 53–65, Dec. 2020.

[12] Y. Hong, J. Wang, Y. Jia, W. Zhang, and X. Wang, "Academic reader: An interactive question answering system on academic literatures," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 9855–9856.

[13] S. A. Al-Azani and C. N. Mahender, "Arabic question-answering system based on deep learning models," in *Innovations in Machine and Deep Learning: Case Studies and Applications*. Cham, Switzerland: Springer, 2023, pp. 133–159.

[14] D. Diefenbach, V. Lopez, K. Singh, and P. Maret, "Core techniques of question answering systems over knowledge bases: A survey," *Knowl. Inf. Syst.*, vol. 55, no. 3, pp. 529–569, Jun. 2018.

[15] P. Banerjee, K. K. Pal, M. Devarakonda, and C. Baral, "Biomedical named entity recognition via knowledge guidance and question answering," *ACM Trans. Comput. Healthcare*, vol. 2, no. 4, pp. 1–24, Oct. 2021.

[16] Z. Abbasiantaeb and S. Momtazi, "Text-based question answering from information retrieval and deep neural network perspectives: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 11, no. 6, p. e1412, Nov. 2021.

[17] A. Barlybayev and B. Matkarimov, "Development of system for generating questions, answers, distractors using transformers," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 14, no. 2, pp. 1851–1863, Apr. 2024.

[18] M. Mohammed and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0230442.

[19] T. N. Manjunath, D. Yogish, S. Mahalakshmi, and H. K. Yogish, "Smart question answering system using vectorization approach and statistical scoring method," *Mater. Today, Proc.*, vol. 80, pp. 3719–3725, Jan. 2023.

[20] M. A. Calijorne Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 6, pp. 635–646, Jul. 2020.

[21] S. Bhattacharjee, R. Haque, G. M. de Buy Wenniger, and A. Way, "Investigating query expansion and coreference resolution in question answering on BERT," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, Jun. 2020, pp. 47–59.

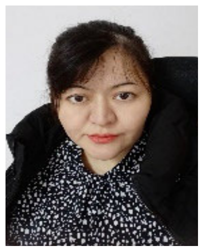
[22] M. A. Rondeau and T. J. Hazen, "Systematic error analysis of the Stanford question answering dataset," in *Proc. Workshop Mach. Reading Question Answering*, Jul. 2018, pp. 12–20.

[23] V. Krishnamoorthy, "Evolution of reading comprehension and question answering systems," *Proc. Comput. Sci.*, vol. 185, pp. 231–238, Jan. 2021.
[24] K. Jiang, D. Wu, and H. Jiang, "FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, Jun. 2019, pp. 318–323.
[25] S. Zheng, M. Yang, Y. Huang, and P. Lin, "Utilization of question categories in multi-document machine reading comprehension," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[26] M. Dubey, D. Banerjee, A. Abdelkawi, and J. Lehmann, "LC-QuAD 2.0: A large dataset for complex question answering over Wikidata and DBpedia," in *Proc. Int. Semantic Web Conf.*, 2019, pp. 69–78.
[27] K. Liu and Y. Feng, "Deep learning in question answering," in *Deep Learning in Natural Language Processing*. Singapore: Springer, 2018, pp. 185–217.
[28] P. Wang, L. Ji, J. Yan, D. Dou, N. D. Silva, Y. Zhang, and L. Jin, "Concept and attention-based CNN for question retrieval in multi-view learning," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 1–24, Jul. 2018.
[29] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," 2020, *arXiv:2010.12309*.
[30] R. Baradaran, R. Ghiasi, and H. Amirkhani, "A survey on machine reading comprehension systems," *Natural Lang. Eng.*, vol. 28, no. 6, pp. 683–732, Nov. 2022.
[31] Y. Liu, X. Zhang, F. Huang, X. Tang, and Z. Li, "Visual question answering via attention-based syntactic structure tree-LSTM," *Appl. Soft Comput.*, vol. 82, Sep. 2019, Art. no. 105584.
[32] A. Aldabergen, B. Kynabay, and A. Zhamanov, "Question answering model construction by using transfer learning," in *Proc. 16th Int. Conf. Electron. Comput. Comput. (ICECCO)*, Nov. 2021, pp. 1–6.
[33] G. Haisa, G. Altenbek, H. Aierzhati, and K. Kenzhekhan, "Research on classification of kazakh questions integrate with multi-feature embedding," in *Proc. 2nd Int. Conf. Electron., Commun. Inf. Technol. (CECIT)*, Dec. 2021, pp. 943–947.
[34] K. Tohma and Y. Kutlu, "Challenges encountered in Turkish natural language processing studies," *Natural Eng. Sci.*, vol. 5, no. 3, pp. 204–211, Nov. 2020.
[35] K. Tohma, H. I. Okur, Y. Kutlu, and A. Sertbas, "Sentiment analysis in Turkish question answering systems: An application of human–robot interaction," *IEEE Access*, vol. 11, pp. 66522–66534, 2023.
[36] A. Bilakhanova, A. Ydyrys, and N. Sultanova, "Kazakh language-based question answering system using deep learning approach," *SDU Bull., Natural Tech. Sci.*, vol. 62, no. 1, pp. 113–121, 2023.
[37] M. Shymbayev and Y. Alimzhanov, "Extractive question answering for Kazakh language," in *Proc. IEEE Int. Conf. Smart Inf. Syst. Technol. (SIST)*, May 2023, pp. 401–405.
[38] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," *Frontiers Artif. Intell.*, vol. 6, Mar. 2023, Art. no. 1023281.
[39] D. Rakhimova, D. Kassymova, and D. Isabaeva, """Research and development of a question-answer system based on the BERT model for the Kazakh language," *Bull. Ser. Phys. Math. Sci.*, vol. 76, no. 4, pp. 119–127, Dec. 2021.
[40] C. B. Gemirter and D. Goularas, "A Turkish question answering system based on deep learning neural networks," *J. Intell. Syst., Theory Appl.*, vol. 4, no. 2, pp. 65–75, Sep. 2021.
[41] S. Schweter. (2020). *BERTurk-BERT Models for Turkish*. [Online]. Available: https://github.com/stefan-it/turkish-bert
[42] F. Soygazi, O. Çiftçi, U. Kök, and S. Cengiz, "THQuAD: Turkish historic question answering dataset for reading comprehension," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2021, pp. 215–220.
[43] B. Mansurov and A. Mansurov, "UzBERT: Pretraining a BERT model for Uzbek," 2021, *arXiv:2108.09814*.
[44] F. Adilova, R. Davronov, and R. Safarov, "UzRoberta: An Uzbek language pre-trained model," *Universum, Tech. Sci.*, vol. 115, no. 10, pp. 28–32, Oct. 2023.
[45] G. Tolegen, A. Toleu, R. Mussabayev, B. Zhumazhanov, and G. Ziyatbekova, "Generative pre-trained transformer for kazakh text generation tasks," in *Proc. 19th Int. Asian School-Seminar Optim. Problems Complex Syst. (OPCS)*, Aug. 2023, pp. 1–5.
[46] F. T. Adilova, R. R. Davronov, and S. I. Kushmuratov, "Tailoring MT5 for the Uzbek language: A compact model for NLP applications," *Int. J. Theor. Appl. Issues Digit. Technol.*, vol. 5, no. 3, pp. 7–16, 2023.

ASSEL MUKANOVA received the Ph.D. degree in computer science from L. N. Gumilyov Eurasian National University, Kazakhstan, in 2015. Currently, she is an Associate Professor with the Higher School of Information Technology and Engineering, Astana International University. Her research interests include artificial intelligence, natural language processing, ontology modeling, and intellectual knowledge base.

ALIBEK BARLYBAYEV received the Ph.D. degree in computer science from L. N. Gumilyov Eurasian National University, Kazakhstan, in 2015. Currently, he is the Director of the Research Institute of Artificial Intelligence, L. N. Gumilyov Eurasian National University. He is also an Associate Professor with the Higher School of Information Technology and Engineering, Astana International University. His research interests include NLP, the use of neural networks in word processing, smart textbooks, fuzzy logic, stock market price prediction, and information security.

AIZHAN NAZYROVA received the Ph.D. degree from the Educational Program in Computer Science, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan, in 2023. She is currently a Senior Lecturer with the Higher School of Information Technologies and Engineering, Astana International University. Her research interests include the development and design of information systems, the theoretical substantiation of computational algorithms, and artificial intelligence.



LYAZZAT KUSSEPOVA received the Ph.D. degree from the Educational Program in Computer Science, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan, in 2023. She is currently a Senior Lecturer with the Higher School of Information Technologies and Engineering, Astana International University. Her research interests include container technologies, cloud computing, information security, and machine learning.

BAKHYT MATKARIMOV received the Ph.D. and Dr.Sci. degrees from the Institute of Mathematics, Almaty, Kazakhstan.

His research interests include bioinformatics, computer vision, neural networks, and question-and-answer systems. In 2022, he was recognized as the Best Researcher of the Republic of Kazakhstan, which indicates the impact of his work on the scientific landscape. His commitment to the development of science was noted, in 2016, when he received the Award of the Republic of Kazakhstan for merits in this field.

GULNAZYM ABDIKALYK was born in Kazakhstan. She is currently pursuing the master's degree with the Higher School of Information Technology and Engineering, Astana International University. Her research interests include artificial intelligence, natural language processing, data mining, and data science.