Navigating Complexity: A Tailored Question-Answering Approach for PDFs in Finance, Bio-Medicine, and Science

Teerath Kumar *, Rutu Bhujbal, Kislay Raj, Arunabha M. Roy *

Posted Date: 17 October 2024

doi: 10.20944/preprints202410.1395.v1

Keywords: Question-Answering; Bidirectional Encoder Representations; Bio-medicine; Finanace; Transformers)

 doi:10.20944/preprints202410.1395.v1

Article

Navigating Complexity: A Tailored Question-Answering Approach for PDFs in Finance, Bio-Medicine, and Science

Teerath Kumar [1,2,*,†], Rutuja Bhujbal [2,†], Kislay Raj [1] and Arunabha M. Roy [3]

[1] School of Computing, Dublin City, University, Ireland; teerath.menghwar2@mail.dcu.ie, kislay.raj2@mail.dcu.ie
[2] School of Computing, National College of Ireland, Dublin, Ireland ; rutujabhujbal01@gmail.com
[3] Aerospace Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA; arunabhr.umich@gmail.com
* Correspondence: teerath.menghwar2@mail.dcu.ie
† These authors contributed equally to this work.

Abstract: Understanding complex Portable Document Format (PDF) files, such as research papers, clinical reports, and scientific manuals, is often a time-consuming endeavor. While significant progress has been made in developing question-answering (QA) systems that yield contextually relevant responses, the creation of a comprehensive end-to-end machine learning model capable of addressing intricate questions remains a formidable challenge. These systems typically rely on substantial labeled training data to effectively train their foundational models for specific tasks. However, assembling such datasets is particularly challenging for complex documents, including annual reports from major technology companies. In this paper, we address this issue by developing a QA system specifically designed for PDF documents, focusing on the domains of finance, biomedicine, and scientific literature. We manually curated datasets from these areas for evaluation purposes and utilized pre-trained Bidirectional Encoder Representations from Transformers (BERT) models from the Hugging Face library. The models were evaluated using the F1 score, achieving a notable score of 44% with the BERT Large model.

Keywords: question-answering; bidirectional encoder representations; bio-medicine; finanace; transformers

1. Introduction

Deep learning has emerged as a transformative force across a diverse array of domains, including image processing [1–7], natural language processing (NLP) [8–11], and audio processing [12–15]. In particular, the realm of question-answering systems has witnessed remarkable advancements in recent years. These sophisticated algorithms are instrumental in efficiently extracting pertinent information from vast quantities of text, enabling users to receive precise and contextually relevant answers to their queries. Despite the strides made in end-to-end training, the challenge of answering complex questions with a singular machine-learning model persists. This research paper aims to bridge this gap by developing a question-answering system tailored specifically for PDF files within the realms of scientific, financial, and biomedical literature [16–19]. The intricate nature of these documents often necessitates exhaustive analysis, making them difficult to comprehend. A robust question-answering system has the potential to alleviate this burden, allowing users to swiftly access the information they need. Previous studies have demonstrated that machine learning models can successfully retrieve information from extensive documents, with their performance typically assessed using metrics such as exact matches and F1 scores. Among these, Bidirectional Encoder Representations from Transformers (BERT) have shown exceptional prowess in deep learning tasks, including quality control and text summarization, requiring minimal modifications beyond pre-training with just one additional output layer. This capability could be particularly beneficial in medical contexts, assisting healthcare professionals in researching disease symptoms and treatment options. Notably, pre-trained BERT models like Bio-BERT and Sci-BERT have outperformed traditional models, aligning with findings in prior research on unsupervised biomedical question-answering pre-training.

This study aspires to enhance our understanding of the effective utilization of transformers in the development of question-answering software for the specified PDF documents. Ultimately, a well-

 doi:10.20944/preprints202410.1395.v1

designed QA system presents a more intelligent, efficient, and user-friendly approach to extracting information from extensive PDFs, such as annual reports, academic articles, and medical documents.

Research Question: The challenges outlined above lead us to the following research question:

In the financial, biomedical, and scientific domains, how can BERT transformers be employed to effectively answer questions and retrieve knowledge from PDF files? This research aims to implement a QA system using pre-trained BERT models tailored for these domains and evaluate their performance to identify the most effective model. An overview of the question-answering system utilizing domain-specific pre-trained BERT on documents is depicted in Figure 1.

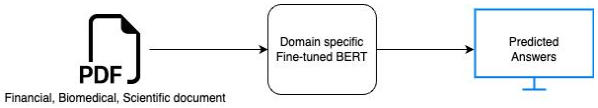

Figure 1. Question-answering system with fine-tuned BERT

The structure of this paper is organized as follows: Section 2 discusses related work, Section 3 outlines the research methodology, Section 4 elaborates on design specifications, Section 5 illustrates the implementation aspects, Section 6 examines evaluation results, and finally, Section 7 highlights conclusions and future work.

2. Related Work

BERT (Bidirectional Encoder Representations from Transformers) has emerged as a groundbreaking framework in the field of Natural Language Processing (NLP), achieving remarkable results across a multitude of tasks. Its design, characterized by conceptual simplicity and empirical robustness, positions BERT as a prime candidate for advancing question-answering systems. This research focuses on developing a BERT-based question-answering system tailored for PDF documents, aiming to navigate the intricate challenges posed by the extraction of nuanced information from this ubiquitous format. Given that PDF files are prevalent in both academic and professional contexts, their diverse structures and complex formatting often hinder effective information retrieval. By leveraging BERT's capabilities, we seek to enhance comprehension and accessibility in the domain of PDF-based question-answering systems.

Our literature review aims to explore existing research that integrates BERT models for question-answering in PDFs, identifying gaps and opportunities for innovation. This examination is intended to propel the development of intelligent systems for document comprehension, providing critical insights into the design and implementation of a BERT-based QA system optimized for PDF files.

A seminal contribution to the foundation of this work is the introduction of the Transformer architecture by Vaswani et al. [20]. This architecture revolutionized sequence transduction models by utilizing attention mechanisms, effectively overcoming the limitations associated with conventional recurrent neural networks. The Transformer architecture significantly outperforms traditional encoder-decoder frameworks in terms of speed and efficacy, particularly in language translation tasks, where it outshines even previously established ensemble models.

Further exploration into the capabilities of pre-trained language models was undertaken by Zhao et al. [21], who assessed their generalization across a spectrum of question-answering datasets. These datasets, varying in complexity, present challenges that require models to navigate different levels of reasoning. This study involved training and fine-tuning various pre-trained language models on diverse datasets, revealing that models like RoBERTa and BART consistently outperform others. Notably, the BERT-BiLSTM architecture demonstrated enhancements over the baseline BERT model, emphasizing the significance of bidirectionality in developing robust systems for nuanced reasoning.

In a distinct application, Müller et al. (2023) introduced Covid-Twitter-BERT (CT-BERT), which was pre-trained on Twitter data related to Covid-19. This study reinforces the effectiveness of domain-specific models, demonstrating how tailored architectures can better address specific tasks [23]. An-

 doi:10.20944/preprints202410.1395.v1

other innovative approach, presented by Huang et al. [24], expands the BERT architecture to account for inter-cell connections within tables. By utilizing a substantial table corpus from Wikipedia, this method retrains parameters to better capture associations between tabular data and surrounding text, thereby enhancing the efficiency and accuracy of question-answering in documents.

Despite BERT's notable success in language comprehension, its adaptation for language generation tasks remains a challenge. The research conducted by Zhang et al. [25] introduced a novel method termed C-MLM, which modifies BERT for target generation tasks. This innovative approach employs BERT as a "teacher" model, guiding traditional Sequence-to-Sequence (Seq2Seq) models, referred to as "students." This teacher-student dynamic significantly enhances the performance of Seq2Seq models in various text generation tasks, yielding substantial improvements over existing Transformer baselines.

Further contributions to BERT's capabilities are highlighted in the work of Devlin et al. [26], where a dual-context training method is employed. This technique jointly trains the model on both left and right contexts across all layers, enabling the pre-training of deep bidirectional representations from unlabeled text. By fine-tuning the pre-trained BERT model with minimal architectural modifications, advanced models for tasks such as question answering and language inference can be developed, achieving an impressive F1 score of 93 percent on the SQuAD v1.1 dataset.

In the realm of disaster management, a pioneering automated approach utilizing BERT for extracting infrastructure damage information from textual data has been proposed [28]. This innovative method, trained on reports from the National Hurricane Center, demonstrates high accuracy in scenarios involving hurricanes and earthquakes, outperforming traditional methods. The approach involves two key steps: Paragraph Retrieval using Sentence-BERT, followed by information extraction with a BERT model. The model was trained on 533 question-answer pairs extracted from hurricane reports, achieving F1 scores of 90.5 percent and 83.6 percent for hurricane and earthquake scenarios, respectively.

In the domain of document classification, an extensive study [29] has pioneered the application of BERT, achieving state-of-the-art results across four diverse datasets. Despite initial concerns regarding computational demands, the proposed BERT-based model surpasses previous baselines by utilizing knowledge distillation to create smaller bidirectional LSTMs, achieving comparable performance with significantly reduced parameters.

Furthermore, BERTSUM, a simplified variant of BERT specifically designed for extractive summarization, has demonstrated its capability in enhancing summarization tasks [30]. This research underscores BERT's robust architecture and extensive pre-training dataset, providing compelling evidence for its effectiveness in extractive summarization, especially given recent advancements that had previously plateaued.

In parallel, the introduction of novel data augmentation techniques leveraging distant supervision has been explored to enhance BERT's performance in open-domain question-answering tasks [31]. This approach addresses challenges related to noise and genre mismatches in distant supervision data, illustrating the sensitivity of models to diverse datasets and hyper-parameters.

A comprehensive survey [32] analyzed various adaptations of BERT, including BioBERT for biomedical texts, Clinical BERT for clinical notes, and SciBERT for scientific literature. These models have showcased superior performance in their respective fields, with recommendations for future research focusing on their application in tasks such as summarization and question answering [33].

Moreover, research tackling the overload of biomedical literature has employed contextual embeddings from models such as XLNet, BERT, BioBERT, SciBERT, and RoBERTa for keyword extraction, yielding significant improvements in F1 scores [35]. Studies have also highlighted the efficacy of BioBERT and SciBERT in processing biomedical text [36], while another investigation noted that ALBERT outperforms BERT with fewer parameters and faster training times in natural language understanding benchmarks [38].

The comprehensive literature review reveals that researchers have achieved significant accuracy with BERT-based models across various NLP tasks, particularly in the domain of question-answering

 doi:10.20944/preprints202410.1395.v1

[21,28]. Notable studies have illustrated the efficacy of models like Roberta for processing complex PDF documents, as well as the importance of domain-specific adaptations such as BioBERT and SciBERT in the biomedical context. The findings underscore the versatility and robustness of BERT, while also illuminating ongoing challenges, including model sensitivity to varying datasets, scalability issues, and the complexities involved in applying BERT to generative tasks. These insights lay the groundwork for future research aimed at refining BERT-based models and enhancing their applicability across diverse domains.

3. Methodology

This section describes methods for implementing a PDF-based question-answering system using BERT base models. BERT is an excellent choice for question-answering (QA) on PDF documents for several reasons. Due to its extensive pre-training on a huge corpus of text material, BERT can acquire an in-depth contextual understanding of language. Understanding the context is essential to understanding the rich and varied content that may be found in PDF documents. Its bidirectional attention mechanism considers both the left and right context for each word in a document[26]. This approach is effective for capturing dependencies and relationships within the text, which is essential for accurate question-answering. A pre-trained BERT model can be used as the base model for QA answering. Further fine-tuning the pre-trained model with question-answer pairs specific to PDFs implements the QA system. BERT has a substantial number of parameters, and fine-tuning the pre-trained model with a small collection of question-answer pairs would result in overfitting. To avoid that, a fine-tuned BERT is used, which was trained on the SQAuD data set.

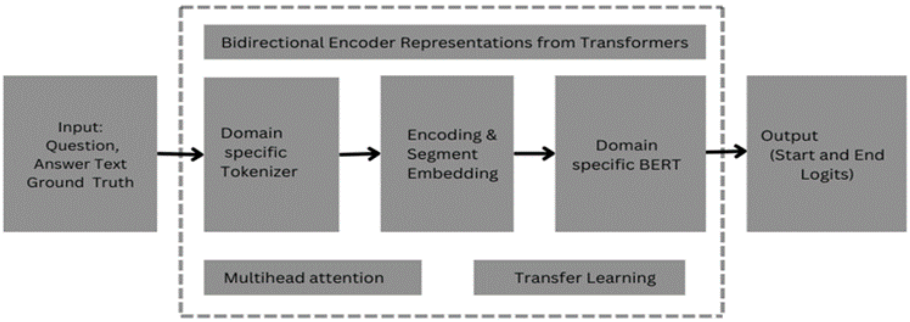

Figure 2. Question answering system using fine-tuned BERT

3.1. Data Collection

Links to download the PDFs are given in the configuration manual.

3.1.1. Financial Domain

For the financial domain, annual reports from Amazon were collected as representative documents. The PDFs were obtained from official sources, ensuring the authenticity and relevance of the financial data.

3.1.2. Biomedical Domain

In the biomedical domain, research papers related to COVID-19 and diabetes were selected for analysis. The data set includes papers from reputable journals and conferences. ensuring a diverse and comprehensive coverage of biomedical information. PDFs were downloaded from Google Scholar.

 doi:10.20944/preprints202410.1395.v1

3.1.3. Scientific Domain

Scientific literature documents were sourced from various research papers related to question-answering systems. These papers were selected to represent the breadth of scientific literature and were obtained from Google Scholar.

3.2. Question-Answer Pairs Dataset Creation

3.2.1. Financial Domain

For the financial domain, a data set of question-answer pairs was manually curated. Questions were formulated to cover various aspects of financial reports and the corresponding Answers were extracted from relevant sections of the annual reports. The data set includes the 'question,' 'context,' and 'ground truth' columns, where 'context' represents The document chunk and 'ground truth' provide the correct answer.

3.2.2. Biomedical Domain

In the biomedical domain, a similar approach was taken to create question-answer pairs related to COVID-19 research papers. Questions were designed to capture key biomedical information and answers were extracted from the respective document chunks. The data set structure includes 'question,' 'context,' and 'ground truth' columns.

3.2.3. Scientific Literature Domain

The creation of question-and-answer pairs for the scientific literature domain followed a similar methodology. Questions were tailored to cover diverse scientific topics, and answers were extracted from relevant chunks of scientific papers. The data set structure includes 'question,' 'context,' and 'ground truth' columns.

3.3. Pre-Processing

When employing BERT or other transformer-based models for question answering, pre-processing is essential. Pre-processing ensures that the input text is appropriate for BERT models that are made to handle text in a specific way. The following justifies the requirement for pre-processing:

3.3.1. Text Extraction

For the financial domain, a data set of question-answer pairs was manually curated. Questions were formulated to cover various aspects of financial reports and the corresponding Answers were extracted from relevant sections of the annual reports. The data set includes the 'question,' 'context,' and 'ground truth' columns, where 'context' represents the document text chunk and 'ground truth' provide the correct answer.

3.3.2. Lower-Casing and Stripping

To maintain consistency and reduce redundancy, all text was converted to lowercase. Leading and trailing white spaces were removed to enhance the uniformity of the data. BERT was trained on a large amount of lowercase text. For the optimal performance of the BERT, lower-casing of the text was a necessary pre-processing step.

3.3.3. Sentence Cleaning

Prior to model training and evaluation, it is crucial to pre-process the raw text data to enhance the quality and relevance of information. This involves cleaning sentences to ensure uniformity and remove noise. The following functions were employed for sentence cleaning:

 doi:10.20944/preprints202410.1395.v1

- clean_sentence(): This function is designed to clean up individual sentences. Converts the sentence to lowercase. Removes special characters using regular expressions. Optionally removes stop-words, leveraging the gensim library's remove stop-words function.
- get_cleaned_ sentences(): This function applies the clean sentence function to a list of sentences. The optional parameter removes the stop-words flag and controls whether stop-words are removed from the sentences.

3.3.4. Convert Sentences into Tokens

BERT employs a particular technique that divides text into smaller pieces known as tokens. The input text is split into words or sub-words, and an embedding vector is given to each token. The PDF text is split into tokens using the nltk_sent_tokenize method, which makes tokens of the text.

3.3.5. Chunking Strategy

To overcome the token limit of BERT (512 tokens), the PDFs were pre-processed by breaking them into smaller chunks. Each chunk was then split into tokens using the appropriate BERT-based model for the respective domain (BERT base for financial, SciBERT for scientific literature, and BioBERT for biomedical).

3.4. Model Implementation

3.4.1. Model Selection

A critical first step in implementing a QA system is choosing suitable pre-trained models. In this section, details of the models chosen for each domain and the rationale behind these selections are given and they were hugely inspired by the literature review conducted [9,10]. BERT base and BERT large models were utilized for the financial domain. For the scientific literature domain, SciBERT, a BERT model pre-trained on scientific text, was employed. In the biomedical domain, BioBERT, pre-trained on biomedical literature, was used. For the financial domain, two variants of BERT models were utilized: BERT Base Model: A base BERT model was employed to capture general financial information and nuances. BERT Large Model: A larger version of BERT was utilized to grasp more complex financial patterns and relationships within the text.

- Financial Domain For the financial domain, two variants of BERT models were utilized: BERT Base Model: A base BERT model was employed to capture general financial information and nuances. BERT Large Model: A larger version of BERT was utilized to grasp more complex financial patterns and relationships within the text.
- Biomedical Domain In the biomedical domain, a specialized BERT model pre-trained on biomedical literature, known as BioBERT, was chosen. BioBERT is appropriate for the study of COVID-19 research publications since it is designed to comprehend the distinct terminologies and ideas found in biomedical texts.

- Scientific Literature Domain For the scientific literature domain, we utilized SciBERT, a BERT model pre-trained on a diverse range of scientific texts. SciBERT is designed to capture the intricacies of scientific language, making it suitable for extracting information from research papers and scientific literature.

3.4.2. Model Fine-Tuning

To adjust a pre-trained BERT model to a specific task or domain, fine-tuning entails training the model on a domain-specific data set. With hundreds of millions to more than 300 million parameters, BERT is an extensive neural network architecture. Thus, Overfitting would occur if a BERT model were trained from scratch on a small data set. A refined, pre-trained BERT model that was trained on a sizable data set is preferable. Using data from the Stanford Question Answering data set (SQuAD), the BERT model has been improved.

 doi:10.20944/preprints202410.1395.v1

3.5. Question Answering Setup

The task of answering questions was framed as identifying relevant information within the chunks. Domain-specific BERT models for QA were used for question-answer pairs created for each data set.

3.6. Evaluation

3.6.1. Metrics

The models' performance was assessed using the F1 score, an accepted measure for question answering. Initially, a confidence score was also used to check how confident the model was in predicting answers. F1 score is a popular and extensively used measurement in quality assurance for classification problems. In cases where we value recall and precision equally, it is appropriate. The foundation of the F1 score is the number of words that are shared between the prediction and the truth; recall is the ratio of shared words to the total number of words in the ground truth, and precision is the ratio of shared words to the total number of words in the prediction.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Figure 3. Calculation of F1 score

3.6.2. Cross-Domain Evaluation

To assess the models' generalizability, cross-domain evaluation was performed by testing fine-tuned BERT large models on datasets from other domains. This helps understand the adaptability and transferability of the models across diverse types of documents. When the performance of fine-tuned BERT was tested on the Biomedical and Scientific domains, the following insights were driven: 1. For the biomedical domain, both BERT large and Bio-clinical BERT gave partial answers for some question-answer pairs and no answers for a few pairs. 2. For the scientific domain, both BERT large and SciBERT models predicted partially correct answers.

3.6.3. Comparative Analysis

Comparisons were made between the performance of the BERT large model within each domain. Additionally, insights were drawn from the cross-domain evaluation to identify potential areas for improvement.

4. Design Specification

In this section, the foundational elements underpinning the implementation of the BERT-based QA system, catering specifically to the distinct characteristics of the financial, biomedical, and scientific domains.

4.1. Techniques

BERT-based QA system integrates several key techniques to address the unique challenges posted by diverse domains: Domain-Specific Fine-tuned BERT: For each of the domains, a domain-specific, fine-tuned BERT model was employed.

- Domain-Specific Fine-Tuned BERT: For each of the domains, a domain-specific fine-tuned BERT model was employed.
- Transfer Learning: Transfer learning in BERT (Bidirectional Encoder Representations from Transformers) involves leveraging pre-trained models on large corpora and fine-tuning them for specific downstream tasks. Google's BERT algorithm has demonstrated impressive results across a range of natural language processing (NLP) applications. The key idea behind transfer

 doi:10.20944/preprints202410.1395.v1

learning in BERT is to utilize the pre-trained knowledge encoded in the model's parameters and adapt it to a particular task or domain with limited labeled data [19].

4.2. Architecture

4.2.1. Multi-Head Attention Mechanism

Multi-Head Attention Mechanism: A multi-head attention mechanism in the architecture [20] enables the model to focus on different parts of the input text at the same time. This is especially beneficial for capturing complex relationships and context within diverse domain-specific documents.

4.2.2. Domain-Specific Embeddings

We utilize domain-specific embeddings to augment the pre-trained BERT embeddings. These embeddings are tailored to the vocabulary and context prevalent in the financial, biomedical, and scientific domains.

4.3. Framework

Implementation is built on the PyTorch framework, providing a robust and flexible platform for deep learning.

- PyTorch Transformers Library: The PyTorch Transformers library was used, which facilitates seamless integration with pre-trained BERT models. This library offers a comprehensive set of tools for tokenization, model configuration, and training.

4.4. Algorithm Description

4.4.1. Algorithm Functionality

BERT-based QA system for financial, biomedical, and scientific domains introduces the following functionalities: Document Chunking Strategy: Given the potentially lengthy and complex nature of documents in these domains, our system employs a document chunking strategy to handle large texts efficiently, ensuring that relevant context is preserved.

4.4.2. Algorithm Requirements

To implement and deploy a QA system successfully, certain requirements must be met: Pre-processing Modules: Custom pre-processing modules are designed to handle data cleaning, tokenization, and embedding generation. Hardware Acceleration: For optimal performance, the system benefits from hardware acceleration, such as GPUs provided by Google Collab to expedite training and inference.

4.5. Tools and Languages

The implementation leveraged the following tools and languages: Programming Language: Python's extensive libraries and versatility in the fields of data science and machine learning led to its selection as the main programming language. Machine Learning Frameworks: Machine learning models were implemented and trained with the help of the scikit-learn, TensorFlow, PyTorch, and Hugging Face Transformers libraries.

5. Implementation

In the final stage of the implementation, the input sequences are prepared for processing by the model, adhering to the model's specific input requirements and constraints.

5.1. Domain-Specific Implementation

Each domain requires a tailored approach to model implementation due to the distinct characteristics of the data. Below, we provide an overview of the strategies employed in each domain.

 doi:10.20944/preprints202410.1395.v1

5.1.1. Financial Domain

The financial domain involves the analysis of annual reports from Amazon. The fine-tuned BERT base and BERT large models were used on a curated dataset of financial question-answer pairs. The stages of implementation of the QA system are as follows:

- With pipeline library: When models were implemented with pipeline library, confidence scores for both BERT base and BERT large models were extremely low, even though the answers were correct.
- With tokenization and segmentation: Models were implemented with different approaches where pre-processed input question and answer text were tokenized using the pre-trained tokenizer. The tokenized input is then segmented into question-and-answer segments. A pre-trained model was trained with the tokenized and segmented input to estimate the beginning and ending positions of the response within the input text. Post-processing is used to handle any spaces at the start of the answer tokens after model inference. The final answer is reconstructed by concatenating these tokens.

- With chunking strategy: Chunking strategy refers to the process of breaking down a large document, such as a PDF, into smaller chunks or segments to be processed by a model. Due to the limitation of 512 tokens Bert models were not efficient for long documents as they will only consider the first 512 tokens. To overcome that limitation input text was stripped into chunks of 512 tokens and then fed to the model in a loop. While chunking can be effective in handling lengthy documents, it comes with certain limitations: Context Discontinuity: Breaking a document into chunks may result in the loss of contextual information that spans across different chunks. BERT models use context to interpret words, so if a question's pertinent context is divided into two chunks, the model's performance might be impacted. Answer Span Across Chunks: Sometimes a question's answer can be found in more than one section. If the model processes each chunk independently, it might miss the context necessary to identify the correct answer span that extends beyond a single chunk. Incoherent Context: The chunks processed in isolation might not provide coherent context, leading to potential misunderstandings by the model. Since BERT is meant to record contextual relationships between words, breaking up the text into smaller sections might cause this continuity to be broken. Increased Complexity: Chunking introduces additional complexity into the pre-processing and post-processing stages. Managing the boundaries of chunks and ensuring a seamless flow of information between them requires careful handling.

- With a Curated Data set of question-answer pairs: A data set of ten examples was created manually from PDFs containing question, context, and ground truth columns. This data set in CSV format was then read as a data frame and fed to the model to calculate the F1 score. This strategy overcomes the following limitations of the chunking method: Context Preservation: The curated data set contains question-answer pairs carefully crafted to ensure that the context necessary for answering the questions is preserved. In contrast, chunking large documents may introduce discontinuities in context, potentially affecting the model's performance. Reduced Complexity: Utilizing a curated data set might simplify the training process compared to managing the complexities introduced by chunking. Dealing with context boundaries, overlaps, and potential information loss associated with chunking can be challenging. A curated data set of question-answer pairs has additional advantages as follows: Training Data Quality: If a curated data set is well-constructed and diverse, it provides a clean and controlled environment for training the model. The model learns from specific examples that are explicitly designed for the task, which can be beneficial in terms of generalization to similar scenarios. Task Relevance: If a task is well-represented in the curated data set, and the questions and answers cover a diverse range of scenarios, a model may perform better compared to a model trained on chunks

 doi:10.20944/preprints202410.1395.v1

of documents. This is particularly true if the curated data set is domain-specific or tailored to the types of documents. Reduced Complexity: Utilizing a curated data set might simplify the training process compared to managing the complexities introduced by chunking. Dealing with context boundaries, overlaps, and potential information loss associated with chunking can be challenging. Evaluation: Curated data sets often come with predefined evaluation metrics and benchmarks, like ground truth, making it easier to assess the model's performance and compare it against other models in the field. Efficiency: Training on a curated data set may be computationally more efficient than training on large, chunked documents, especially if the documents are extensive.

- Fine-tuning of DistilBERT on SQAuD data set: To fine-tune the pre-trained BERT model, the Trainer class from the PyTorch library was utilized. A small subset of the SQAuD data set was loaded from the data sets library and was split into train test data sets using the train test split method. Then DistilBERT, a distilled version of BERT was loaded to process question and answer. The data set was pre-processed to truncate the context and map the answer tokens to the context. Map function from the data set library was used to apply pre-processing to the entire data set. A batch of examples were created using Data Collator. The next step was to define hyper-parameters in training arguments such as learning rate, number of epochs, and weight decay. After that, the trainer was given training arguments that included the model, data set, tokenizer, and data collator. The train function was called to fine-tune the model. This fine-tuned model was saved and used for inference for the financial data set.

5.1.2. Biomedical Domain

In the biomedical domain, fine-tuned BioBERT was applied to COVID-19 research papers. The Curated data set for biomedical PDFs was tested on the model to predict the answers. A data set containing question, context, and ground truth columns was used to calculate the F1 score.

5.1.3. Scientific Literature Domain

For the scientific literature domain, SciBERT was employed to analyze scientific research papers. A curated scientific dataset was used to calculate the F1 score.

6. Evaluation

This section presents an in-depth evaluation of the findings from the experimental research conducted in each domain. The analysis focuses on the most relevant findings that contribute to addressing the research question.

6.1. Financial Domain

6.1.1. Case Study 1

Implementation of QA system with QA pipeline library. In this case study, the implementation of a Question Answering (QA) system using a dedicated QA pipeline library is explored. The Hugging Face Transformers library was used to perform question-answering tasks using two different models: ”bert-large-uncased-whole-word-masking-fine-tuned-squad” and ”bert-base-uncased.” Pre-processed text from Amazon's annual report was fed to the QA pipeline as context and model were evaluated with a confidence score that measures model confidence. Comparison of the question-answering performance of two different BERT models on a specific question and context helps evaluate how the choice of model can impact the quality of answers provided by the question-answering system. The large model's answer was more relevant and contextually appropriate for the given question about the document's topic. The higher score of 0.44 indicates a higher confidence level compared to the base model, which provided a less relevant answer with a significantly lower score.

 doi:10.20944/preprints202410.1395.v1

Table 1. Confidence score of BERT base and BERT large

Model	Score
BERT base	3.44E-05
BERT large	0.441348344

6.1.2. Case Study 2

: Implementation of QA system with Tokenization and segment embeddings of text and questions were tokenized using the tokenizer's encoding method. The [SEP] token index separated the question and answer segments. Segment IDs were created, assigning 0s to segment A (question) and 1s to segment B (answer). The tokenized input and segment IDs were passed to the model to obtain outputs. The start and end indices of the predicted answer were determined, and the answer span was constructed by concatenating the corresponding tokens. The score was calculated as the maximum value of the start index, but it does not provide a direct measure of the model's confidence or certainty in the predicted answer. When a small text from a PDF was tested, the model predicted the answer correctly (refer to Table 2). Further functions were modified to incorporate the F1 score to measure the model's predictions. The limitation was that BERT could only consider 512 tokens. So, this method was not useful for longer documents.

Table 2. Predicted answer for question pair with Tokenization strategy

Question	Context	Predicted Answer
How much was the net sales in the year 2022?	Net sales increased 13% to $143.1 billion in the third quarter, compared with $127.1 billion in the third quarter of 2022.	"$127.1 billion"

6.1.3. Case Study 3

Implementation of QA with chunking strategy. Further input sequences were divided into chunks of 510 and special tokens [CLS] and [SEP] were added to separate the question and answer. Zero-padding was done to ensure consistent sizes. For each chunk, the answer question function was called with the question, and the chunk's tokens were converted back to a string as the answer text. Table 3 shows the predicted answer by this strategy.

Table 3. Predicted answer for question pair with chunking strategy

Question	Context	Predicted Answer
How does AWS help Amazon grow in the year 2022?	PDF text from Amazon's quarterly report	Segment sales increased 12% year-over-year to $23.1 billion. Operating income increased to $11.2 billion in the third quarter, compared with $2.5 billion in the third quarter of 2022. North America segment operating income was $4.3 billion, compared with an operating loss of $0.4 billion in the third quarter of 2022.

6.1.4. Case Study 4

Implementation of QA with curated data set. A curated data set from the PDF text was created manually to test the model's performance for multiple questions. This approach overcomes the limitation of the chunking strategy which could cause a loss of context and it was more efficient in evaluating long text as well.

 doi:10.20944/preprints202410.1395.v1

Table 4. Predicted answers and F1 scores for curated financial data set

Question	Context	Ground Truth	Predicted Answer	F1 Score
What were Amazon's net sales in the first quarter of 2023?	PDF Text	Net sales increased 9% to $127.4 billion in the first quarter, compared with $116.4 billion in the first quarter of 2022.	$127.4 billion	0.039
How much did net sales increase compared to the first quarter of 2022?	PDF Text	Excluding the $2.4 billion unfavorable impact from year-over-year changes in foreign exchange rates throughout the quarter, net sales increased 11% compared with the first quarter of 2022.	9%	0.0
How did North America segment sales change year-over-year?	PDF Text	North America segment sales increased 11% year-over-year to $76.9 billion.	Foreign exchange rates	0.199
What was the operating income for the AWS segment?	PDF Text	AWS segment operating income was $5.1 billion, compared with an operating income of $6.5 billion in the first quarter of 2022.	$5.1 billion	0.0
How did the operating cash flow change for the trailing twelve months?	PDF Text	Operating cash flow increased 38% to $54.3 billion for the trailing twelve months, compared with $39.3 billion for the trailing twelve months ended March 31, 2022.	Net sales increased 9%	0.074

6.1.5. Case Study 5

Implementation of QA with fine-tuned DistilBERT. The next step was to see if fine-tuning the BERT model improves the score. The distilling version of BERT was loaded and fine-tuned with hyper-parameters learning rate=1e-5, num train epochs=3, per device train batch size=8, per device eval batch size=8. When the fine-tuned model was inference for simple QA pairs, the confidence score was 0.250225812,(Table 5). The question-answer pairs of the financial data set were inferences to evaluate the performance. The results are given in Table 6.

Table 5. Result for small question-answer pair on fine-tuned BERT

Question	Context	Predicted Answer	Score
What are different search engines?	BLOOM has 176 billion parameters and can generate text in 46 natural languages and 13 programming languages.	176 billion	0.250225812

Table 6. Results for Financial dataset on fine-tuned BERT

Question	Context	Predicted Answer	Score
What were Amazon's net sales in the first quarter of 2023?	PDF text	$127.4 billion	0.07002584
How much did net sales increase compared to the first quarter of 2022?	PDF text	$127.4 billion	0.070099174
What was the impact of foreign exchange rates on net sales?	PDF text	$116.4 billion	0.070099174
How did North America segment sales change year-over-year?	PDF text	$116.4 billion	0.080088
How did the operating cash flow change for the trailing twelve months?	PDF text	$127.4 billion	0.03641737

 doi:10.20944/preprints202410.1395.v1

6.2. Bio-Medical Domain

6.2.1. Case Study 6

Implementation of QA using pre-trained Bio-BERT with curated data set. PDF text from the Covid research paper was fed to the Bio-Clinical BERT pre-trained on biomedical and clinical text. A curated data set of 10 question-answer pairs was tested on the model and evaluated with an F1 score. The results are given in Table 7.

Table 7. Comparison of F1 scores of BioBERT and BERT Large for biomedical data set

QA Pairs	Bio-ClinicalBERT F1-Score	Bio-ClinicalBERT Average F1	BERT Large F1-Score	BERT Large Average F1 Score
1	0.035	0.033	0.028	0.043
2	0.037		0.083	
3	0.030		0.038	
4	0.034		0.040	
5	0.032		0.122	
6	0.026		0.022	
7	0.031		0.022	
8	0.038		0.025	
9	0.023		0.024	
10	0.040		0.025	

6.2.2. Case Study 7

Implementation of QA using pre-trained BERT large with curated data set. For cross-domain evaluation, the biomedical data set was then tested on a fine-tuned BERT large model to assess the model's generalizability to other domains and to investigate which model performs the best for biomedical documents.

6.3. Scientific Domain

6.3.1. Case Study 8

Implementation of QA using pre-trained Sci-BERT with curated data set. Pre-trained Sci-BERT trained on a large corpus of scientific literature, including scholarly articles, research papers, and other documents from the bio-medical and life sciences domains. The model was tested for F1 scores.

6.3.2. Case Study 9

Implementation of QA using pre-trained BERT large with curated data set. To test the pre-trained BERT model's generalizability, a scientific data set was also evaluated on the BERT large model. Results are given in Table 8.

Table 8. Comparison of F1 scores of SciBERT and BERT Large for scientific literature data set

Q A Pairs	SciBERT F1-Score	SciBERT Avg F1	BERT Large F1-Score	BERT Large Avg F1
1	0.050	0.053	0.026	0.053
2	0.029		0.110	
3	0.091		0.058	
4	0.070		0.045	
5	0.044		0.054	
6	0.062		0.029	
7	0.038		0.031	
8	0.036		0.071	
9	0.054		0.034	
10	0.060		0.068	

 doi:10.20944/preprints202410.1395.v1

7. Discussion

Developing a QA system for PDF is a challenging task since PDF may contain complex text, tables, images, or complex layouts. PDFs related to the financial, biomedical, and The scientific sector is even more complex and time-consuming to comprehend. To utilize pre-trained BERT models for the respective domains, experiments were carried out to implement a QA system for the chosen PDFs. From case study 1, BERT Large gives a 44% score (see Table 1) for the pre-processed financial PDF text. The research was successful in implementing a QA system for the smaller texts. As we can see from Case Study 2, the model predicts the answer correctly (see Table 2). The chunking strategy implemented in Case Study 3 successfully overcomes the limitation of 512 tokens in the BERT model (see Table 3). The limitation of case study 3 was overcome in case study 4 with a curated data set that preserves the context (see Table 4). Case Study 5 implemented fine-tuning of pre-trained DistilBERT. The model's confidence score was slightly higher for the simple question-answer pairs (see Table 5) than for the complex texts (see Table 6). However, as the research progressed to design an end-to-end QA system on longer PDFs, the following limitations were found during the experiments:

1. Chunking text into 512 tokens is only useful for small PDFs. Amazon's annual reports used for analysis are 16 pages long and the prototype developed here lacks the implementation for longer PDFs. Additionally, this could result in context loss and incoherence in answer generation while processing multiple chunks.
2. Low F1 scores for the curated data set for respective domains as shown in Table 9 suggest that the proposed research needs optimization and should consider fine-tuning the curated data set.
3. A data set was created for each domain using only a few pages of the PDFs.Creating data sets manually for fine-tuning and evaluation is challenging.
4. When cross-domain evaluation was conducted in case studies 6,7,8 and 9 did not show much difference. As described in the literature review, previous studies show domain-specific BERT models have achieved significant results for respective domains.

Table 9. Average F1 scores for the respective domains.

Domain	BERT Model Used	Average F1 Score
Financial	BERT Large Uncased	0.03913
Biomedical	Bio-ClinicalBERT	0.033
Scientific	SciBERT	0.053

8. Conclusion and Future Work

This study intended to utilize the pre-trained BERT models for implementing a QA system on PDFs from various domains. Several strategies were used to implement a QA system for financial, scientific, and bio-medical domains. The proposed research successfully implemented a question-answering pipeline with a pre-trained BERT base and BERT large models. For longer documents, chunking the long text into chunks of 512 and extracting answers from the chunks was implemented successfully. Data sets were created manually for evaluation for the chosen domains with question, context, and ground truth columns. These data sets were tested on different BERT models like BioClinical BERT, SciBERT, BERT large, and DistilBERT. This research poses few limitations such as lower confidence score of BERT models even after fine-tuning with hyper-parameters. The creation of correct data sets manually from PDFs was also challenging and needs to be addressed for better evaluation of the models. This research holds the potential to utilize personalized chatbots for various fields like education, medicine, and finance. This research can be extended in the future for the improvisation of the model's confidence score and the creation of question-answer pairs from complex PDFs.

 doi:10.20944/preprints202410.1395.v1

References

1. Aleem, S., Kumar, T., Little, S., Bendechache, M., Brennan, R. & McGuinness, K. Random data augmentation based enhancement: a generalized enhancement approach for medical datasets. *ArXiv Preprint ArXiv:2210.00824*. (2022)
2. Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Tataei Sarshar, N., Tirkolaee, E., Ali, S., Kumar, T. & Bendechache, M. ME-CCNN: Multi-encoded images and a cascade convolutional neural network for breast tumor segmentation and recognition. *Artificial Intelligence Review*. pp. 1-38 (2023)
3. Roy, A., Bhaduri, J., Kumar, T. & Raj, K. A computer vision-based object localization model for endangered wildlife detection. *Ecological Economics, Forthcoming*. (2022)
4. Kumar, T., Park, J., Ali, M., Uddin, A. & Bae, S. Class specific autoencoders enhance sample diversity. *Journal Of Broadcast Engineering*. **26**, 844-854 (2021)
5. Kumar, T., Turab, M., Talpur, S., Brennan, R. & Bendechache, M. FORGED CHARACTER DETECTION DATASETS: PASSPORTS. *DRIVING LICENCES AND VISA STICKERS*.
6. Kumar, T., Mileo, A., Brennan, R. & Bendechache, M. RSMDA: Random Slices Mixing Data Augmentation. *Applied Sciences*. **13**, 1711 (2023)
7. Kumar, T., Brennan, R. & Bendechache, M. Stride Random Erasing Augmentation. *CS & IT Conference Proceedings*. **12** (2022)
8. Kumar, T., Park, J., Ali, M., Uddin, A., Ko, J. & Bae, S. Binary-classifiers-enabled filters for semi-supervised learning. *IEEE Access*. **9** pp. 167663-167673 (2021)
9. Singh, A., Ranjbarzadeh, R., Raj, K., Kumar, T. & Roy, A. Understanding EEG signals for subject-wise definition of armoni activities. *ArXiv Preprint ArXiv:2301.00948*. (2023)
10. Turab, M., Kumar, T., Bendechache, M. & Saber, T. Investigating multi-feature selection and ensembling for audio classification. *ArXiv Preprint ArXiv:2206.07511*. (2022)
11. Kumar, T., Turab, M., Raj, K., Mileo, A., Brennan, R. & Bendechache, M. Advanced Data Augmentation Approaches: A Comprehensive Survey and Future directions. *ArXiv Preprint ArXiv:2301.02830*. (2023)
12. Kumar, T., Turab, M., Mileo, A., Bendechache, M. & Saber, T. AudRandAug: Random Image Augmentations for Audio Classification. *ArXiv Preprint ArXiv:2309.04762*. (2023)
13. Chandio, A., Shen, Y., Bendechache, M., Inayat, I. & Kumar, T. AUDD: audio Urdu digits dataset for automatic audio Urdu digit recognition. *Applied Sciences*. **11**, 8842 (2021)
14. Kumar, T., Park, J. & Bae, S. Intra-Class Random Erasing (ICRE) augmentation for audio classification. *Korean Society Of Broadcasting And Media Engineering Conference Proceedings*. pp. 246-249 (2020)
15. Park, J., Kumar, T. & Bae, S. Search for optimal data augmentation policy for environmental sound classification with deep neural networks. *Journal Of Broadcast Engineering*. **25**, 854-860 (2020)
16. Roy, A., Bhaduri, J., Kumar, T. & Raj, K. WilDect-YOLO: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*. **75** pp. 101919 (2023)
17. Khan, W., Raj, K., Kumar, T., Roy, A. & Luo, B. Introducing urdu digits dataset with demonstration of an efficient and robust noisy decoder-based pseudo example generator. *Symmetry*. **14**, 1976 (2022)
18. Chandio, A., Gui, G., Kumar, T., Ullah, I., Ranjbarzadeh, R., Roy, A., Hussain, A. & Shen, Y. Precise single-stage detector. *ArXiv Preprint ArXiv:2210.04252*. (2022)
19. Singh, A., Raj, K., Kumar, T., Verma, S. & Roy, A. Deep learning-based cost-effective and responsive robot for autism treatment. *Drones*. **7**, 81 (2023)
20. Adhikari, A., Ram, A., Tang, R. and Lin, J. (2019). Docbert: Bert for document classification, arXiv preprint arXiv:1904.08398
21. K. Pearce, T. Zhan, A. Komanduri, and J. Zhan, "A Comparative Study of Transformer-Based Language Models on Extractive Question Answering," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.03142
22. W. Zaghouani, I. Vladimir, and M. Ruiz, "COVID-Twitter-BERT: A natural language processing model to analyze COVID-19 content on Twitter." [Online]. Available: https://github.com/digitalepidemiologylab/covid-twitter-bert
23. E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.03323

24. V. Zayats, K. Toutanova, and M. Ostendorf, "Representations for Question Answering from Documents with Tables and Text," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.10573
25. Y.-C. Chen, Z. Gan, Y. Cheng, J. Liu, and J. Liu, "Distilling Knowledge Learned in BERT for Text Generation," Association for Computational Linguistics.
26. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805
27. S. Wadhwa, K. R. Chandu, and E. Nyberg, "Comparative Analysis of Neural QA models on SQuAD," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.06972
28. Y. Kim, S. Bang, J. Sohn, and H. Kim, "Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers," Autom Constr, vol. 134, Feb. 2022, doi: 10.1016/j.autcon.2021.104061.
29. A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for Document Classification," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.08398
30. Y. Liu, "Fine-tune BERT for Extractive Summarization," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.10318
31. W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data Augmentation for BERT Fine-Tuning in Open-Domain Question Answering," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06652
32. A. H. Mohammed and A. H. Ali, "Survey of BERT (Bidirectional Encoder Representation Transformer) types," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jul. 2021. doi: 10.1088/1742-6596/1963/1/012173.
33. A. H. Mohammed and A. H. Ali, "Survey of BERT (Bidirectional Encoder Representation Transformer) types," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jul. 2021. doi: 10.1088/1742-6596/1963/1/012173.
34. I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.10676
35. A. Celikten, A. Ugur, and H. Bulut, "Keyword extraction from biomedical documents using deep contextualized embeddings," in 2021 International Conference on Innovations in Intelligent Systems and Applications, INISTA 2021 - Proceedings, Institute of Electrical and Electronics Engineers Inc., Aug. 2021. doi: 10.1109/INISTA52262.2021.9548470.
36. V. Kommaraju et al., "Unsupervised Pre-training for Biomedical Question Answering," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.12952
37. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11942
38. M. Namazifar, A. Papangelis, G. Tur, and D. Hakkani-T¨ur, "Language model is all you need: Natural language understanding as Question answering," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7803–7807. doi: 10.1109/ICASSP39728.2021.9413810.
39. C. Tao, S. Gao, M. Shang, W. Wu, D. Zhao, and R. Yan, "Get the point of my utterance! Learning towards effective responses with multi-head attention mechanism," in IJCAI International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, 2018, pp. 4418–4424. doi: 10.24963/ijcai.2018/614.