A QUESTION ANSWERING SYSTEM USING DEEP LEARNING TECHNIQUES IN THE EDUCATION DOMAIN

EĞİTİM ALANINDA DERİN ÖĞRENME TEKNİKLERİNİ KULLANAN BİR SORU CEVAPLAMA SİSTEMİ

ZEYNEP ŞANLI

PROF. DR. İLYAS ÇİÇEKLİ

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Master of Science

in Computer Engineering

September 2024

ABSTRACT

A QUESTION ANSWERING SYSTEM USING DEEP LEARNING TECHNIQUES IN THE EDUCATION DOMAIN

Zeynep Şanlı

Master of Science, Computer Engineering

Supervisor: Prof. Dr. İlyas ÇİÇEKLİ

September 2024, 59 pages

Integrating advanced AI-driven question answering (QA) systems into educational settings offers significant potential for enhancing learning experiences. This study focuses on developing and optimizing an educational QA system using the T5-base model, a versatile text-to-text transformer known for its robust performance in natural language processing tasks. In addition to T5, other prominent large language models (LLMs) such as GPT-3, GPT-4 and BERT are also evaluated to compare several vital metrics comprehensively. By employing deep learning techniques such as Transformer architecture and sequence-to-sequence (Seq2Seq) models, the QA system is designed to provide contextually relevant and accurate responses to educational queries. The T5 model is fine-tuned and optimized through experiments to enhance its performance and responsiveness. The results indicate that, despite its smaller size, the T5-base model effectively generates meaningful answers, demonstrating its potential utility in educational applications. This research evaluates the effectiveness of the T5-base model and provides a benchmark for assessing other LLMs in educational QA applications. The evaluation results emphasize the need for a balanced approach in model selection, considering factors such as performance, resource efficiency, and the specific requirements of educational

environments. This study contributes to creating more innovative and effective educational tools by enhancing the understanding of AI-driven QA systems in education.

Keywords: Question Answering (QA) Systems, T5-base Model, Deep Learning, LLMs

ÖZET

EĞİTİM ALANINDA DERİN ÖĞRENME TEKNİKLERİNİ KULLANAN BİR SORU CEVAPLAMA SİSTEMİ

Zeynep Şanlı

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Prof. Dr. İlyas ÇİÇEKLİ

Eylül 2024, 59 sayfa

Eğitim ortamlarına gelişmiş yapay zeka destekli soru cevaplama (QA) sistemlerinin entegrasyonu, öğrenme deneyimlerini geliştirme potansiyeline önemli ölçüde katkı sağlar. Bu çalışma, doğal dil işleme görevlerinde sağlam performansıyla tanınan çok yönlü bir metinden metne dönüştürücü olan T5-base modeli kullanarak bir eğitim QA sistemini geliştirmeye ve optimize etmeye odaklanmaktadır. T5'e ek olarak, GPT-3, GPT-4 ve BERT gibi diğer önde gelen büyük dil modelleri (LLM) de birkaç önemli metriği kapsamlı bir şekilde karşılaştırmak için değerlendirilmektedir. Dönüşüm mimarisi ve dizi-diziden (Seq2Seq) modeller gibi derin öğrenme tekniklerini kullanarak, QA sistemi eğitim sorgularına bağlamsal olarak alakalı ve doğru yanıtlar sağlamak üzere tasarlanmıştır. T5 modeli, performansını ve tepki süresini artırmak amacıyla deneylerle ince ayar yapılarak optimize edilmiştir. Sonuçlar, daha küçük boyutuna rağmen, T5-base modelinin anlamlı yanıtlar üretebildiğini ve eğitim uygulamalarında potansiyel faydasını göstermektedir. Bu araştırma, T5-base modelinin etkinliğini değerlendirir ve eğitim QA uygulamalarında diğer LLM'leri değerlendirmek için bir ölçüt sağlar. Değerlendirme sonuçları, model seçiminde performans, kaynak verimliliği ve eğitim ortamlarının özel gereksinimleri gibi faktörleri dikkate alarak dengeli bir yaklaşım gereksinimini vurgular. Bu çalışma, eğitimde yapay

zeka destekli QA sistemlerinin anlaşılmasını artırarak daha yenilikçi ve etkili eğitim araçları oluşturulmasına katkıda bulunur.

Keywords: Soru Cevap Sistemi, T5-base Model, Derin Öğrenme, Büyük Dil Modeli

CONTENTS

Page

ABSTRACT i
ÖZET iii
CONTENTS v
TABLES vi
FIGURES vii
ABBREVIATIONS viii
1. INTRODUCTION 1
1.1. Scope Of The Thesis 3
1.2. Contributions 3
1.3. Organization 5
2. BACKGROUND OVERVIEW 6
3. RELATED WORK 9
4. SYSTEM DESING AND DEVELOPMENT 15
4.1. T5 Model 16
4.2. Propesed System 17
4.2.1. System Overview 18
4.2.2. Training Setup 18
4.2.3. Fine-Tuning T5 Model 18
4.3. Dataset 21
5. TESTING AND EVALUATION 23
5.1. Testing Environment and Data 23
5.2. Performance Evaluation and Results 23
5.2.1. Metrics 24
5.2.2. Results 28
6. CONCLUSION 43
6.1. Future Directions 45

TABLES

Page

Table 4.1 Parameters tested in Fine-tuning 20

Table 5.1 Results of T5 model and fine-tuned models 29

Table 5.2 Load time in seconds 29

Table 5.3 Memory usage of model [GB] 30

Table 5.4 Latency of models [seconds] 31

Table 5.5 Total execution time, *squad dataset 32

Table 5.6 Total execution time, *trivia dataset 32

Table 5.7 Average EMs 34

Table 5.8 Number of unanswered questions 35

Table 5.9 Average F1 Score 37

Table 5.10 Test results of Squad-v2 38

Table 5.11 Test results of Trivia 39

Table 5.12 Test results of EduSpecialized 40

FIGURES

Page

Figure 4.1 Workflow diagram of T5 QA System 15

Figure 4.2 Sample Data 22

ABBREVIATIONS

QA	:	**Q**uestion **A**nswering
NLP	:	**N**atural **L**anguage **P**rocessing
DL	:	**D**eep **L**earning
ML	:	**M**achine **L**earning
LLM	:	**L**arge **L**anguage **M**odel
AI	:	**A**rtificial **I**ntelligence
GPT	:	**G**enerative **P**re-trained **T**ransformer
SVMs	:	**S**upport **V**ector **M**achines
BERT	:	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
RoBERTa	:	A **R**obustly **O**ptimized **BERT** **P**retraining **A**pproach
DistilBERT	:	**Distil**led **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
T5	:	**T**ext-**T**o-**T**ext **T**ransfer **T**ransformer
Flan-T5	:	**F**ine-Tuned **Lan**guage Model - **T**ext-**T**o-**T**ext **T**ransfer Transformer
Seq2Seq	:	**Seq**uence to **Seq**uence

1. INTRODUCTION

The evolution of Question Answering (QA) system technology is a testament to the rapid advancements in artificial intelligence and natural language processing (NLP). From the rudimentary Baseball, developed by David L. Waltz, to contemporary models like GPT and T5, QA systems have undergone significant transformations. BASEBALL, often regarded as the first QA system to answer questions about baseball games using a rules-based approach [1]. The field has since progressed from foundational rule-based systems that lacked flexibility to sophisticated AI-driven models that offer enhanced responsiveness and contextual understanding [2].

The transition from rule-based systems to machine learning (ML) models marked a significant milestone in QA system development. Early ML models such as decision trees and support vector machines (SVMs) improved intent classification but required extensive training data and were prone to overfitting. The introduction of sequence-to-sequence (Seq2Seq) models further advanced the field by improving the coherence of generated responses, albeit with challenges in maintaining long-term context and computational demands [3], [4]. The emergence of the Transformer model, with its self-attention mechanism, revolutionized QA development by enabling more context-aware and scalable interactions [5].

NLP techniques underpin the functionality of modern QA systems, allowing them to process and generate human-like text. Foundational techniques such as tokenization, stemming, and lemmatization are essential for input processing, while word embeddings like Word2Vec and GloVe capture semantic relationships in vector space [6], [7]. The development of models like BERT and GPT further enhanced the ability of QA systems to understand and generate contextually relevant responses [8], [9].

As QA systems have evolved, their potential applications in educational settings have become increasingly evident. Modern AI-driven QA systems, particularly those based on large language models (LLMs) such as GPT-3, GPT-4, and T5, are demonstrating

exceptional capabilities in providing instant, accurate responses to educational queries. By automating routine question-answering tasks, these systems can support personalized learning experiences, enhance student engagement, and reduce the cognitive load on educators [10], [11].

QA systems have shown considerable promise in enhancing learning experiences in educational settings. Early educational QA systems, such as those highlighted by Brewer, primarily relied on rule-based systems to provide essential instructional support [12]. However, as NLP techniques have advanced, so too have the capabilities of educational QA systems. Modern AI-driven models can now provide real-time, context-aware feedback, significantly improving learning outcomes and engagement [13], [14]. These advances have led to the use of models like T5, BERT, GPT-3.5-TURBO, etc., in testing and developing educational QA systems, where each model contributes different accuracy, efficiency, and resource usage strengths.

This thesis explores developing and implementing educational QA systems using the T5-base model. The T5 model, known for its versatility and strong performance across various NLP tasks, is particularly suited for question-answering applications in educational contexts. This study aims to optimize the T5 model for generating meaningful responses to educational queries to enhance users' learning experience. By leveraging advanced deep learning techniques and fine-tuning methodologies, this research seeks to address the challenges of implementing effective educational QA systems and contribute to the ongoing evolution of AI in education.

This chapter introduces an overview of QA systems in section 1.1. Section 1.2 examines QA systems types and their application areas. Section 1.3 explains the role of deep learning in QA systems. Finally, Section 1.4 discusses the purpose and scope of this thesis on this subject.

1.1. Scope Of The Thesis

This thesis aims to explore the development and implementation of an educational QA systems using advanced deep learning techniques, specifically focusing on the T5-base model. The primary focus is on the educational domain, where the QA systems are designed to assist in providing real-time, context-aware responses to student queries. The initial implementation supports the English language, with plans for potential extension to Turkish in future work. The selection of the T5 model, particularly the T5-base variant, is justified based on its performance, versatility, and computational efficiency. In addition to T5-base, this thesis also examines the performance of several other prominent models, including Bert, DistilBERT, RoBERTa, GPT-3.5-turbo, Flan-T5 and GPT-4o-mini to provide a comprehensive evaluation and comparison across various metrics. The thesis delves into the architectural features of the T5 model and its suitability for educational applications. The study involves applying deep learning techniques, including Transformer architecture and seq2seq models, for training and fine-tuning the QA systems model. The research aims to optimize these techniques to enhance the QA systems's ability to generate meaningful educational responses. The comparative analysis across the different models offers insights into their strengths and limitations in educational contexts, helping to determine the most effective approach for implementing AI-driven QA systems in education. Detailed methodologies for data preprocessing, model training, fine-tuning, and evaluation are provided. The thesis assesses the QA systems's performance using specific metrics to ensure its effectiveness in an educational context. Additionally, the scope includes a discussion on the potential impact of the developed QA systems system on educational practices and student engagement, as well as future work that may involve extending language support and exploring additional deep learning models and techniques.

1.2. Contributions

This thesis makes several significant contributions to the field of educational technology and AI-driven learning systems:

- **Development of a Robust Educational QA systems**: Creating a QA systems system using the T5-base model tailored for educational applications. Creating QA systems involves innovative use of deep learning techniques, particularly Transformer architecture and seq2seq models, to enhance the QA systems's responsiveness and contextual understanding. Additionally, the thesis explores the application of other prominent models, including Bert, DistilBERT, RoBERTa, GPT-3.5-turbo, Flan-T5 and GPT-4o-mini providing a comparative analysis of their performance in educational contexts.

- **Model Optimization for Educational Queries**: Significant efforts have been made to optimize the T5 model for generating accurate and relevant responses to educational queries. This optimization includes fine-tuning the model with a specific dataset designed for educational purposes. The thesis also examines the fine-tuning processes for other models to identify the most efficient and effective approach for educational QA systems.

- **Evaluation Framework**: Establishing a comprehensive evaluation framework for assessing the QA systems' performance. This framework uses specific metrics to measure the accuracy and relevance of the QA systems' responses in an educational setting. The evaluation framework is applied to the T5-base model and the other models tested in this study, providing a benchmark for future research in educational AI systems.

- **Implementation Methodology**: A detailed methodology for implementing and integrating the QA systems system into educational environments. Integration of the QA system to educational environments includes steps for data preprocessing, model training, fine-tuning, and real-time deployment. The methodology is designed to be adaptable across different AI models, ensuring broad applicability in diverse educational settings.

- **Language Support Extension**: Laying the groundwork for extending the QA systems's language support, initially focused on English, with plans to include Turkish.

The language support enhances the accessibility and applicability of the system in diverse educational contexts. Further, the study explores the potential for extending language support across the other models evaluated, aiming to create a multilingual educational tool that can be deployed in various linguistic environments.

1.3. Organization

The organization of the thesis is as follows:

- Chapter 1 presents our motivation, contributions, and the scope of the thesis.
- Chapter 2 provides an overview of QA systems, application types, and development methods.
- Chapter 3 gives information about related work.
- Chapter 4 gives information about system design and development and data preprocessing.
- Chapter 5 demonstrates the results, testing, and evaluation.
- Chapter 6 provides a summary of the thesis.

2. BACKGROUND OVERVIEW

Advancements in machine learning (ML) and deep learning (DL) have significantly influenced the development of Question-Answering (QA) systems. Firsy QA systems were largely rule-based, they depend on predefined scripts and pattern matching to respond to user queries. However, these early systems were rigid so that they were insufficient to handle the complexities of natural language.

The introduction of ML brought a major change in the development of QA systems. ML algorithms, such as support vector machines (SVMs) and decision trees, allowed systems to analyze more extensive datasets and learn from examples rather than relying on predefined rules. This change enabled QA systems to improve accuracy and adapt to a wider range of inputs [15].

The Impact of Deep Learning on QA Systems

The most significant advancements in QA systems have come with the rise of Deep Learning (DL). DL models, particularly those based on neural networks, have revolutionized how QA systems process and generate natural language. Unlike traditional ML models, DL models can automatically learn representations from raw data, allowing them to understand the deeper semantic meanings of words and sentence meanings of words and sentences [16].

Transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), are primary examples of the DL's impact on QA systems. These models utilize self-attention mechanisms to weigh the relevance of each word in a sentence relative to others, enabling a more nuanced understanding of context [4]. This capability is crucial for QA systems, allowing them to generate more accurate and contextually appropriate responses, even for complex queries [9].

Deploying DL models in QA systems has also significantly improved their ability to handle unstructured data, which is common in real-world applications. For example, a DL-powered

QA system can process and answer questions based on large volumes of text from various sources, such as books, articles, and online content, making it much more versatile than earlier systems [10].

Educational Applications of ML and DL in QA Systems

Integrating ML and DL into QA systems in the educational domain led to substantial improvements in how students and educators interact with these tools. ML techniques allow educational QA systems to learn from vast datasets of educational content, enabling them to provide personalized learning experiences [17]. For example, an ML-powered QA system can analyze a student's past performance and tailor its responses to address specific areas where the student needs improvement [18].

DL further enhances this by enabling QA systems to understand and generate more complex language structures, which is particularly beneficial in educational settings where precise and contextually relevant answers are crucial [15]. The use of DL models allows these systems to provide more detailed explanations, generate practice questions, and even assist with essay writing by offering suggestions for improvement [9].

Moreover, DL models' ability to process and analyze unstructured data means that educational QA systems can pull information from various sources, including textbooks, academic papers, and online resources, to provide comprehensive answers. This capability improves the accuracy of the information provided and helps students develop a deeper understanding of the material [19].

Challenges and Future Directions

Despite the significant advancements brought by ML and DL, challenges still need to be addressed in the development of QA systems. One of the main challenges is the computational cost associated with training and deploying DL models. These models require large amounts of data and processing power, which can be a barrier for institutions with limited resources [16]. Additionally, the black-box nature of DL models can make it difficult

to understand how they arrive at specific answers, which can be problematic in educational contexts where transparency is essential [20].

Future research will likely focus on improving the efficiency and interpretability of DL models in QA systems. These improvements include developing more efficient training algorithms, reducing the computational requirements of these models, and creating tools that make it easier to understand and explain the decisions made by DL-powered QA systems [4]. There is also growing interest in exploring the potential of hybrid models that combine the strengths of ML and DL with rule-based approaches to create even more robust and versatile QA systems [19].

3. RELATED WORK

The evolution of Question Answering (QA) systems has been deeply intertwined with advancements in machine learning (ML) and deep learning (DL) techniques. From their inception, QA systems have aimed to automate answering questions in natural language, providing accurate and contextually relevant information. Over time, the methodologies behind these systems have evolved from rule-based approaches to sophisticated ML and DL techniques, significantly enhancing their performance and applicability across various domains, particularly in education.

The earliest QA systems, such as BASEBALL, developed by David L. Waltz, was designed to answer inquiries about baseball statistics, demonstrating early uses of QA for accessing specific database information [1]. Users would ask structured questions in natural language, and the system would match these questions to predefined patterns to find the correct answer from the database. For example, if a user asked, "Who won the championship last year?" BASEBALL would understand the question and retrieve the appropriate answer from the database. However, the limitations of the system were that it could only work on a limited set of information and that questions had to be asked in a specific format. These limitations have led to significant improvements in the evolution of QA systems over time.

In the 1970s, systems like SHRDLU introduced more sophisticated rule-based processing, allowing interactions within a micro-world of geometric shapes. SHRDLU utilized a more complex set of syntactic and semantic rules to parse user inputs, enabling it to understand and manipulate objects within its constrained environment [21]. However, like BASEBALL, SHRDLU was fundamentally limited by its reliance on manually crafted rules, which needed to be more scalable to more complex, open-domain environments. Another system that emerged around the same time, Lunar, is a natural language processing system that answers geological questions about samples brought back from the Moon [22]. This system has been an essential step as a domain-specific QA system used in scientific research.

The limitations of previously used approaches led to the adoption of statistical methods in the 1990s, which marked a significant shift in the development of QA systems. Introducing systems like START, developed by Boris Katz at MIT, was pivotal. START is one of the first web-based question-answer systems. It combined natural language processing (NLP) with information retrieval (IR) techniques, leveraging statistical models to parse queries and retrieve relevant information from structured and unstructured text [23]. Unlike earlier systems, which relied on fixed rules, START utilized statistical models to determine the most relevant pieces of information based on the query, thereby improving the accuracy and relevance of the answers. This system, which spreads access to information to a wider area, laid the foundations of modern web-based QA systems.

The advent of machine learning techniques, particularly in the late 1990s and early 2000s, further revolutionized QA systems. ML models, which could be trained on large datasets to recognize patterns and make predictions, offered a more scalable and flexible approach to QA. These models moved away from manually encoded rules and instead learned to generate answers based on the relationships and patterns discovered in the training data [24]. Early ML-based QA systems often employed techniques like support vector machines (SVMs), decision trees, and logistic regression to classify and retrieve information relevant to a given query.

However, it was the emergence of deep learning in the 2010s that truly transformed QA systems, enabling them to handle the complexities of language with unprecedented accuracy. Deep learning, particularly through the use of neural networks, allowed for the development of models that could process vast amounts of text and learn intricate representations of language [25].

The introduction of deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), provided the foundation for more advanced QA systems. RNNs, and specifically long short-term memory networks (LSTMs), were particularly well-suited for QA tasks because of their ability to capture dependencies across sequences of text. LSTMs helped models understand context by maintaining a memory of previous words

in a sentence, thereby enabling the generation of more contextually appropriate responses [26]. These models formed the backbone of early DL-based QA systems and were often used in conjunction with attention mechanisms to improve focus on relevant parts of the input text [27].

One of the most significant advancements in DL-based QA systems came with the introduction of transformer models, which revolutionized NLP by allowing models to consider the entire input sequence simultaneously rather than processing it sequentially as in RNNs. The transformer architecture, introduced by Vaswani et al. in 2017, forms the basis of many state-of-the-art QA systems today. Transformers utilize a self-attention mechanism that enables models to weigh the importance of different words in a sentence, regardless of their position, thus capturing long-range dependencies more effectively than RNNs or LSTMs [4].

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. in 2019, was a groundbreaking model. BERT's bidirectional approach allowed it to consider the context of a word by looking at both its preceding and succeeding words, leading to a more nuanced understanding of language [28]. BERT was pre-trained on large corpora using masked language modeling (MLM) and next-sentence prediction (NSP) tasks, which helped it develop a deep understanding of the relationships between words in a sentence. Once pre-trained, BERT could be fine-tuned on specific tasks, including QA, by adding a superficial classification layer on top of the pre-trained model. This fine-tuning process allowed BERT to achieve state-of-the-art performance on various QA benchmarks, such as the Stanford Question Answering Dataset (SQuAD) [29].

The success of BERT inspired the development of other transformer-based models, including GPT (Generative Pre-trained Transformer) by OpenAI. GPT-2 and GPT-3, introduced in 2019 and 2020, respectively, took a different approach by focusing on autoregressive language modeling, where the model generates text one word at a time based on the previous context [9], [10]. GPT-3, with its 175 billion parameters, demonstrated remarkable capabilities in generating human-like text and answering questions across a wide range of

topics without any task-specific fine-tuning. This ability to generalize from pre-trained knowledge allowed GPT-3 to perform well in zero-shot and few-shot learning scenarios, where it was given little to no task-specific data during training. However, despite their impressive capabilities, GPT models come with significant challenges, particularly regarding their resource-intensive nature. The training and deployment of models like GPT-3 require substantial computational power, which limits their accessibility for many organizations, especially in educational settings where resources may be limited [10].

In this thesis, the T5-base model is compared with other prominent large language models like GPT-3 and BERT in the context of educational QA applications. The focus is evaluating the balance between model performance and computational efficiency, particularly in environments with limited resources. While GPT-3 offers superior accuracy and generative capabilities, its high computational costs make it less feasible for widespread use in educational settings. On the other hand, models like T5-base, which are more resource-efficient, are analyzed for their potential to provide a more practical solution for educational QA systems.

To address these challenges, research has increasingly focused on developing more efficient transformer-based models that deliver high performance with lower computational demands. One such model is the T5 (Text-To-Text Transfer Transformer), which was introduced by Google in 2020. T5 adopts a unified text-to-text framework, where all NLP tasks are framed as text generation problems. This approach simplifies fine-tuning the model for specific tasks, including QA, by converting tasks like classification, translation, and summarization into text generation tasks [11]. T5 is trained on a large and diverse corpus using a denoising autoencoder objective, where the model learns to predict missing or corrupted text spans. Combined with task-specific fine-tuning, this pre-training approach allows T5 to achieve robust performance across a wide range of NLP tasks.

In educational settings, T5 has shown promise as a QA system due to its ability to be fine-tuned on domain-specific datasets, thereby improving its accuracy and relevance in answering educational queries. Fine-tuning involves training the model on a smaller,

domain-specific dataset after it has been pre-trained on a large corpus, allowing it to adapt to the nuances of the target domain [30]. For example, a T5 model fine-tuned on a dataset of educational materials can provide more accurate answers to questions related to those materials, making it a valuable tool for educators and students alike.

In this thesis, the T5-base model is fine-tuned on an educational dataset and evaluated against other models, such as GPT-3 and BERT, to determine its effectiveness in providing accurate and contextually relevant answers in an educational setting. The study focuses on various metrics, including accuracy, response time, and computational resource usage, to comprehensively compare these models.

Deploying QA systems in education has highlighted the need for models operating efficiently in real-world conditions. Studies comparing different models in educational contexts have evaluated various factors, including accuracy, response time, computational resource usage, and ease of deployment [31]. These studies often emphasize the trade-offs between model performance and resource efficiency, particularly in resource-constrained environments. For instance, while GPT-3 may offer superior accuracy and generative capabilities, its high computational costs make it less feasible for widespread school use. In contrast, models like T5 which are more resource-efficient, provide a more balanced solution for educational QA applications.

This thesis carefully analyzes these trade-offs to determine the most appropriate model for deployment in an educational setting. The study explores the potential of T5-base as viable alternatives to larger, more resource-intensive models, particularly in their ability to deliver high-quality educational content without overwhelming computational demands.

The integration of QA systems into educational environments has had a profound impact on the learning experience. These systems provide students with immediate access to information and support self-directed learning, enabling students to explore complex topics independently. Furthermore, by automating the process of answering routine questions, QA systems free educators to focus on more interactive and personalized teaching methods, thereby enhancing the overall educational experience [32].

However, the success of QA systems in education depends on several factors, including the quality of the training data, the appropriateness of the chosen model, and the ability to fine-tune the model for specific educational domains. Ensuring that QA systems are trained on diverse and high-quality datasets is crucial for maintaining the accuracy and relevance of the answers provided [33]. Additionally, the ethical implications of deploying AI-driven QA systems in education must be carefully considered, particularly regarding potential biases and the impact on traditional teaching methods. Transparent oversight and human-in-the-loop approaches are essential to ensure that these systems complement rather than replace human educators [34].

This thesis critically examines the potential biases and ethical considerations associated with AI-driven QA systems, particularly in the context of educational applications. The study explores how these systems can be designed and implemented to minimize bias and ensure that they support rather than undermine educators' roles.

In conclusion, the evolution of QA systems from rule-based approaches to sophisticated deep learning models like BERT, GPT-3 and T5 reflects the rapid advancements in AI and NLP. Applying these models in education offers significant potential for enhancing the learning experience by providing students with timely and accurate information. This research contributes to the ongoing discourse by evaluating the effectiveness of the T5-base model in educational QA applications and comparing it with other prominent large language models. Future research should continue to explore the balance between model performance. Moreover, resource efficiency should be addressed, particularly in resource-constrained settings, and ethical considerations associated with using AI in education.

4. SYSTEM DESING AND DEVELOPMENT

The main goal of this thesis is to develop and optimize an educational Question Answering (QA) system using the T5-base model. A simple illustration of this model shown in Figure 4.1. T5 is a highly versatile text-to-text transformer. It especially has strong performance in natural language processing tasks. Also, this system compared system with other LLMs. This QA system aims to produce context-aware, precise answers to educational queries. This model is compared with other advanced deep learning techniques, such as sequence-to-sequence (Seq2Seq), and Transformer architecture models in terms of performance.

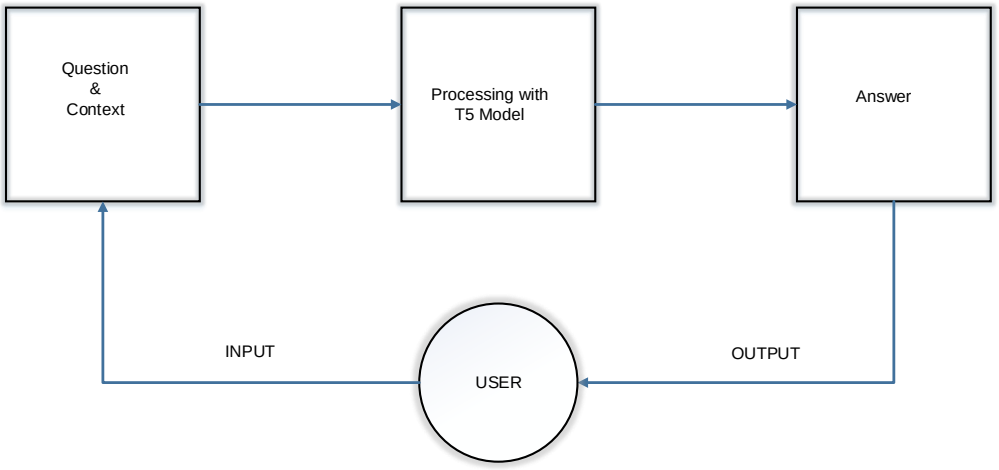

Figure 4.1 Workflow diagram of T5 QA System

In Section 4.1, a detailed explanation of the selection of the T5-base model is provided. The reasons why it is well-suited for the educational context, detailing its architectural

features and stating that. After that, the research compares and evaluates base models withith other prominent large language models (LLMs), such as GPT-3, GPT-4 and BERT offering a comprehensive comparison across many fundamental metrics, such as performance and resource efficiency. These evaluations and comparisons help to determine which model is the most effective and qualified for the specific requirements of educational settings. Section 4.2 summarizes the T5-base model's training set-up of the T5-base model. This section covers the fine-tuning process, which techniques are applied to fine-tune the model, and the hyperparameters used to optimize the model's ability to generate relevant and accurate answers. In Section 4.3, the datasets used in this study are described detailiy.

Currently, the QA system supports English but, it also considers extending the QA system to support Turkish in future developments, aiming to widen its applicability in diverse educational contexts. The study concludes by underlining the importance of selecting a balanced approach in model selection, taking hardware and resource constraints and the model performance while highlighting the contributions of this study to the progress of AI-driven QA systems in education.

4.1. T5 Model

The T5 (Text-to-Text Transfer Transformer) model was selected for this project because it is well-rounded and performs well across various NLP tasks, such as text generation, translation, summarization, and question-answering. The T5 model treats all NLP tasks as a text-to-text problem, allowing for a unified approach to tasks where both input and output are text strings. This technique simplifies the model architecture and training process, making it an ideal candidate for developing a question-answering (QA) system.

The T5 model, has considered as a standard for many state-of-the-art NLP models. It presented by [11], is based on the Transformer architecture. As mentioned earlier, the T5 model is pre-trained for various text-based tasks. For this reason, it is well-suited for fine-tuning specific applications, such as generating questions from a specific context. It employs an encoder-decoder framework, which means the encoder gets and processes the

input text, and the decoder gives the output text generated. Considering all these features, the T5 model was selected for this study.

Model Architecture The T5 model uses the Encoder-Decoder structure. This structure consists of an Encoder and a Decoder with several layers of attention mechanisms and feed-forward neural networks. The encoder and decoder each have a stack of identical layers, typically consisting of multi-head self-attention and feed-forward layers. The model can be configured with various layers and attention heads this allows scalability, relying on the main points such as the available computational resources and the complexity of the task. If the Encoder and Decoder sections are examined more closely:

- Encoder: This component processes the input text and creates a rich, contextual representation. It consists of multiple self-attention layers that help capture the dependencies between various parts of the input text.
- Decoder: It generates the output text based on the encoder's representations. The decoder component also employs self-attention and cross-attention layers that attend to the encoder's outputs, ensuring that the generated text is relevant to the input, like the encoder component.

As it allows the model to capture complex relationships and generate coherent and contextually relevant responses, this architecture is ideal for tasks that requires understanding and generating text. In addition to that by providing task-specific prompts and training on relevant datasets, the model can be fine-tuned to specific tasks. For this study, the model has been fine-tuned to generate answers based on the questions and context given from the dataset.

4.2. Propesed System

In this section, firstly, system overview, training setup of the selected model, fine-tuning process of the T5 base model are presented. After that, datasets used in this study are detailed.

4.2.1. System Overview

This study proposes the selection and fine-tuning of an advanced AI-driven Question Answering (QA) system, particularly for educational environments. The proposed system is streamlined using the T5-based model, which is a powerful and versatile text-to-text converter for natural language processing (NLP) tasks. The goal of this system is to provide accurate, contextually relevant answers to educational queries with given context, enhancing both student learning experiences and the efficiency of educators while taking performance and resource needs and restrictions topics into account. In addition to that, another goal of this study is to compare the various MLLs on different bases, like performance and resource dependency. The MLLs used to evaluate the T5 base model and its fine-tuned versions are: Bert, Gpt-3.5-Turbo, Gpt-4o, Roberta, Distilbert and Flan-T5.

4.2.2. Training Setup

The training setup includes configuring the model's hyperparameters and the computational environment required for training. For this study, tests run on the CPU, but it may be useful to utilize GPUs to accelerate the training and testing processes, given the intensity of such tasks. If we give information about the hardware on which the tests are performed, on a computer with a 2.6 GHz 6-core Intel Core i7 processor and 16 GB 2400 MHz DDR4 memory. During the development, Python Anaconda Spyder IDE was used, and libraries like PyTorch, Hugging Face's Transformers, and OpenAi library were used to implement and train the T5 model. Determining an optimal training duration through experimentation and balancing model performance with computational cost are considered during the training phase.

4.2.3. Fine-Tuning T5 Model

The fine-tuning process includes adjusting the model's parameters to improve its performance on the intended task. In this study, a pre-trained model is trained on a specific

dataset, aiming to optimize the model's performance for the intended task. This section details the steps taken to fine-tune the T5 model and the training setup used for the proposed system.

- Task-Specific Prompts: The T5 model is provided with task-specific prompts to guide its learning process. For this study, the input is be formatted as context: ¡CONTEXT¿", "question: ¡QUESTION¿, and the model is trained to generate the corresponding answer. A special context specific training data was created to further improve the model on a specific topic.

- Layer freezing: The T5 architecture has 24 layers, 12 in the encoder section and 12 in the decoder section. With the layer freezing method, some layers of the pre-trained model are frozen (i.e., the weights of these layers are not updated during training) and the remaining layers are trained. In this way, the information that the model has previously learned is preserved and only certain layers are updated with new data. In this study, the first 6 layers are frozen. Since the first layers usually learn the basic structures of the language, freezing these layers preserves the general grammar abilities of the model. These layers are usually already learned well enough and do not need to be retrained. Layer freezing reduces the computational load during training and perhaps prevents the model from over-learning. This is especially useful while working with limited data or resources.

- Hyperparameter optimization: This is another method used to fine-tune the model on a specific task. To find the optimal hyperparameters, several optimization techniques can be employed. First one is Grid search which involves exhaustively searching through a predefined set of hyperparameters. Each combination is evaluated, and the best performing set is selected. Secondly, Bayesian optimization that uses a probabilistic model to select the most promising hyperparameters based on past evaluations. This method is more sophisticated and can find optimal hyperparameters more quickly. Last one and the one that is used in this study is Random search, samples hyperparameters randomly from a predefined distribution. It is less computationally intensive than

grid search and can often find good hyperparameters more efficiently. At Table 4.1 hyperparameters monitored at this study are listed.

Parameter	**Value/Value Range**
Batch Size	8, 4, 3
Number of Train Epochs	3 - 8
Warmup Steps	0, 50, 500
Evaluation Strategy	Non, Steps
Gradient Accumulation Step	2 - 4
Weight Decay	0.0, 0.01
Early Stopping	3

Table 4.1 Parameters tested in Fine-tuning

Fine-tuning process involves adjusting the pre-trained T5 model on the specific dataset to optimize its performance for question generation. The steps included:

- Loading Pre-Trained Model: Initializing the T5 model with weights pre-trained on a large corpus of text data. The first results were obtained without making any changes to the T5-base model and compared with ground truth data. Thus, the effect of the changes made on the model can be seen.
- Setting Up Training Parameters: Configuring hyperparameters to control the training process. This step may require repeating and monitoring many experiments. The model should be saved in order to test the fine tuned model with different test data. The test results were compared with the values optained running base model. The results are compared to evaluate how the model responds to parameter changes.

4.3. Dataset

For this research, **SQuAD-v2** and **TriviaQA**, two widely recognized datasets, accessed via the Hugging Face platform. Three thousand data from each dataset were randomly selected to conduct a thorough evaluation, ensuring a representative and diverse subset for testing the models' capabilities.

SQuAD-v2 (Stanford Question Answering Dataset v2.0) [35] is a well-established dataset in the natural language processing (NLP) community designed to evaluate machine reading comprehension systems. Unlike its predecessor, SQuAD-v2 includes unanswerable questions alongside those with answers directly derived from the context, making it a more challenging and realistic benchmark. This feature allows evaluation of the model's ability to extract information and determine when an answer is unavailable from the given context. In this study, only the answerable questions were processed.

TriviaQA [36] is another well-known dataset used in this research, consisting of question-answer pairs where the answers are derived from an extensive collection of web documents. TriviaQA is known for its complex, real-world questions, often requiring the model to comprehend and synthesize information from multiple sentences or even paragraphs within the context. This dataset was preprocessed to make it compatible with the method and other datasets used in this study.

In addition to these established datasets, **custom**, refered as EduSpecialized, datasets were created for this thesis. This dataset was compiled carefully from eBooks [37], [38] that focuses on natural language processing (NLP) and deep learning, chosen for their relevance to the field. These datasets consist of 2750 data. The datasets comprises several key fields: *ID*, *title*, *context*, *question*, and *answer*. The *ID* field uniquely identifies each entry, while the *title* and *context* provide the necessary background information. The *question* field contains queries related to the provided context, and the *answer* field holds the correct responses per the source material. This custom dataset was particularly valuable in training the T5 model and testing the performances of the models on datasets that specialized in specific

educational content, offering insights into their applicability in domain-specific scenarios. Figure 4.2 shows sample data rows from the dataset.

ID: 1
Title: The Challenge of Information Overload
Context: In 2003, it was estimated that the annual production of books amounted to 8 Terabytes. This enormous volume of information presents a significant challenge for natural language processing as the field aims to develop computational techniques that can efficiently sift through large bodies of text.
Question: What is the estimated annual production of books in 2003?
Answer: 8 Terabytes

ID: 2
Title: Coping with the Information Explosion
Context: The only feasible way for humans to manage the massive influx of information, which is increasingly available electronically, is by utilizing computational techniques that help filter and process vast amounts of data.
Question: How is it suggested humans cope with the information explosion according to the text?
Answer: By utilizing computational techniques that help filter and process vast amounts of data.

ID: 3
Title: Advances in NLP Applications
Context: NLP technologies have seen significant advancements, with applications in machine translation, speech recognition, and sentiment analysis becoming increasingly sophisticated and integral to various industries.
Question: What are some key areas where NLP technologies have advanced?
Answer: Machine translation, speech recognition, and sentiment analysis.

ID: 4
Title: The Role of AI in NLP
Context: Artificial intelligence plays a crucial role in the development of NLP solutions, providing the foundation for systems that can understand, interpret, and generate human language in a way that is both meaningful and contextually appropriate.
Question: What role does artificial intelligence play in NLP?
Answer: Providing the foundation for systems that can understand, interpret, and generate human language in a way that is both meaningful and contextually appropriate.

Figure 4.2 Sample Data

This research aimed to provide a comprehensive evaluation of the models across a range of question types and complexities, thereby ensuring robust and generalizable findings. It utilizes both established benchmarks and a custom-built dataset for this purpose.

5. TESTING AND EVALUATION

5.1. Testing Environment and Data

At this study, some key aspects and strategies for fine-tuning the T5-base is applied: Layer freezing, context related dataset preperation, hyperparameter optimizations used in fine tuning phase. These methods affect the performance of the system. For this reason the models tested in context releavent test data to see whether there is a significant improvement effect on the performance of the model. At the comparison and evaluation phase other LLMs used with the T5 and the fine-tuned T5 models. These LLMs are Gpt3.5, Gpt4o-mini, Bert, Distilbert, Roberta and Flan T5 models.

5.2. Performance Evaluation and Results

In order to evaluate the performance of the Question-Answering (QA) system, a range of metrics are used, such as performance and resource efficiency. Latency and memory usage are the metrics used critically to assess the system's operational feasibility, especially in real-time applications and when there are limited computational resources. Exact Match and F1 Score provide insight into the model's accuracy. While the exact match metric indicates the percentage of precisely correct predictions, the F1 Score measure balance of precision and recall to offer a more nuanced view of performance. It is also observed that models sometimes have problems answering questions. The number of questions left unanswered by the model is an important metric in measuring the reliability of the system. To understand how well the model captures the semantic meaning of the input and whether it produces meaningful answers, it is necessary to evaluate deeper semantic fit. For this purpose, Cosine similarity and BERT score are used to measure semantic similarity between the predicted answer and the ground truth.

BLEU Score and ROUGE Score evaluate the quality of the generated text compared to ground truth text. BLEU Score focuses on the overlap of n-grams. In particular, it shows

how well the model's answer matches the reference answer in terms of word order. ROUGE measures the extent to which the model's response overlaps with the reference response (at the word, sentence level). In QA systems, it evaluates the extent to which the model's response overlaps with the reference response. This metric is particularly useful when considering the length and structural similarity of the response. All in all, these metrics provide a complete evaluation of the QA system, covering aspects from operational efficiency to the semantic quality of the answers generated. This comparison provides quantitative scores that reflect various aspects of the QA system's performance, such as accuracy, relevance, and fluency, thereby guiding further improvements in the system. All in all, these metrics provide a complete evaluation of the QA system, covering aspects from operational efficiency to the semantic quality of the answers generated. This comparison provides quantitative scores that reflect various aspects of the QA system's performance, such as accuracy, relevance, and fluency, thereby guiding further improvements in the system. The evaluation of the results is presented and analyzed to determine the model's effectiveness on different test sets. The results are compared with several other LLMs, such as Bert, the base model and the fine-tuned models to identify the strengths and weaknesses of each model. Performance evaluation metrics are discussed in more detail below.

5.2.1. Metrics

Cosine Similarity: Cosine similarity is a widely used metric to measure the similarity between two non-zero vectors in an inner product space. It is calculated as the cosine of the angle between the two vectors, which can be represented as:

$$cosine\ similarity = \frac{\mathbf{A} * \mathbf{B}}{||\mathbf{A}|| * ||\mathbf{B}||}$$

where **A** and **B** are the vectors being compared.

In traditional cosine similarity, the vectors **A** and **B** are often term frequency vectors, or TF-IDF vectors, constructed from the word counts or term frequencies within each text. Each dimension of these vectors represents a unique word in the vocabulary, and the value in

each dimension represents the frequency or weighted frequency (TF-IDF) of that word in the document. While this approach captures the relative importance of words within documents, it has several limitations. First, the vectors are typically high-dimensional and sparse, leading to inefficiencies in storage and computation. Secondly, traditional cosine similarity doesn't capture the semantic meaning of the words. For instance, synonyms or contextually related words are treated as completely distinct, which might not reflect true textual similarity. Sentence embeddings are dense vector representations of sentences that capture semantic meaning. These embeddings are generated using deep learning models such as BERT, GPT, or models specifically designed for sentence embeddings like Sentence-BERT (SBERT). Unlike traditional word vectors, sentence embeddings consider the context and semantics of the entire sentence. The process of using sentence embeddings for cosine similarity involves two main steps. Firstly, each sentence is converted into a dense vector using a pre-trained model. Secondly, the cosine similarity is computed between these dense vectors. The cosine similarity for sentence embeddings is calculated similarly:

$$cosine\,similarity = \frac{\mathbf{E}_1 * \mathbf{E}_2}{||\mathbf{E}_1|| * ||\mathbf{E}_2||}$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ are the embedding vectors for the two sentences. Using sentence embeddings for cosine similarity offers several advantages. Sentence embeddings are dense vectors, typically of lower dimensions compared to traditional term-frequency vectors, which makes them more efficient for storage and computation. More importantly, sentence embeddings capture semantic relationships between words and phrases, allowing the similarity measure to reflect true textual similarity better. In summary, while traditional cosine similarity uses high-dimensional and sparse vectors that capture term frequency, cosine similarity using sentence embeddings leverages dense vectors that capture the semantic meaning of entire sentences. This makes the latter approach more effective in understanding and comparing the true meaning of textual content. As a result, this metric is used as a performance evaluation metric in this study.

BERT-SCORE: Bert Score is a sophisticated metric used to evaluate the performance of Question-Answering (QA) systems by assessing the semantic similarity between the model's predicted answers and the reference answers. Unlike traditional metrics that rely on exact word matches, the BERT Score leverages the power of contextual embeddings generated by the BERT model.

When using the BERT Score, each word in both the predicted and the reference answers is represented as a dense vector, capturing the word's meaning within its specific context. The similarities of these vectors are calculated, and are compared across the two texts. This metric evaluates the meaning and context of the words used.

The BERT Score is advantageous in the context of QA systems because it recognizes when two answers are semantically similar, even if they do not share the same words. This makes it particularly valuable for evaluating models that generate natural language responses, where slight variations in phrasing should not necessarily count as errors if the underlying meaning is preserved.

In this study, the BERT Score used to measure the QA system's ability to generate meaningful and contextually suitable answers. By focusing on semantic similarity rather than exact matches, BERT Score allowed us to capture the true effectiveness of the model in understanding and responding to queries, making it a crucial component of our overall assessment strategy.

F1 Score: It is a crucial metric used in the evaluation of Question-Answering (QA) systems, especially in scenarios where both precision and recall are important. The F1 Score is the harmonic mean of precision and recall. It provides a single measure that balances the trade-off between these two metrics.

- **Precision** is the ratio of correctly predicted positive examples to the total predicted positive instances. In the context of a QA system, this means how many of the answers provided by the model are correct.

- **Recall**, on the other hand, is the ratio of correctly predicted positive instances to the total actual positive instances, indicating how many of the correct answers were successfully identified by the model.

The F1 Score combines these two measures, precision and recall into one, giving equal weight to both precision and recall. This is particularly useful in QA tasks where it's important not only to retrieve correct answers (precision) but also to ensure that as many correct answers as possible are found (recall).

The F1 Score is particularly important in scenarios with an imbalance between classes, when the costs of false positives and false negatives are both significant. In our QA system evaluation, the F1 Score provided a balanced view of the model's performance, ensuring that the model is not only accurate in its predictions but also comprehensive in its ability to retrieve relevant answers.

The F1 score was used as an evaluation metric, both to minimize errors and to take into account the capacity to obtain the highest possible number of correct answers. For this reason F1 Score is an essential part of performance evaluation.

Latency refers to the time taken to generate a response after receiving a query, the question. It is critical in real-time applications, where faster response times directly improve user experience. The lower the latency, the more efficient the system is.

Memory Usage measures the amount of memory consumed by the model during inference. This includes loading the model and processing the input data. Efficient memory usage is important for deploying the QA system on devices with limited resources, ensuring the model can operate smoothly across various environments.

Exact Match (EM) is a severe metric that calculates the percentage of predictions that exactly match the reference answers. It is critical for assessing the model's precision, with higher Exact Match scores reflecting the system's ability to produce completely accurate answers.

Number of Unanswered Questions is performance metrics that count the number of questions that could not be extracted from the context. In contrast, the ground truth answer exists in the model and cannot generate or extract the answer.

ROUGE: In question answering systems, the ROUGE metric assesses content coverage, ensuring that a systems response encompasses all crucial information present in a reference answer. This is vital, especially in scenarios like educational QA, where the completeness of information provided can significantly impact learning outcomes and user satisfaction. ROUGE effectively measures how much essential content from the reference is captured in the chatbot's response, highlighting areas where critical information might be missing.

BLEU: BLEU is utilized in QA to evaluate the precision and grammatical accuracy of a system's responses by measuring the overlap of words and phrases with those in reference answers. It ensures that the QA system's answers are not only factually correct, but also relevant, detailed, and aligned with the context provided by the question. For these question-and-answer systems, these metrics may be appropriate for evaluating performance. In practice, these metrics would be applied by comparing the responses generated by the system against a set of pre-defined or human-generated responses that are thought ideal or acceptable.

5.2.2. Results

Firstly, the T5 model is fine-tuned and results of this process can be seen in Table 5.1. These test get from running model on a customized context dataset. The T5-1 setup provides an overall performance increase by allowing a larger model update, provides a more balanced model despite the slight increase in latency. The T5-2 setup leeds to a better understanding of language and sentence structures with deeper layers of optimization and Early Stopping. However, this comes with higher latency and slight decreases in some metrics. The fine-tuning operations have generally improved the performance of the model. The T5-1 and T5-2 models have higher F1, Bleu, Rouge, and Bert scores compared to the T5-Base model. However, these improvements have led to an increase in latency. Therefore, if speed

and resource consumption are not critical, the T5-1 and T5-2 models can be preferred. If speed and resource usage are critical as well as performance, the T5-1 model can stand out as a balanced option. Significantly, the significant increase in the Bleu score shows that the T5-2 model produces the best results in terms of grammar, but it should be kept in mind that this model runs slower.

Model/ Metrics	Memory Usage	Latency	F1 Score	Bleu Score	Rouge Score	Bert Score	Cosine Similarity
T5-Base	1.14	0.43	0.44	0.2948	0.5189	0.9209	0.6964
T5-1	1.14	0.45	0.51	0.3617	0.6158	0.9397	0.7852
T5-2	1.14	0.49	0.51	0.6367	0.6152	0.9387	0.7776

Table 5.1 Results of T5 model and fine-tuned models

The following section includes a detailed assessment and comparison of the models across various metrics. This evaluation helps understand how each model performs on three different datasets and under which conditions each model may be most suitable. The comparison and evaluation phase used the T5, Gpt3.5, Gpt4o-mini, Bert, Distilbert, Roberta, and Flan T5 models. Table 5.2 gives information about the load times of models; nearly all models load quickly,

Model	Load Time (sec)
T5	1.85
Bert	1.05
Gpt-3.5-turbo	0.30
Gpt-4o-mini	0.55
Roberta	1.03
Distilbert	0.59
Flan-T5	1.80

Table 5.2 Load time in seconds

Performance measurement and evaluation will be discussed below as order:

Memory Usage

The analysis of memory usage across diverse models Table 5.3 on different datasets reveals important distinctions in resource consumption, which is crucial for understanding the

efficiency and scalability of these models in various applications. Although, T5-1 and T5-2 models show that fine-tuning operations optimize the model parameters and slightly reduce memory usage compared to T5-base, these three have similar performance in terms of memory usage. T5 models may have higher requirements in terms of memory usage, which requires more powerful hardware but in scenarios that require more in-depth analysis it can be preferred.

Bert model required quite high memory usage in all three datasets. This is related to the model having a large number of parameters and its complex structure. Roberta model shows lower memory usage compared to Bert, but still has an above-average memory requirement. Bert and Roberta models can be used in a wide range of training domains due to their high accuracy and ability to understand language. However, the high memory requirement requires powerful hardware when working with more extensive data.

Distilbert offers lower memory usage compared to Bert, while Flan T5-Small also shows a moderate memory requirement. Distilbert's small and efficient structure optimizes memory usage, while Flan T5-Small offers relatively low memory requirements as a lighter version of T5. Distilbert and Flan T5-Small are balanced options in terms of memory usage. In education, they can offer solutions that can work with fast response and low hardware requirements. They can be especially suitable for classroom applications and teaching support systems.

DataSet/Model	**T5-Base**	**T5-1**	**T5-2**	**Bert**	**GPT-3.5**	**GPT-4o-mini**	**Roberta**	**Distilbert**	**Flan T5-Small**
Squad	1.27	1.18	1.18	1.54	0.12	0.12	0.76	0.31	0.78
Trivia	1.27	1.19	1.27	1.65	0.20	0.19	0.91	0.93	0.78
EduSpecialized	1.14	1.14	1.14	1.45	0.11	0.11	0.66	0.48	0.57

Table 5.3 Memory usage of model [GB]

GPT-3.5 and GPT-4o-mini models have the lowest memory usage. They showed very low memory requirements in all three data sets. This shows that these models can work more lightly and efficiently. They are ideal for those looking for fast and low-resource solutions

in educational environments. It can also be preferred in applications that require large-scale data processing or fast response.

Latency (Response Time)

The analysis of the time taken by different models to produce answers on various datasets is shown in Table 5.4. It provides important insights into response times, which are vital for applications that require real-time or near-real-time processing. In the field of education, real-time usage is an important metric.

DataSet/ Model	**T5-base**	**T5-1**	**T5-2**	**Bert**	**GPT 3.5**	**GPT-4o mini**	**Roberta**	**Distilbert**	**FlanT5-Small**
Squad	0.54	0.40	0.40	0.58	0.94	1.61	0.13	0.12	0.30
Trivia	0.54	0.67	0.56	1.35	1.22	1.10	0.34	0.33	0.30
Edu Specialized	0.43	0.45	0.50	0.22	0.98	1.05	0.70	0.04	0.20

Table 5.4 Latency of models [seconds]

T5 models, especially the fine-tuned versions, T5-1 and T5-2, have low latency values on the Squad dataset, indicating that fine-tuning processes speed up the model. However, a slight increase in latency values is observed on the Trivia dataset, which may be due to the complexity of the data. On the specialized dataset, T5 models exhibit similar latency values. In general, T5 models have strong language understanding and response generation capabilities in training environments, but their performance may slow down slightly on complex datasets. In general, T5 models have strong understanding and response general capabilities in training environments, but their performance may slow down slightly on complex datasets. Total time elapsed to answer questions of the Squad dataset can be seen at Table 5.5.

Model	Total Duration(min)
Bert	28.86
Distilbert	5.18
Roberta	6.55
Gpt-3.5-turbo	47.24
T5-base, T5-1, T5-2	26.78, 11.80, 11.67
Flan-T5	14.83
Gpt-4o-mini	80.66

Table 5.5 Total execution time, *squad dataset

For the Bert model, it exhibits high latency values, especially on the Trivia dataset because of its complex structure and the difficulty of the dataset. The Trivia dataset reveals different latency dynamics mainly due to its complexity. Total time elapsed to answer questions Trivia can be seen at Table 5.6. Bert's low latency value on the Edu Specialized dataset indicates that the model works efficiently on specialized datasets. Roberta exhibits a more balanced performance. Roberta's more balanced latency performance provides an advantage in training environments by providing more consistent performance across various datasets.

Model	Total Duration(min)
Bert	67.61
Distilbert	16.92
Roberta	16.92
Gpt-3.5-turbo	61.11
T5-base, T5-1, T5-2	27.02, 19.52, 16.46
Flan-T5	14.83
Gpt-4o-mini	54.76

Table 5.6 Total execution time, *trivia dataset

GPT models generally have higher latency values than other models, which indicates that they require complex calculations. GPT models may be suitable for tasks requiring extensive data processing and detailed analysis. However, due to their high latency values, they may be disadvantaged in educational scenarios that require fast responses.

When it comes to the Distilbert and FlanT5-Small models, they have low latency values. They mainly have lowest latency values on the Edu Specialized and Squad datasets. Their low latency makes them advantageous in speed and efficiency and can provide fast and effective responses in educational environments, especially in classroom applications and teaching support systems.

Exact Match and Unanswerable Questions

When we examine the models comparatively, the T5 models (T5-Base, T5-1, T5-2) stand out as the models with the highest Exact Match (EM) rates, especially in the Squad dataset. Exact match results for each model can be monitored in Table 5.7. While the T5-2 model exhibited the best performance with 78%, all T5 models gave similar results in the Trivia dataset (between 0.43 and 0.44). In the EduSpecialized dataset, the fact that T5-1 and T5-2 performed better than T5-Base reveals the effect of fine-tuning processes. Although Bert exhibited good performance with a 72% EM rate in the Squad dataset, its success decreased in the Trivia and EduSpecialized datasets. The 41% EM rate in the Trivia dataset shows that Bert exhibiteslower success in this dataset. On the other hand, Bert's higher number of unanswered questions than the T5 models is an important factor affecting the overall accuracy performance.

GPT-3.5-turbo and GPT-4o-mini models did not show any Exact Match scores on the Squad and Trivia datasets. This shows that these models are unable to produce answers or provide exact matches with correct answers on the relevant datasets. However, GPT-3.5-turbo achieved a 17% EM rate on the EduSpecialized dataset, indicating that it has shown some success on this dataset. Although GPT models are powerful for more complex and broader knowledge tasks, they underperformed the other models in these tests. Roberta and Distilbert models showed average performance on the Squad dataset with 71% and 65%

Model/Dataset	Squad	Trivia	EduSpecialized
T5-Base	0.77	0.44	0.26
T5-1	0.76	0.43	0.35
T5-2	0.78	0.43	0.35
Bert	0.72	0.41	0.28
Gpt-3.5-turbo	0	0	0.17
Gpt-4o-mini	0	0	0
Roberta	0.71	0.27	0.33
Distilbert	0.65	0.27	0.20
Flan-T5-small	0.70	0.27	0.24

Table 5.7 Average EMs

EM rates, respectively. However, the performance of both models decreases on the Trivia and EduSpecialized datasets. Roberta's performance on the EduSpecialized dataset with 33% EM rate is higher than Distilbert (20%), but overall both models yielded lower results compared to the T5 models.

Flan-T5-small performed similarly to T5 models with an EM rate of 70% on the Squad dataset. However, its performance deteriorated on the Trivia and EduSpecialized datasets, achieving lower Exact Match rates than the other T5 models. Overall, Flan-T5-small, despite being a smaller and lighter model, provided moderate performance on language comprehension tasks.

According to the results in the Table 5.8, when the number of unanswered questions of the models in different datasets is examined, a general performance comparison can be made. The T5 models (T5-Base, T5-1, T5-2) are particularly successful in the Squad and EduSpecialized datasets, as they encountered very few unanswered questions. However, the T5-1 model needed more help in the Trivia dataset. Bert has a very high loss rate, especially in the Trivia dataset (651 unanswered questions), which shows that it needs help in complex datasets. There are also many unanswered questions in the Squad and EduSpecialized datasets. The GPT-3.5-turbo and GPT-4o-mini models did not show any unanswered questions in all datasets, which reveals their strong performance based on extensive knowledge. Roberta and Distilbert, on the other hand, encountered many unanswered questions, especially in the Trivia and EduSpecialized datasets. This shows

that these models experience performance loss in specific datasets. Flan-T5-small, on the other hand, generally gave successful results but needed help to answer a few questions in the Trivia dataset.

In general, GPT models exhibit the most robust performance due to not encountering any unanswered questions. While T5 models also gave good results in general, models such as Bert and Distilbert had difficulty with more questions. In education, models that can produce accurate and comprehensive answers should be preferred; in this context, GPT models and T5 models with a low number of unanswered questions stand out.

Model/Dataset	**Squad**	**Trivia**	**EduSpecialized**
T5-Base	1	1	0
T5-1	1	11	0
T5-2	1	1	0
Bert	26	651	29
Gpt-3.5-turbo	0	0	0
Gpt-4o-mini	0	0	0
Roberta	47	69	33
Distilbert	62	69	47
Flan-T5-small	1	2	0

Table 5.8 Number of unanswered questions

T5 models and Bert have the highest Exact Match rates, while GPT models did not show any EM scores on some datasets. Based on this, the performance of some models is likely affected by the datasets with solid performance. When the number of unanswerable questions and Exact Match results are evaluated together, T5 models show a more balanced performance. In contrast, GPT models show low Exact Match values despite the questions they could not answer. These results show that T5 models are more suitable for higher accuracy and response rates in educational applications.

F1 Score, BLEU and ROUGE Scores

The F1 score evaluation across different models on the Squad, Trivia, and EduSpecialized datasets highlights key performance differences. These results can be seen at Table 5.9. In the Squad dataset, T5 models (T5-Base, T5-1, T5-2) are the best-performing models with an F1 score of 88%. A high F1 score indicates that these models can provide accurate and complete answers. Roberta performs very close to T5 models with 80%, while Distilbert also provides a good result with 76%. Flan-T5-small is a robust model in the Squad dataset with 81%. While Bert performs at a moderate level (55%) in the Squad dataset, GPT models (GPT-3.5-turbo and GPT-4o-mini) have the lowest F1 scores in Squad, indicating that they are lacking in terms of accuracy and sensitivity in this dataset.

In the Trivia and EduSpecialized datasets, T5-1 and T5-2 models generally provide better results. Bert and Roberta exhibit moderate F1 scores in the EduSpecialized dataset. GPT models have low F1 scores on all datasets and underperform on complex and specialized datasets.

This comparison shows that high F1 scores indicate that the models provide accurate and complete answers; that is, the model produces correct results and finds as many correct results as possible, while low F1 scores indicate that these models miss many correct answers or give incomplete answers. Models to be used in training environments should provide both accuracy and sensitivity in a balanced manner; in this context, T5 models and Roberta are generally the most suitable options.

Model/Dataset	Squad	Trivia	EduSpecialized
T5-Base	0.88	0.5	0.26
T5-1	0.88	0.48	0.51
T5-2	0.88	0.49	0.51
Bert	0.55	0.04	0.44
Gpt-3.5-turbo	0.17	0.09	0.28
Gpt-4o-mini	0.15	0.08	0.21
Roberta	0.80	0.31	0.43
Distilbert	0.76	0.31	0.40
Flan-T5-small	0.81	0.30	0.22

Table 5.9 Average F1 Score

The evaluation of the ROUGE and BLEU metrics across the three datasets (Squad-v2, Trivia, and EduSpecialized) highlights the varying capabilities of different models in generating accurate and relevant text responses. The measurements for these metrics are seen in Table 5.10 for Squad-v2, Table 5.11 for Trivia data, and Table 5.12 for custom dataset.

The Bleu score shows how close a model's predicted answer is to the reference sentences. For the Squad dataset, in this metric, T5-1 (62.10%) and T5-Base (61.81%) showed the highest performance. Roberta (56.71%) and Flan-T5 (56.25%) also showed strong results, showing that they were successful in language generation. Distilbert (50.98%) showed an average performance, while Bert (36.61%) was slightly lower. On the other hand, GPT-3.5 (5.38%) and GPT-4o-mini (4.30%), had shallow Bleu scores, indicating that these models were inadequate in terms of language prediction.

The Rouge score measures the overlap of the predicted answer with the reference sentences and evaluates the similarities in sentence structure. In this metric, T5-1 (88.43%) and T5-Base (88.29%) again had the best results. Flan-T5 (81.27%) and Roberta (81.00%) models are also successful regarding Rouge. Distilbert (77.81%) shows an average performance, while Bert (55.84%) has a lower Rouge score. GPT-3.5 (21.95%) and

GPT-4o-mini (19.87%) need to improve in language understanding and sentence structure prediction with very low results.

METRICS (Avg.)	SQUAD-V2								
	Bert	Distilbert	Roberta	Gpt-3.5	T5-base	Flan-T5	Gpt-4o-mini	T5-1	T5-2
Cosine Similarity	0.9800	0.9613	0.9761	0.8495	0.9262	0.9079	0.8478	0.9285	0.9289
Bleu Score	0.3661	0.5098	0.5671	0.0538	0.6181	0.5625	0.0430	0.6210	0.3637
Rouge Score	0.5584	0.7781	0.8100	0.2195	0.8829	0.8127	0.1987	0.8843	0.8842
Bert Score	0.9405	0.9477	0.9566	0.8492	0.9772	0.9672	0.8443	0.9771	0.9772

Table 5.10 Test results of Squad-v2

Regarding Bleu scores in the Trivia dataset, T5-Base (27.28%) received the highest score, the model with the predicted answer closest to the reference sentences among the models used. T5-2 (26.49%) and T5-1 (26.14%) models also achieved strong Bleu scores. Distilbert (17.19%) and Roberta (17.12%) also achieved moderate results, while Flan-T5 (15.52%) performed lower. Bert (3.00%), GPT-3.5 (1.47%), and GPT-4o-mini (1.32%) models failed to provide the expected language prediction performance in this dataset with shallow Bleu scores.

Regarding the Rouge score in Trivia dataset, T5-Base (49.58%) again received the highest result, showing that the predicted answers were the most successful regarding structural overlap with the reference sentences. T5-2 (48.72%) and T5-1 (47.90%) also showed similar strong performance. Distilbert (30.69%) and Roberta (30.69%) provided average results, while Flan-T5 (30.60%) showed close performance to these two models. GPT-3.5 (13.72%) and GPT-4o-mini (13.84%) failed to provide successful results in terms of sentence structures, with low Rouge scores. Bert (4.14%) had the lowest Rouge score in the Trivia dataset. Regarding the Bleu and Rouge metrics, T5 models have superior language modeling ability on the Trivia dataset, while other models lag.

METRICS (Avg.)	TRIVIA								
	Bert	Distilbert	Roberta	Gpt-3.5	T5-base	Flan-T5	Gpt-4o-mini	T5-1	T5-2
Cosine Similarity	0.7860	0.7855	0.7318	0.8456	0.7027	0.5803	0.8458	0.6925	0.6972
Bleu Score	0.03	0.1719	0.1712	0.0147	0.2728	0.1552	0.0132	0.2614	0.2649
Rouge Score	0.0414	0.3069	0.3069	0.1372	0.4958	0.3060	0.1384	0.4790	0.4872
Bert Score	0.6784	0.8845	0.8845	0.8261	0.9173	0.8712	0.8253	0.9159	0.9159

Table 5.11 Test results of Trivia

In the EduSpecialized dataset, T5-1 (36.17%) and T5-2 (36.37%) models have the highest Bleu scores and show the best performance in terms of language production. T5-Base (29.48%) and Roberta (29.07%) also yielded successful results. Bert (28.02%) and Distilbert (23.02%) have moderate Bleu scores. Flan-T5 (21.47%) and GPT-3.5 (12.86%) showed lower performance, while GPT-4o-mini (6.78%) was weak in language production with shallow Bleu scores.

When looking at Rouge scores, T5-1 (61.58%) and T5-2 (61.52%) again show the highest performance. These models are the most successful in matching sentence structures with reference sentences. Roberta (%52.23), Bert (%51.81), and T5-Base (%51.89) also provide quite successful results in terms of Rouge scores. Distilbert (%48.67) and GPT-3.5 (%32.56) show moderate performance, while Flan-T5 (%25.65) and GPT-4o-mini (%23.20) have relatively low Rouge scores and are weak in language comprehension and sentence structure generation.

METRICS (Avg.)	EduSpecialized								
	Bert	Distilbert	Roberta	Gpt-3.5	T5-base	Flan-T5	Gpt-4o-mini	T5-1	T52
Cosine Similarity	0.9281	0.9034	0.8219	0.8694	0.6964	0.4162	0.8604	0.7852	0.7776
Bleu Score	0.2802	0.2302	0.2907	0.1286	0.2948	0.2147	0.0678	0.3617	0.3637
Rouge Score	0.5181	0.4867	0.5223	0.3256	0.5189	0.2565	0.2320	0.6158	0.6152
Bert Score	0.9132	0.9002	0.8135	0.9061	0.9209	0.8734	0.8745	0.9397	0.9387

Table 5.12 Test results of EduSpecialized

Cosine Similarity, BERT Score

The measurements for these metrics are seen in Table 5.10 for Squad-v2, Table 5.11 for Trivia data, and Table 5.12 for custom dataset. While the Bert score was high in some tests, the Cosine similarity is relatively low, which means that the context is correct in the prediction process, but the words are expressed differently. Similarly, when the Cosine similarity was high, and the Bert score was low, the word choices in the prediction process were the same or similar to the reference sentence. However, it means the context is incorrect, incomplete, or meaningless.

In the Squad-v2 dataset, all models show strong performance in terms of Cosine Similarity and BERT Score. T5 models lead with the highest BERT Score, indicating that its generated responses are very close to the reference text in terms of meaning and linguistic structure. Roberta and Flan-T5 also perform well, with BERT Scores showing their capability to produce high-quality responses. Cosine Similarity is similarly high across these models, with Bert achieving 0.9800, indicating very high semantic alignment between the generated and reference texts. GPT-3.5-turbo and GPT-4o-mini have lower BERT Scores and Cosine Similarity, around 0.8495 and 0.8478 respectively, suggesting they are less effective in maintaining semantic consistency.

For the Trivia dataset, the performance diverges more noticeably. T5 again leads with a BERT Score of 0.9173. However, its Cosine Similarity drops which means that it generates responses are generally accurate. But the answers may differ more in word choice or phrasing. Bert and Distilbert also maintain strong BERT Scores, though their Cosine Similarity suggests that Bert struggles more with semantic alignment in this dataset. Roberta and Flan-T5 score similarly, with moderate Cosine Similarity around 0.7318 and BERT Scores in the high 0.88s; this indicates they are reliable but less consistent performance than T5-Base. GPT-3.5-turbo and GPT-4o-mini show strong Cosine Similarity; their lower BERT Scores suggest a trade-off between semantic similarity and overall textual quality.

In the EduSpecialized dataset, T5 models performs well with the highest BERT Score and a moderate Cosine Similarity . This suggests that while T5-Base maintains good alignment with the reference text, the specialized nature of the dataset presents more challenges in achieving high semantic similarity. Bert and Distilbert, while having slightly lower BERT Scores (around 0.84 and 0.83, respectively), also show a significant drop in Cosine Similarity, indicating that these models may struggle with the unique demands of specialized content. Roberta and Flan-T5 show similar patterns, with their BERT Scores indicating good overall performance, but their lower Cosine Similarity suggests a need for further fine-tuning to handle specialized datasets better.

Across all datasets, T5-Base consistently demonstrates strong performance in both BERT Score and Cosine Similarity, making it a reliable model for generating high-quality, semantically aligned text. Roberta and Flan-T5 also show strong capabilities, particularly in less specialized contexts like Squad-v2, but they may require further adjustment to excel in more challenging or specialized datasets. Bert and Distilbert perform well in terms of BERT Score but show more variability in Cosine Similarity, particularly in more complex datasets like Trivia and EduSpecialized, indicating that while they can generate text with good overall quality, maintaining semantic consistency can be challenging. GPT-3.5-turbo and GPT-4o-mini, though they perform adequately in terms of Cosine Similarity, have lower BERT Scores, suggesting that while their responses may be semantically similar to the reference texts, they may lack the overall textual quality seen in other models.

Based on the comprehensive analysis of various metrics, including memory usage, latency, exact match, F1 scores, ROUGE, BLEU, Cosine Similarity, and BERT Score, T5-Base stands out as the most versatile model for educational applications. Its consistently high performance across datasets suggests it is well-suited for tasks requiring accuracy and efficiency, such as automated grading, personalized tutoring, and content generation in educational platforms. T5-Base's strong F1 scores, high ROUGE and BLEU metrics, and reasonable memory usage and latency, make it ideal for real-time applications where precise and semantically rich responses are critical.

Roberta and Flan-T5-small also show potential, particularly when quick, low-latency responses are needed, such as in classroom interactive systems or chatbots for student support. These two models offer a good balance between performance and resource efficiency, making them suitable for limited computational resources.

While effective in less complex tasks, Bert and Distilbert may be better suited for specific use cases like educational content tagging or simpler question-answering systems where the demand for high exact matches and nuanced understanding is lower. Their variability in performance across datasets suggests they struggle with more specialized or complex educational content.

GPT-3.5-turbo and GPT-4o-mini, despite their lower scores in many metrics, could be leveraged in scenarios where the focus is on generating creative content or brainstorming, where semantic accuracy is less critical. However, their higher latency and memory usage might limit their applicability in real-time educational tools.

Overall, T5-Base is the most reliable choice for comprehensive educational applications requiring a blend of accuracy, efficiency, and adaptability, while Roberta and Flan-T5-small offer strong alternatives for more resource-constrained environments.

6. CONCLUSION

In conclusion, this thesis demonstrates the potential of leveraging advanced deep learning techniques to develop a sophisticated Question Answering (QA) system for educational purposes. By employing and evaluating multiple models, including T5-Base, fine-tuned versions of the T5-base, Roberta, Bert, and others, the research has shown that it is possible to create a QA system capable of providing accurate, context-aware responses to student queries. This may result in an improvement in the learning experience.

The T5 model, in particular, is a highly practical choice due to its consistent performance across various metrics, including F1 score, ROUGE, BLEU, and BERT Score. This makes it well-suited for tasks requiring both precision and adaptability. Additionally, T5 models consistently outperformed others like GPT-3.5 and GPT-4o-mini in accuracy and response quality, especially in specialized datasets such as EduSpecialized and Trivia. This model's ability to handle diverse and complex queries with relatively low latency and efficient memory usage highlights its potential for integration into real-time educational tools. When evaluated from many perspectives, it provides a balanced use. This study made specific modifications to the T5 model, particularly in its fine-tuning approach, layer freezing, and optimization for educational queries. These changes have resulted in notable improvements in accuracy and adaptability across specialized datasets, showcasing the model's enhanced ability to deliver precise and contextually accurate responses. Fine-tuning the T5 model for specific educational domains can help students and educators access information quickly and accurately. The T5 model is a well-balanced choice regarding the resources it requires, the speed of generating answers, and its accuracy.

One of the key insights from this study is the difference between surface-level similarity and proper contextual understanding, as highlighted by the evaluation of Cosine Similarity and BERT Scores. While some models like GPT-3.5 and GPT-4o-mini exhibited relatively high Cosine Similarity, their lower BERT Scores revealed weaknesses in understanding the deeper context of the questions. This emphasizes the importance of selecting models that not

only provide semantically similar responses but also capture the underlying meaning, which is crucial in an educational context. T5 models consistently achieved high BERT Scores, demonstrating their ability to balance surface similarity and contextual understanding, making them superior in handling complex educational queries.

The development process, including model selection, training, and fine-tuning, has been meticulously documented, providing a valuable resource for future research and development in this field. The evaluation framework established in this thesis ensures that the QA system's performance can be systematically considered and improved over time, particularly as educational content and students' and educators' needs evolve.

Furthermore, this work emphasizes the broader implications of AI-driven educational tools, suggesting that such systems can significantly contribute to personalized learning and student engagement. While the T5 model currently offers a robust solution, future research directions include expanding the system's language capabilities, exploring the integration of more specialized models like Flan-T5-small for resource-constrained environments, and investigating the potential of emerging models to enhance performance further. Models like Flan-T5-small or Distilbert, despite their lower memory usage and relatively efficient processing speeds, could be valuable in environments where computational resources are limited.

Overall, this study contributes to the constant evolution of AI in education, demonstrating the feasibility and benefits of integrating sophisticated QA systems into learning environments. In this study, it is seen how important the model selection is depending on the existing problem and resources. Likewise, performance evaluation in developing tools to support and enhance the educational experience effectively is crucial. This work highlights the balance between semantic similarity (Cosine Similarity) and deeper contextual understanding (BERT Score) when developing QA systems, ensuring that educational tools are accurate and context-aware in delivering information to students.

6.1. Future Directions

The future of Question Answering system implementation in education lies in exploring more efficient and context-aware models, while also focusing on reducing computational costs and improving the interpretability of AI decisions during interactions. Investigating hybrid approaches that combine rule-based systems with advanced AI-driven models could provide a balanced solution, enhancing both the reliability and adaptability of the system. Additionally, further research into optimizing models like T5-Base for specialized educational content and expanding their language capabilities could significantly enhance their effectiveness in diverse learning environments.

REFERENCES

[1] B.F. Green, A.K. Wolf, C. Chomsky, and K. Laughery. Baseball: An automatic question-answerer. In *Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference*, pages 219–224. **1961**.

[2] Sotiris B Kotsiantis, Ioannis Zaharakis, and Panayiotis Pintelas. Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 160(1):3–24, **2007**.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112. **2014**.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008. **2017**.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. New Orleans, Louisiana, **2018**.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*. **2013**.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Doha, Qatar, **2014**.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, **2018**. OpenAI.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **2019**.

[10] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, **2020**.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, **2020**.

[12] Ernest W. Brewer. Chatbots in the classroom: A survey of educational chatbots for higher education. *Journal of Educational Technology Systems*, 38(4):309–325, **2010**.

[13] Li Deng and Dong Yu. Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, **2014**.

[14] Guang Ruan and Li Zeng. Application of artificial intelligence chatbots in education: A systematic review. *International Journal of Educational Technology in Higher Education*, 16(1):1–15, **2019**.

[15] Daniel Jurafsky and James H Martin. *Speech and Language Processing*. Pearson, 3rd edition, **2020**.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, **2015**.

[17] Yuchen Wang and Li Deng. Deep learning techniques for spoken language understanding. *Proceedings of the IEEE*, 109(6):934–965, **2021**.

[18] Percy Liang, Christopher Potts, and Daniel Jurafsky. Learning latent semantic structures with neural networks. *Artificial Intelligence*, **2021**.

[19] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai: A survey. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298, **2021**.

[20] Heung-Yeung Shum, Xiaodong He, and Di Li. Eliza: An ai-driven conversational agent for mental health support. *Nature Reviews Neuroscience*, 19(11):683–695, **2018**.

[21] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, **1972**.

[22] W.A. Woods, R.M. Kaplan, and B.L. Nash-Webber. The lunar sciences natural language information system: Final report. Technical Report BBN Report 2378, Bolt Beranek and Newman Inc., **1972**.

[23] Boris Katz. Annotating the world wide web using natural language. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 136–143. **1997**.

[24] Daniel Jurafsky and James H Martin. *Speech and language processing*, volume 3. Pearson, **2008**.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, **2016**.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, **1997**.

[27] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, **2015**.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, **2019**.

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, **2016**.

[30] Zeynep Sanli. Development and implementation of an educational question-answering system using the t5-small model, **2024**. Unpublished Master's Thesis.

[31] Yao Zhang, Jie Fu, Zhen Han, and Yun Zhao. Comparative study of transformer-based models in educational qa systems. *Journal of Educational Data Mining*, 15(2):102–123, **2023**.

[32] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1–10. **2020**.

[33] Jesse Dodge, Q Vera Liao, Yiling Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13. **2021**.

[34] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. **2021**.

[35] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, **2018**.

[36] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, **2017**.

[37] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Stanford University and University of Colorado at Boulder, Stanford, CA and Boulder, CO, third edition, **2023**. Draft of January 7, 2023. Comments and typos welcome!

[38] Palash Goyal, Karan Jain, and Sumit Pandey. *Deep Learning for Natural Language Processing: Creating Neural Networks with Python*. Apress, Bangalore, Karnataka, India, **2018**. ISBN 978-1-4842-3684-0. doi:10.1007/978-1-4842-3685-7. ISBN-13 (electronic): 978-1-4842-3685-7.