Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi
Pamukkale University Journal of Engineering Sciences

Transformer-Based question answering systems for higher education: A comparative study of turkish and multilingual models

Yükseköğretim için transformer tabanlı soru-cevap sistemleri: Türkçe ve çok dilli modellerin karşılaştırmalı bir incelemesi

Halenur Sazak[1], Muhammed Kotan[1]*

[1]Department of Information Systems Engineering, Sakarya University, Sakarya, Türkiye.
halenurs@sakarya.edu.tr, mkotan@sakarya.edu.tr

Received/Geliş Tarihi: 09.10.2024 Revision/Düzeltme Tarihi: 14.07.2025 doi: 10.5505/pajes.2025.44459
Accepted/Kabul Tarihi: 05.09.2025 Research Article/Araştırma Makalesi

Abstract

This study presents a question answering system developed for higher education using transformer-based models. Five pretrained models were evaluated including BERTurk Base cased/uncased, ELECTRA-Turk, mBERT and XLM-R. The models were fine-tuned on the THQuAD dataset and tested on a frequently asked questions dataset constructed from official university sources and student queries. In addition to standard evaluation metrics such as Exact Match and F1 score, an extended evaluation approach was applied to better capture semantically appropriate answers. ELECTRA-Turk achieved the highest F1 score of 0.8936 and an Exact Match score of 0.8478. The results show that transformer-based approaches can effectively support automated question answering in academic domains and improve information access for students.

Keywords: Question-Answering System, Transformer Models, Information Retrieval, Turkish Language Models, Natural Language Processing

Öz

Bu çalışma, yükseköğretimde kullanılmak üzere geliştirilen bir soru cevap sistemi sunmaktadır. BERTurk Base cased/uncased, ELECTRA-Turk, mBERT ve XLM-R olmak üzere beş önceden eğitilmiş model değerlendirilmiştir. Modeller, THQuAD veri kümesi üzerinde ince ayarlanarak eğitilmiş ve resmi üniversite kaynakları ile öğrenci sorularından oluşturulan bir sıkça sorulan sorular veri kümesi üzerinde test edilmiştir. Tam Eşleşme ve F1 gibi standart metriklere ek olarak anlam açısından doğru yanıtları da dikkate alan genişletilmiş bir değerlendirme yöntemi uygulanmıştır. En yüksek F1 skoru 0.8936 ve Tam Eşleşme skoru 0.8478 ile ELECTRA-Turk modeli tarafından elde edilmiştir. Bulgular, transformer tabanlı yaklaşımların akademik ortamda otomatik soru cevaplama görevlerinde etkili olduğunu ve öğrenci odaklı bilgiye erişimi artırabileceğini göstermektedir.

Anahtar kelimeler: Soru-Cevap Sistemi, Transformer Modeller, Bilgi Edinme, Türkçe Dil Modelleri, Doğal Dil İşleme

1 Introduction

In today's digital educational landscape, the role of Artificial Intelligence (AI) in information retrieval is becoming increasingly significant. Question answering systems (QAS), particularly those based on deep learning and transformer architectures, have become valuable tools for enabling natural language interaction with structured and unstructured information sources.

QAS allows users to submit questions in natural language and receive direct answers, reducing the need for manual browsing or interaction with support personnel. They apply natural language processing (NLP) and machine learning techniques to interpret user queries and identify appropriate responses [1]. Moreover, these systems can generate personalized responses through individualized analyses.

Due to their adaptability, QAS are now widely used across domains such as customer support, healthcare, legal services, and education, where they help automate information access and reduce the burden on human staff. In the healthcare industry, QAS systems can be used to answer patients' medical queries and provide information on specific symptoms, making it easier to access healthcare services and information. Law firms and consulting businesses can use QAS to help clients with legal issues and research. The systems can be utilized in the industrial sector, and technical support can be used to answer users' inquiries while facilitating troubleshooting processes. In the context of education, QAS has the potential to support students, instructors, and administrators by addressing questions related to academic programs, course requirements, mobility procedures, regulations, and institutional services. Educational institutions increasingly use QAS to streamline communication, reduce administrative workload, and improve the accessibility of reliable information.

This study presents the design and evaluation of a Turkish-language question answering system for use in higher education institutions. The system is based on transformer-based language models and is intended for prospective applicants, currently enrolled undergraduate and graduate students, and university personnel. Its objective is to provide accurate answers to common questions efficiently, while also reducing repetitive workloads for administrative staff.

The motivation behind this study stems from common challenges such as students lacking sufficient information about departments, communication difficulties, and current students not receiving adequate and timely responses to their academic queries. The proposed QAS aims to address these issues by facilitating online rapid access to information, optimizing the information retrieval process, and enabling

*Corresponding author/Yazışılan Yazar

users to access needed information anytime and anywhere. Additionally, the system is designed to provide specialized support for higher education institutions, enhancing their overall efficiency by allowing staff and students to access relevant information swiftly.

The proposed system first receives and analyzes the user's question and its context. It then performs a targeted search across predefined links and documents to gather the relevant information. Upon retrieving the context, the system divides it into manageable chunks and searches within these sections to identify the most suitable answer. This process allows for a more accurate extraction of information, particularly from lengthy or complex documents. The most relevant answer span is selected using a transformer-based model and is stored for potential reuse, allowing the system to improve over time based on recurring questions.

Unlike traditional keyword-based search systems, the proposed approach benefits from contextual understanding by identifying answer spans based on semantic meaning. To evaluate performance, we fine-tuned five pretrained transformer models (BERTurk Base, BERTurk Uncased, ELECTRA-Turk, mBERT and XLM-R) using the THQuAD dataset [2] and tested them on a university-specific dataset composed of frequently asked questions. In addition to standard metrics such as exact match and F1 score, we employed a relaxed evaluation protocol that allows semantically accurate but lexically non-identical answers to be accepted.

The remainder of the paper is organized as follows. Section 2 reviews related work in the field of question answering. Section 3 presents the methodology, including the models used, dataset characteristics, and evaluation metrics. Section 4 reports and analyzes the experimental results obtained from the fine-tuned models. Section 5 concludes the study with a summary of findings and implications for future applications.

2 Literature review

As AI technologies continue to advance, there has been a significant increase in research focused on designing intelligent systems that can handle complex, domain-specific queries. QAS have received considerable attention for their ability to retrieve accurate answers from structured or unstructured text.

QAS offers significant potential to facilitate learning and enhance student engagement by providing immediate, accurate responses to queries [3]. Smutny and Schreiberova [4] examined the use of educational chatbots integrated into platforms like Facebook Messenger. Similarly, Okonkwo and Ade-Ibijola [3] presented a systematic review of 53 articles focusing on the use of Chatbots in education. The review provides an overview of prior studies, highlighting the benefits and challenges associated with Chatbot implementation in educational settings and identifying future research.

Using Ontology and SWRL (Semantic Web Rule Language) technology, Yousuf and Imran [5] developed an automated QA system for academia, achieving a 48% F-Measure accuracy in responding to academic queries. Fulmal et al. [6] employed Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) neural networks to determine question similarities and retrieve the best answers based on the highest similarity scores, with BiLSTM showing higher accuracy. In another paper, Madabushi et. al [7] combined fine grained Question Classification with a Deep Learning model for Answer Selection. Clarizia et al. [8] highlighted the development of a prototype Chatbot designed for the educational domain, focusing on the creation of specific architecture and a model to manage communication and provide accurate answers to students. The system employs NLP techniques and domain ontologies to detect questions and deliver appropriate responses to students.

Durmus et. al. [9] proposed a new metric, FEQA, based on automatic question answering to assess the faithfulness of summaries compared to their source documents, correlating more closely with human scores, especially for highly abstract summaries. Ardaç and Erdoğmuş [10] developed question-answering models utilizing text mining and deep networks, including a specific model, BERTDuQuA (BERT Düzce University Question Answering), designed for answering questions in Turkish. Derici et al. [11] introduced a question-answering framework for student queries in Turkish, using reliable resources and multi-document summaries, with multilingual support through translation, achieving a 50–60% success rate in providing accurate answers. Akyon et al. [12] used a fine-tuned multilingual T5 transformer in a multitask setting for question answering, generating and answer extraction tasks using Turkish QA datasets.

While the advancements are promising, some studies have raised ethical concerns and challenges related to AI use in education. Kooli [13] explores the ethical challenges and opportunities of integrating AI systems and chatbots into education and research, emphasizing their potential while highlighting the need for creative assessment methods, ethical guidelines, and sustainable adaptation to technological advancements.

Recent studies have also used transformer-based models for complex tasks like sustainability report analysis. In a study by BERTurk and BM25 models, these models were employed for the automatic categorization of Turkish user comments on e-commerce platforms, achieving high accuracy in categorizing comments [14]. Similarly, the Retrieval Augmented Generation (RAG) approach has been applied to sustainability reports to evaluate the ESG factors of companies in the BIST Sustainability-25 index, with BM25 outperforming BERTurk in terms of performance for information extraction from documents [15]. Furthermore, BERTurk-based models have been effectively fine-tuned for question-answering (QA) tasks in Turkish medical texts. For example, a dataset of 8200 question-answer pairs, gathered from Turkish Wikipedia and the Thesis Center of the Higher Education Council of Turkey, was used to fine-tune BERTurk models. The models were evaluated using metrics such as Exact Match (EM) and F1 score, with the BERTurk (cased, 128k) model achieving the best results, showing the effectiveness of pre-trained models for Turkish medical text processing and QA tasks [16].

These recent developments highlight the significant role of BERTurk and other transformer-based models in improving question-answering systems and text categorization tasks in Turkish.

This study systematically evaluates multiple transformer models on both a large-scale Turkish QA dataset (THQuAD)

and a custom dataset of university FAQs. We further introduce a relaxed evaluation scheme to account for semantically correct but non-identical predictions.

3 Methodology

Transformer-based models have become the foundational architecture for state-of-the-art NLP tasks [17]. Introduced by Vaswani et al. [18], the transformer architecture relies on a self-attention mechanism, which enables the model to consider all positions in the input sequence simultaneously. This design facilitates efficient training and allows for effective modeling of long-range dependencies in text [19]. An overview of the original transformer architecture is presented in Figure 1.

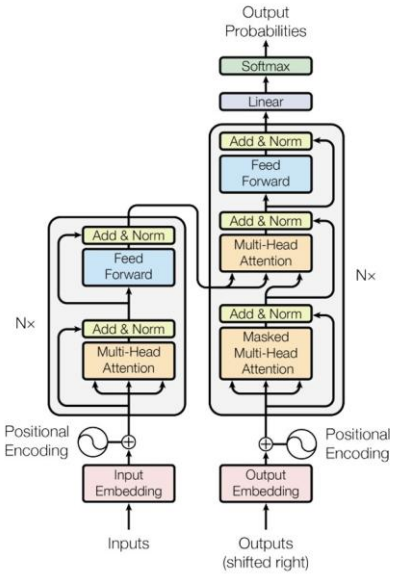

Figure 1. Transformers Architecture [18]

The transformer consists of two main components: an encoder and a decoder. The encoder processes input sequences through a stack of self-attention and feedforward layers to produce contextualized token representations. The decoder generates output sequences by attending to both the encoder outputs and previously generated tokens.

The Transformer architecture consists of multiple stacked encoder layers, each incorporating key components to enhance contextual learning. At the core of each encoder layer is the multi-head self-attention mechanism, which allows the model to compute relationships between all tokens in a sentence simultaneously. This is achieved through Query, Key, and Value matrices, enabling the model to capture both local and long-range dependencies. Following the self-attention layer, a feed-forward network (FFN) is applied to each token independently, consisting of two linear transformations with a non-linear activation function. This helps the model learn complex representations beyond simple attention weights. To ensure stable training and efficient information flow across layers, residual connections are used alongside layer normalization, which normalizes activation and mitigates gradient vanishing issues.

This study proposes a transformer-based QAS for Turkish-language educational content. Five pretrained models were selected for comparison, including both Turkish-specific and multilingual architectures. All models were fine-tuned using the THQuAD dataset, which follows a span-based SQuAD-style format. Following fine-tuning, the models were tested on a custom university FAQ dataset composed of real academic and administrative questions. A chunking strategy was implemented to handle long context passages by dividing them into overlapping segments. To assess performance, the study uses Exact Match, F1 Score and a relaxed evaluation metric that accounts for semantically correct but non-identical answers. A summary of the proposed pipeline is illustrated in Figure 2.

3.1 Pretrained Transformer Models

In late 2018, a team at Google AI Language introduced a new language model called BERT (Bidirectional Encoder Representations from Transformers) [20]. BERT is a model based on the Transformer architecture, which has significantly advanced state-of-the-art performance in NLP tasks. Unlike traditional language models, which process text sequentially, BERT uses a bidirectional approach that considers both the left and right context of all words in a sentence [21]. This ensures that the target word is influenced not only by preceding words but also by the context of the entire sentence. Bert's structure for fine-tuning QA is shown in Figure 3.

Training the BERT model involves pre-training unlabeled data and additional training stages on labelled data for specific application problems. The model's architecture is based on multi-layer bidirectional transformers [23],[24]. BERT utilizes the encoder mechanism of Transformer architecture to generate a language model. It processes token sequences by embedding them into vectors and then processing them through the neural network, resulting in contextualized representations for each token. Additionally, BERT utilizes token and segment embeddings, where each token is represented by a high-dimensional vector, and segment embeddings help distinguish between different sentences in tasks like Next Sentence Prediction (NSP). During pretraining, BERT is optimized using Masked Language Model (MLM) and NSP on large-scale corpora, allowing it to develop deep contextualized word representations. By using this architectural design, BERT effectively models complex linguistic patterns, making it highly efficient for various NLP applications [25].

RoBERTa (Robustly Optimized BERT Approach), an improved variant of the original BERT model, was introduced by Liu et al. [26]. It is based on the same transformer encoder architecture

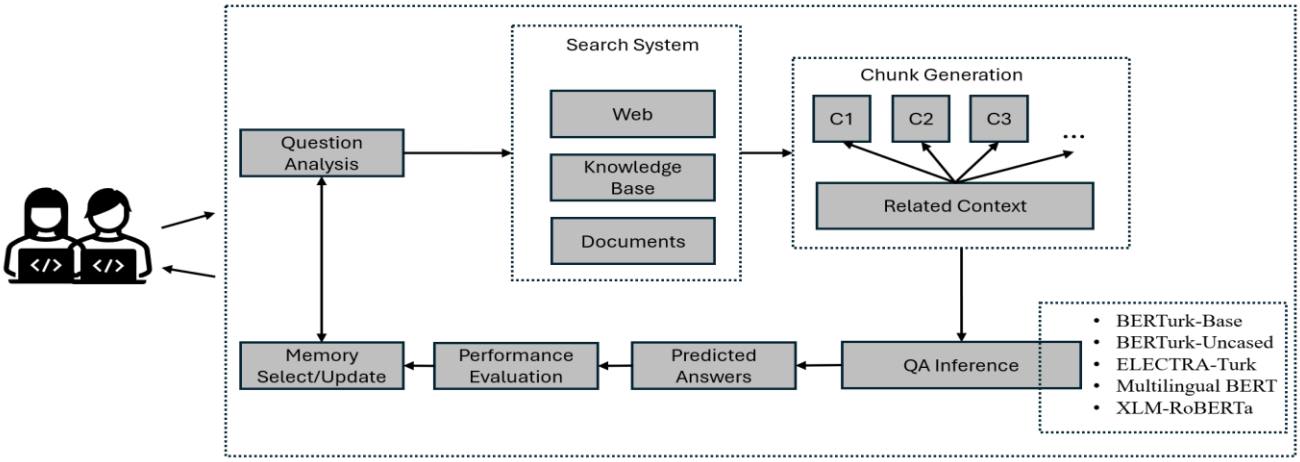

Figure 2. Overview of the Proposed QA System Pipeline. The system processes user questions through multiple stages: (1) Question input and preprocessing, (2) Context retrieval from university documents and FAQ sources, (3) Document chunking for handling long passages, (4) Fine-tuned transformer model processing for answer extraction, and (5) Answer span selection and output generation. The pipeline demonstrates the end-to-end flow from question input to final answer delivery.

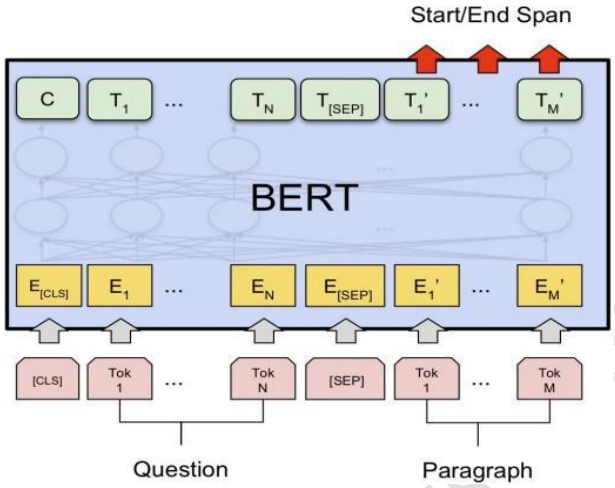

Figure 3. BERT Model [20],[22].

as BERT but differs significantly in its pretraining strategy and hyperparameter configuration. It eliminates the NSP objective. RoBERTa also adopts dynamic masking during training and is pretrained on more data than BERT.

ELECTRA [27] is designed as a more sample and computing efficient alternative to MLM approaches such as BERT. In traditional MLM, only a small subset of input tokens is masked and predicted during training. ELECTRA addresses this inefficiency by using a discriminative pretraining objective called Replaced Token Detection (RTD). A small generator predicts replacements for some input tokens, and a discriminator is trained to detect which tokens in the sequence have been replaced. Unlike MLM, ELECTRA's training signal is defined over all tokens in the input, not just those that are masked. ELECTRA's architecture can be fine-tuned for a wide range of downstream NLP tasks such as classification and question answering.

Transformer models offer a fundamental advantage by allowing rapid and cost-effective fine-tuning in terms of space and time complexities through the transfer of pre-trained foundational language models. During pre-training, the model learns contextual word embeddings or sentence encodings, enabling it to be repurposed for downstream tasks with even small amounts of labeled data. Another advantage is the attention mechanism, which allows the model to better capture long-term dependencies.

To assess the effectiveness of different architectures for Turkish-language question answering, five pretrained models were selected based on their language coverage, tokenizer design, and pretraining strategies. These include both monolingual models trained specifically on Turkish corpora and multilingual models with broader cross-lingual capabilities. All models were obtained from the Hugging Face, fine-tuned using the THQuAD dataset [2], and subsequently tested on a domain-specific university FAQ dataset. Table 1 summarizes the selected models.

Table 1. Key architectural features of the selected models

Model	Hugging Face Model Name	Language Coverage	Layers
BERTurk-Base	dbmdz/bert-base-turkish-cased	Turkish	12
BERTurk-Uncased	dbmdz/bert-base-turkish-uncased	Turkish	12
mBERT	bert-base-multilingual-cased	104 languages	12
ELECTRA-Turk	dbmdz/electra-base-turkish-cased-discriminator	Turkish	12
XLM-R	xlm-roberta-base	100 languages	12

BERTurk-Base [28]: A Turkish BERT model trained on a 35 GB corpus consisting of OSCAR, OPUS, Wikipedia and a special corpus data. It follows the standard 12-layer BERT architecture.

BERTurk-Uncased [28]: Similar to the cased version but performs lowercasing and accent striping during preprocessing. This model uses the same training corpus and architecture as BERTurk-Base.

mBERT [20]: A multilingual version of BERT trained on Wikipedia text from 104 languages.

ELECTRA-Turk [27]: A Turkish-specific ELECTRA model that replaces masked token prediction with a more sample-efficient RTD objective. It was trained on the same data as BERTurk.

XLM-RoBERTa (XLM-R) [29]: XLM-R builds on the RoBERTa architecture by extending it to the multilingual setting. It is trained on 2.5 TB of filtered CommonCrawl data in 100 languages using masked language modeling objective.

3.2 Dataset Details

This study adopts a two-stage dataset pipeline to develop and evaluate transformer-based question answering models for Turkish-language educational applications. In the first stage, model fine-tuning was conducted using the Turkish Historic Question Answering Dataset for Reading Comprehension (THQuAD) [2]. THQuAD supports extractive question answering tasks through context passages, WH-type questions, and annotated answer spans defined by start and end character indices. THQuAD consists of 2,701 passages and 15,554 question-answer pairs. The dataset comprises two thematic sub-corpora: TQuAD [30], which is one of the earliest structured Turkish QA resources, and a second corpus covering Ottoman history.

In the second stage of the study, a custom evaluation dataset was developed to assess model performance on real-world academic and administrative queries in Turkish higher education. The dataset comprises over 200 frequently asked questions collected from the Faculty of Computer and Information Sciences and Sakarya University. These include student handbooks, academic regulations, Erasmus mobility documentation, internship guidelines, and informal queries submitted via departmental emails and social media. Each entry contains a question, a manually annotated answer span, and a corresponding context paragraph extracted from official documents. The dataset follows the SQuAD-style extractive QA format and was used exclusively for evaluation. Sample questions from the dataset shown in Table 2.

3.3 Evaluation

In extractive question-answering systems based on transformer architectures, the task is to identify the start and end positions of the answer span within a given context. The input to the model consists of the concatenated question and context, represented using three embedding types [18]:

- Token Embeddings (E_{token}): Each word or word piece in the text is converted into a token embedding.
- Segment Embeddings (E_{segment}): Used to distinguish between the question and the context.
- Position Embeddings (E_{position}): Used to denote the position of words in the sequence.

$$Input\ Embeddings = E_{token} + E_{segment} + E_{position} \quad (1)$$

In the QA task, the model predicts the start and end positions of the answer span within the context. The model outputs two sets of logits: *S* for the start positions and *E* for the end positions.

$$S_i = StartLogit(h_i)\ , E_i = EndLogit(h_i) \quad (2)$$

where h_i represents the hidden states corresponding to the input tokens.

Table 2. Sample Questions from the Dataset

No	Turkish Question	English Translation
1	Ders seçimi nasıl yapılır?	How is course selection done?
2	Fakülte Staj Komisyonu kimlerden oluşur?	Who are the members of the Faculty Internship Committee?
3	Bilgisayar ve Bilişim Bilimleri Fakültesi öğrencileri kaç gün staj yapmak zorundadırlar?	How many days of internship are Computer and Information Sciences Faculty students required to complete?
4	Stajlar hangi tarihler arasında yapılır?	During which dates are internships conducted?
5	Staj defteri nasıl doldurulur?	How should the internship logbook be filled out?
6	Bilgisayar ve Bilişim Bilimleri Fakültesinde hangi bölümler bulunmaktadır?	Which departments are present in the Faculty of Computer and Information Sciences?
7	Bilişim Sistemleri Mühendisliği bölümünde kaç farklı komisyon bulunmaktadır?	How many different committees are there in the Department of Information Systems Engineering?
8	Genel amaçlı yazılım laboratuvarında kaç bilgisayar bulunmaktadır?	How many computers are available in the general-purpose software laboratory?
9	Kredi ve not aktarım formu nasıl doldurulmalıdır?	How should the credit and grade transfer form be filled out?
10	Bölümün amacı nedir?	What is the purpose of the department?
11	Mezun olabilmek için toplam kaç kredilik ders alınması gerekmektedir?	How many credits of courses must be completed to graduate?
12	Erasmus programına nasıl başvurabilirim?	How can I apply for the Erasmus program?
13	Öğrenci kulüplerine nasıl katılabilirim?	How can I join student clubs?
14	Program veya bölümlerde UMDE faaliyetlerinin izlenmesinden kim sorumludur?	Who is responsible for monitoring UMDE activities in programs or departments?
15	Görüntü işleme alanında çalışan akademisyenler kimlerdir?	Who are the academicians working in the field of image processing?
16	Staj Yönetim Sistemi linki nedir?	What is the link to the Internship Management System?
17	Öğrenci numarasını nereden öğrenebiliriz?	Where can we find the student number?
18	SABİS şifresini unuttuğumda nereye müracaat ederek şifre sıfırlama işlemi yaptırabilirim?	Where should I apply to reset my SABIS password if I forget it?
19	Bitirme alabilmek için Genel Not Ortalaması kaç olmalıdır?	What should be the Grade Point Average (GPA) to take the graduation course?
20	Mezun olabilmek için kaç AKTS gerekir?	How many ECTS credits must be completed to graduate?

In all question answering systems developed in this study, the loss is computed as the average of two cross-entropy losses corresponding to the predicted start and end positions of the answer span. This standard approach is internally handled by the Hugging Face Trainer class in conjunction with transformer-based models. By providing the ground truth start and end indices during training, the models are optimized to accurately locate answer boundaries within the context.

To thoroughly assess the effectiveness of the models utilized in the study, various performance metrics are meticulously calculated. These metrics are used to evaluate the predicted responses against the actual outcomes, providing a comprehensive understanding of the model's accuracy, precision, recall, and overall performance.

Exact Match (EM) checks if the predicted answer is exactly the same as the true answer after stripping whitespace and converting to lowercase.

$$EM(\hat{a}, a) = \begin{cases} 1 & if\ \text{norm}(\hat{a}) = \text{norm}(\text{a}) \\ 0 & otherwise \end{cases} \quad (3)$$

The F1 score is calculated as the harmonic mean of Precision and Recall.

$$Precision = \frac{|common\_tokens|}{|pred\_tokens|} \quad (4)$$

$$Recall = \frac{|common\_tokens|}{|true\_tokens|} \quad (5)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

To address limitations of EM in real-world use cases, we extended the evaluation with a Relaxed Exact Match (REM), where each ground truth question is allowed to have a set of semantically valid reference answers A={a_1,a_2,...,a_k}. The predicted answer $\hat{a}$ is considered correct if it matches any of the normalized answers in this set.

The evaluation metrics were applied under both strict and relaxed conditions. The strict setting requires exact lexical matches, while the relaxed setting accepts semantically equivalent but lexically variant answers, aiming to better reflect real-world applicability and user satisfaction.

$$REM(\hat{a}, A) = \begin{cases} 1 & \text{if } \exists a_i \in A \text{ such that } \text{norm}(\hat{a}) = \text{norm}(a_i) \\ 0 & otherwise \end{cases} \quad (7)$$

Given a predicted answer $\hat{a}$, the Relaxed F1 score is computed by evaluating the standard F1 score against each of the acceptable references in A and selecting the highest score.

$$Relaxed\ F1(\hat{a}, A) = max_{a_i \in A} F1(\hat{a}, a_i) \quad (8)$$

4 Experimental Results

This section presents the empirical evaluation of five transformer-based language models fine-tuned for Turkish question answering. The aim is to assess model effectiveness in responding to real-world questions derived from university settings, both syntactically and semantically. All models were fine-tuned using the THQuAD dataset [2] and subsequently evaluated on a custom-built dataset comprising university-related questions.

All models were fine-tuned and evaluated using Google Colab with a T4 GPU. We used the Hugging Face Transformers library. Training strategy and model selection directly affect generalization performance [31]. Furhtermore, hyperparameters such as optimizer, learning rate, and other training-related settings have a significant impact on the performance of machine learning models [32]. For each model, we adopted a consistent set of hyperparameters to ensure fair comparison. These configurations are summarized in Table 3.

To assess model performance, we used EM and F1 Score metrics. While these metrics provide strong formal guarantees, they often penalize answers that are semantically correct yet lexically divergent from the ground truth.

Recognizing this limitation, we introduced an additional evaluation protocol based on relaxed matching. In this setting, alternate correct variants of the ground truth were manually annotated and incorporated into the evaluation process. The manual annotation process for semantically equivalent answers involved identifying valid variations for each ground truth answer by considering: (1) abbreviations and full forms, (2) numerical representations, (3) synonymous terms commonly used in academic contexts, and (4) alternative phrasings with identical meaning. This annotation process resulted in a reference set that better reflects the natural linguistic variations. The strict and relaxed evaluation scores for all models are presented in Table 4.

Table 3 Fine-Tuning Configuration.

Parameter	Value
Hardware	Google Colab T4 GPU
Number of Epochs	5
Learning Rate	3e-5
Weight Decay	0.01
Batch Size	12
Optimizer	AdamW
Evaluation Strategy	Per Epoch
Early Stopping	No (use the best model at the end)

Table 4. Strict and Relaxed Scores for All Models

Model	EM	F1	Relaxed EM	Relaxed F1
BERTurk-Base	**0.5870**	**0.8411**	0.8261	0.8822
BERTurk-Uncased	0.4783	0.8180	0.8043	0.8795
ELECTRA-Turk	0.4783	0.8370	**0.8478**	**0.8936**
mBERT	0.5652	0.7656	0.7391	0.7975
XLM-R	0.3478	0.7396	0.6087	0.7727

Figure 4 demonstrates the substantial performance differences between model architectures under relaxed evaluation conditions. ELECTRA-Turk leads with the highest Relaxed F1 score of 0.8936, closely followed by BERTurk-Base at 0.8822. The consistent superiority of Turkish-specific models (ELECTRA-Turk, BERTurk-Base, and BERTurk-Uncased) over multilingual alternatives (mBERT and XLM-R) highlights the importance of language-specific pretraining for optimal performance. The relaxed evaluation protocol reveals that models often produce semantically correct answers that differ lexically from ground truth, with performance improvements in F1 scores compared to strict evaluation metrics.

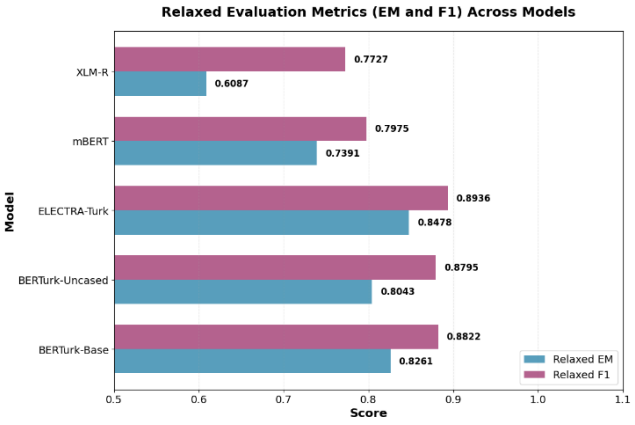

Figure 4. Comparative Performance Analysis Using Relaxed Evaluation Metrics

Among all evaluated models, BERTurk-Base emerged as the top performer under the strict evaluation criteria, achieving an EM score of 0.5870 and an F1 score of 0.8411. ELECTRA-Turk followed closely in F1 (0.8370), though its strict EM score remained at 0.4783, similar to BERTurk-Uncased. This suggests a prevalence of partial matches, where predictions approximated the correct span but did not precisely replicate it.

Models pretrained on Turkish-specific corpora consistently outperform multilingual baselines. The observed discrepancies between EM and F1 scores underscore the importance of considering partial matches in the evaluation of real-world QA systems. For example, a model predicting "240 ECTS" instead of "240 ECTS credits" would be penalized under strict metrics, despite the two answers being functionally equivalent in the given academic context. The relaxed evaluation protocol thus serves as a more realistic benchmark for downstream utility in our settings.

When relaxed evaluation criteria were applied, all models demonstrated substantial performance gains, underscoring the rigidity of the exact match condition. Notably, ELECTRA-Turk surpassed its counterparts under this protocol, achieving a relaxed EM of 0.8478 and a relaxed F1 of 0.8936. This result highlights ELECTRA's ability to semantically align with target answers despite occasional lexical variations. BERTurk-Base and BERTurk-Uncased also delivered robust relaxed scores.
Experimental results demonstrate that transformer-based models, when fine-tuned with domain-specific data, offer reliable and contextually relevant performance for question answering tasks in educational environments. These results highlight the models' capacity to generalize across diverse queries and support efficient information retrieval in real-world academic scenarios.

5 Conclusion

Effectively responding to frequently asked questions in university settings is essential for improving institutional efficiency and facilitating access to accurate information. Students and staff often seek guidance on administrative procedures, academic policies, and operational details. A robust QAS can significantly reduce the workload of administrative units while enhancing the responsiveness and quality of student services.

This study presented a detailed evaluation of transformer-based models for Turkish-language QA tasks within higher education. Five pretrained models, including both monolingual and multilingual variants, were fine-tuned on the THQuAD dataset and tested on a domain-specific corpus compiled from faculty regulations, student handbooks, and informal inquiries.

The results revealed that monolingual models trained specifically for Turkish outperform multilingual alternatives. ELECTRA-Turk achieved the best results with an F1 score of 0.8936 and an Exact Match score of 0.8478. These scores were obtained under an extended evaluation framework that accounts for semantically correct but lexically divergent predictions, providing a more realistic measure of model utility in practical scenarios. BERTurk-Base models also demonstrated strong performance, further validating the effectiveness of linguistically aligned pretraining.
These findings highlight the importance of using domain- and language-specific models to enhance QA performance. From a practical perspective, transformer-based QA systems offer scalable solutions for academic institutions by automating routine information delivery and improving user satisfaction.

While our results demonstrate strong performance within the Turkish higher education domain, several limitations should be acknowledged. The models' generalizability to other domains (e.g., healthcare, legal services) remains unclear due to domain-specific vocabulary and question types in our data. Additionally, the extractive QA approach may not suit applications requiring generative responses or complex reasoning. Future work should explore cross-domain transfer learning and evaluate performance across diverse domains to better understand the broader applicability of transformer-based QA systems.

Future research may extend this work by incorporating generative models for open-ended questions, evaluating the systems in interactive scenarios, and exploring transfer learning approaches for related low-resource languages. These directions can further strengthen the role of AI-driven QA systems in higher education and beyond.

6 Author Contribution Statements

Author 1, Creation of the idea, design, supply of resources and materials, data collection, analysis, literature review, writing.
Author 2, Creation of the idea, design, supply of resources and materials, data collection, analysis, literature review, writing.

7 Ethics committee approval and conflict of interest statement

"Ethics committee permission is not required for the article prepared".

"There is no conflict of interest with any person/institution in the article prepared".

8 References

[1] Guo X, Zhao B, Ning B. "A survey on intelligent question and answer systems". *International Conference on Mobile Computing, Applications, and Services*, 81–88, 2022.

[2] Soygazi, F., Çiftçi, O., Kök, U., & Cengiz, S. "THQuAD: Turkish historic question answering dataset for reading comprehension." In 2021 6th international conference on computer science and engineering (UBMK) (pp. 215-220). *IEEE,*2021.

[3] Okonkwo CW, Ade-Ibijola A. "Chatbots applications in education: A systematic review". *Computers and Education: Artificial Intelligence*, 2, 100033, 2021.

[4] Smutny P, Schreiberova P. "Chatbots for learning: A review of educational chatbots for the facebook messenger". *Computers & Education*, 151, 103862, 2020.

[5] Yousuf M, Jami SI. "An automated question-answering (q/a) system for academic environment". *2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, 1–6, 2022.

[6] Fulmal V, et al. "The implementation of question answer system using deep learning". *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(1S), 176–182, 2021.

[7] Madabushi HT, Lee M, Barnden J. "Integrating question classification and deep learning for improved answer selection". *Proceedings of the 27th International Conference on Computational Linguistics*, 3283–3294, 2018.

[8] Clarizia F, Colace F, Lombardi M, Pascale F, Santaniello D. "Chatbot: An education support system for student". *Cyberspace Safety and Security: 10th International Symposium, CSS 2018*, Amalfi, Italy, 29–31 October 2018, Springer, 291–302, 2018.

[9] Durmus E, He H, Diab M. "Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization". *arXiv preprint arXiv:2005.03754*, 2020.

[10] Ardac HA, Erdogmus P. "Question answering system with text mining and deep networks". *Evolving Systems*, 1–13, 2024.

[11] Derici C, Aydin Y, Yenialaca N, Aydin NY, Kartal G, Ozgur A, Gungor T. "A closed-domain question answering framework using reliable resources to assist students". *Natural Language Engineering*, 24(5), 725–762, 2018.

[12] Akyön FÇ, Cavuşoğlu ADE, Cengiz C, Altinuc SO, Temizel A. "Automated question generation and question answering from turkish texts". *Turkish Journal of Electrical Engineering and Computer Sciences*, 30(5), 1931–1940, 2022.

[13] Kooli C. "Chatbots in education and research: A critical examination of ethical implications and solutions". *Sustainability*, 15(7), 5614, 2023.

[14] Altintas V, Kilinc M. "Automated categorization of Turkish e-commerce product reviews using BERTurk". 8th International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1–6, September 2024, IEEE.

[15]Incidelen M, Aydoğan M. "Developing question-answering models in low-resource languages: A case study on Turkish medical texts using transformer-based approaches". 8th International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1–4, September 2024, IEEE.

[16] Ardic O, Ozturk MU, Demirtas I, Arslan S. "Information extraction from sustainability reports in Turkish through RAG approach". 32nd Signal Processing and Communications Applications Conference (SIU), pp. 1–4, May 2024, IEEE.

[17] Yildirim S. "Fine-tuning transformer-based encoder for turkish language understanding tasks". *arXiv preprint arXiv:2401.17396*, 2024.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. "Attention is all you need". *Advances in Neural Information Processing Systems*, 30, 2017.

[19] Kotan M. "Duygu analizi ve dijital dönüşüm üzerine etkileri". In book: *Dijital Dönüşümler ve Sektörel Yansımaları 2*, 2022.

[20] Devlin J, Chang MW, Lee K, Toutanova K. "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*, 2018.

[21] Alagöz NK, Küçüksille EU. "System of automatic scientific article summarization in Turkish". *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 30(4), 470–481, 2024.

[22] Kenton JD, Chang MW, Toutanova K. "Bert: Pre-training of deep bidirectional transformers for language understanding". *Proceedings of NAACL-HLT*, 1, 2, 2019.

[23] Amer E, Hazem A, Farouk O, Louca A, Mohamed Y, Ashraf M. "A proposed chatbot framework for COVID-19". *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 263–268, 2021.

[24] Koroteev M. "Bert: a review of applications in natural language processing and understanding". *arXiv preprint arXiv:2103.11943*, 2021.

[25] Gürbüz, M., & Kotan, M. "Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT", Acta Infologica, 10.26650/acin.1483488, 2025.

[26] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692,2019.

[27] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555,2020.

[28] Schweter, S. "BERTurk - BERT models for Turkish". Zenodo, 10.5281/zenodo.3770924,2020.

[29] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116,2019.

[30] Turkish NLP Q&A Dataset, Availabel at: https://github.com/TQuad/turkish-nlp-qa-dataset

[31] Saka SO, Cömert Z. "Sentiment analysis based on text with Universal Sentence Encoder and CNN-LSTM models". *8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–4, September 2024, IEEE.

[32] Yücel N, Cömert Ö. "Müşteri duyarlılığını keşfetmek için yapay zeka destekli analiz ile çevrimiçi ürün incelemelerinden anlamlı bilgiler elde etme". *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 35(2), 679–690, 2023.