

Configuration Manual

MSc Research Project
Data Analytics

Rutuja Bhujbal
Student ID: x22123822

School of Computing
National College of Ireland

Supervisor: Prof. Teerath Kumar Menghwar

National College of Ireland
Project Submission Sheet
School of Computing



Student Name:	Rutuja bhujbal
Student ID:	22123822
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Teerath Kumar Menghwar
Submission Due Date:	14th December 2023
Project Title:	Configuration Manual
Word Count:	751
Page Count:	11

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rutuja Bhujbal
Date:	14th December 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission , to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Rutuja Bhujbal
x22123822

1 Introduction

The code used to implement the project "Building a question-answering system to extract information from PDF files using BERT transformers" is described in detail in this file, along with the hardware and software requirements.

2 System configuration

Your second section. Change the header and label to something appropriate.

2.1 Hardware

- Processor: AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz
- 16.0 GB (15.3 GB usable) ,12.7 GB System RAM on Google Colab
- System type: 64-bit operating system, x64-based processor
- Hard Disk Storage: 100GB (Google Drive Storage)

2.2 Software

- Google Colab), Overleaf, Microsoft Excel, Google Scholar
- Browser Engine: Google Chrome/ Microsoft edge.
- Email: Gmail login to access Google Colab .

3 Project Development

3.1 Project management tool

- Google Drive with 100 GB of storage was used for creating a project environment. Domain-specific PDFs used for analysis were stored in separate folders.
- All notebooks were mounted on Google Drive and saved in the Colab Notebooks folder.

3.2 Installed libraries

Libraries required for implementing question-answering systems were installed at the beginning of each notebook. There are a total of six Colab notebooks. Some of the libraries installed were: transformers, Scikit Learn, PyPDF2, genism, pdf Plumber Devlin et al. (2018) Pearce et al. (2021)

```
!pip install pdfplumber # Python library for extracting information from PDF
!pip install PyPDF2 # Library for reading and manipulating PDF files in Python.
!pip install nltk # Library provides tools for tasks such as tokenization, stemming, lemmatization
!pip install -U gensim # Library for topic modeling and document similarity analysis
!pip install accelerate==0.20.3 # Library for GPU acceleration in Python
!pip install transformers -U # Library for working with pre-trained models in transformer-based models

import pdfplumber
import re
import gensim
from gensim.parsing.preprocessing import remove_stopwords
from transformers import pipeline
import torch
from transformers import BertTokenizer
from transformers import BertForQuestionAnswering
from sklearn.metrics import f1_score
import re
import pandas as pd
import nltk
```

Figure 1: Libraries installed in Financial Notebook

4 Design Flow

Perform design guidelines mentioned below such as 1) Data understanding 2) data pre-processing 3) Logic implementation for building question-answering system 4) Evaluation of system on Financial, Biomedical, and scientific PDF Dataset Alsentzer et al. (2019) Beltagy et al. (2019)

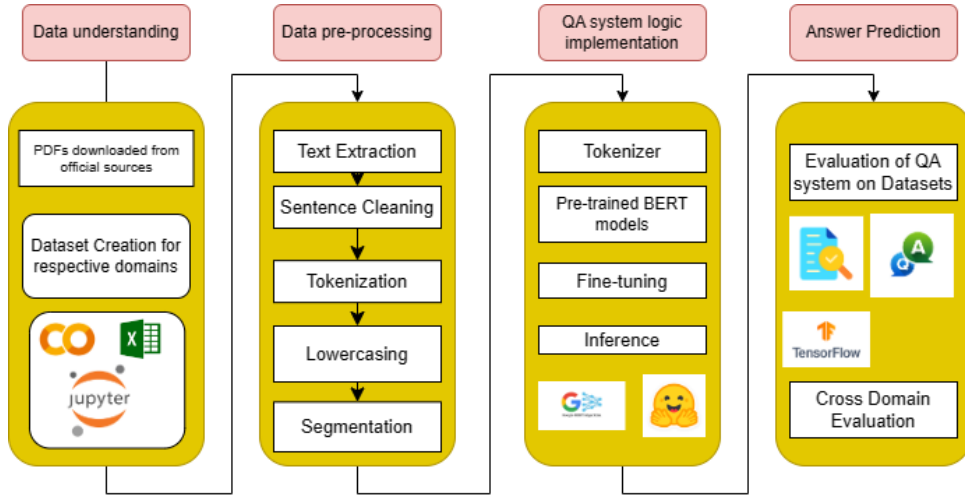


Figure 2: Design flow of building QA system on PDFs

5 Data Collection

PDFs for Financial and Biomedical domains were downloaded from the following sources respectively.

- Amazon’s annual reports: Amazon.com Announces Third Quarter Results

- covid Research paper- Biomedical Research COVID-19 Impact Assessment
- For the scientific literature domain research papers used for the literature review of this research were analyzed. Papers were downloaded from Google Scholar. Google Scholar Search

6 Data Pre-processing

This step involves PDF text extraction, text cleaning, tokenization, and segmentation.

```
[ ] pdf = pdfplumber.open("/content/gdrive/MyDrive/Colab Notebooks/PDF_Folder_Financial/Q1-2023-Amazon-Earnings-Release.pdf")
print(len(pdf.pages))
pdf_txt= ""
for i in range(len(pdf.pages)):
    pdf_txt +=pdf.pages[i].extract_text()
pdf.close()
```

Figure 3: PDF text Extraction

```
import nltk
nltk.download('punkt')
tokens = nltk.sent_tokenize(pdf_txt)
for t in tokens:
    print(t, "\n")
```

Figure 4: Sentence Tokenization

```
def clean_sentence(sentence, stopwords=False):
    sentence = sentence.lower().strip()
    sentence = re.sub(r"[a-z0-9]", '', sentence)
    if stopwords:
        sentence = remove_stopwords(sentence)
    return sentence

def get_cleaned_sentences(tokens, stopwords=False):
    cleaned_sentences = []
    for row in tokens:
        cleaned = clean_sentence(row, stopwords)
        cleaned_sentences.append(cleaned)
    return cleaned_sentences

[ ] cleaned_sentences = get_cleaned_sentences(tokens, stopwords=True)
cleaned_sentences_with_stopwords = get_cleaned_sentences(tokens, stopwords=False)
print(cleaned_sentences)
print(cleaned_sentences_with_stopwords)

['amazoncom announces quarter results seattlebusiness wire october 26 2023amazoncom nasdaq amzn today announced financial results quarter ended septem
['amazoncom announces third quarter results\nseattlebusiness wire october 26 2023amazoncom inc nasdaq amzn today announced financial\nresults for its
```

Figure 5: Sentence cleaning

7 Model Implementation Logic

The stepwise model implementation is given below:

7.1 with pipeline library

Implementation of a QA system with a Pipeline library using a pre-trained BERT base and BERT large

```
# Experiment with different models
qa_pipeline_large = pipeline("question-answering", model="bert-large-uncased-whole-word-masking-finetuned-squad")
qa_pipeline_base = pipeline("question-answering", model="bert-base-uncased")

# Assign cleaned sentences to context
context = " ".join(cleaned_sentences)
question = "What is the topic of the document?"

# Get answers from different models
answer_large = qa_pipeline_large(question=question, context=context)
answer_base = qa_pipeline_base(question=question, context=context)

# Print answers and scores
print("Large Model Answer:", answer_large["answer"])
print("Large Model Score:", answer_large["score"])

print("Base Model Answer:", answer_base["answer"])
print("Base Model Score:", answer_base["score"])
```

Figure 6: Implementation of a QA system with a Pipeline

```
Large Model Answer: corporate corruption racial injustice
Large Model Score: 0.4413483440876007
Base Model Answer: supply chain management support sustainability targets abdul
Base Model Score: 3.550049223122187e-05
```

Figure 7: Outputs produced with pipeline library

7.2 with pre-trained BERT function

Implementation of a QA system with a Pipeline library using a pre-trained BERT base and BERT large

```
def answer_question(question, answer_text):
    input_ids = tokenizer.encode(question, answer_text, max_length=512, truncation=True)
    print('Query has {:,} tokens.\n'.format(len(input_ids)))
    sep_index = input_ids.index(tokenizer.sep_token_id)
    num_seg_a = sep_index + 1
    num_seg_b = len(input_ids) - num_seg_a
    segment_ids = [0]*num_seg_a + [1]*num_seg_b
    assert len(segment_ids) == len(input_ids)
    outputs = model(torch.tensor([input_ids]), token_type_ids=torch.tensor([segment_ids]))
    start_index = torch.argmax(outputs.start_logits)
    end_index = torch.argmax(outputs.end_logits)
    all_tokens = tokenizer.convert_ids_to_tokens(input_ids)
    score = float(torch.max(start_index))
    answer_start = torch.argmax(start_index)
    answer_end = torch.argmax(end_index)
    tokens = tokenizer.convert_ids_to_tokens(input_ids)
    answer = tokens[start_index]
    for i in range(start_index + 1, end_index + 1):
        if tokens[i][0:2] == ' ':
            answer += tokens[i][2:]
        else:
            answer += ' ' + tokens[i]
    return answer, score
```

Figure 8: Implementation of a QA system with a BERT large

7.3 With curated Financial dataset

The modified answer_question function is called with the curated dataset to evaluate the F1 score. Refer manual dataset creation step to generate the Financial dataset question-answer pairs.

```

answer_text = """The ability of the machine to infer knowledge from
the user documents can be tested based on its ability to answer
the question asked. Conventional Artificial Neural network
(ANN) models for knowledge extraction only answer to the
questions which are simple and objective as they don't analyze
the questions and don't try to understand what really the content
of document mean. The proposed question answering system
(QAS) uses deep cases along with ANN to understand the
contents of the documents. We divide the sentences of natural
language into knowledge units and assign deep case to each word
to improve the quality of knowledge extraction."""

question = "How to improve the quality of knowledge extraction?"
ground_truth = "divide the sentences of natural language into knowledge units and assign deep case to each word"

question1 = "What is the capital of France?"
answer_text1 = "The capital of France is Paris."
ground_truth1 = "France"

question2 = "What is Artificial Intelligence?"
answer_text2 = "AI, or Artificial Intelligence, refers to the simulation of human intelligence in machines that are programmed to think and learn like
ground_truth2 = "simulation of human intelligence"

```

Figure 9: set of of question-answer pairs

```

▶ answer_question(question1,answer_text1)

Query has 17 tokens.

('paris', 14.0)

[ ] answer_question(question2,answer_text2)

Query has 33 tokens.

('the simulation of human intelligence', 15.0)

[ ] answer_question(question,answer_text)

Query has 130 tokens.

('divide the sentences of natural language into knowledge units and assign deep case to each word',
105.0)

```

Figure 10: Outputs for set of question-answer pairs

```

[ ] # Loop through rows and call the function
for index, row in df.iterrows():
    user_question = row['Question']
    context = row['Context']
    ground_truth = row['GroundTruth']
    # Call your answer_question function
    ans, f1 = answer_question3(user_question,context,ground_truth)
    # Print the result
    print(f"Question: {user_question}")
    print(f"Context: {context}")
    print(f"Answer: {ans}")
    print(f"F1 Score: {f1}")
    print(f"=" * 50)

Wedding, starring Jennifer Lopez and Josh Duhamel; romantic comedy Somebody I Used to Know, directed by Dave
Franco; comedic thriller The Consultant, starring Christoph Waltz; and Swarm, from co-creator and executive
producer Donald Glover.

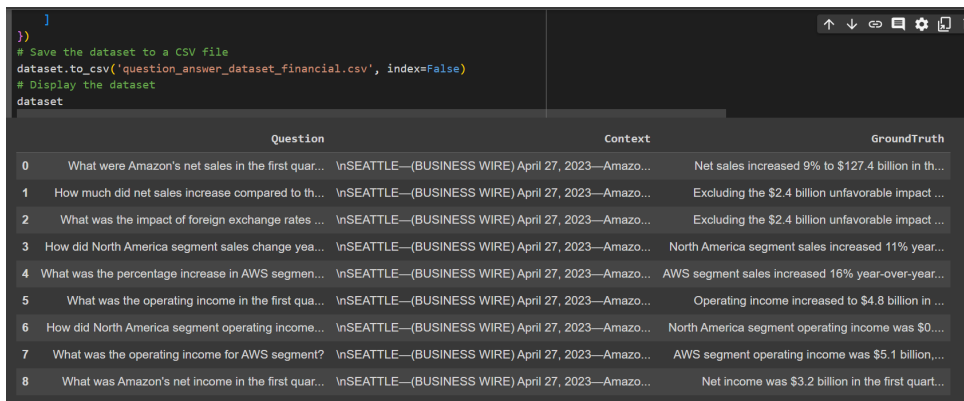
Answer: $ 127 . 4 billion
F1 Score: 0.0392156862745098
=====
The input has a total of 179 tokens.
Question: How did North America segment operating income compare to the first quarter of 2022?
Context:
SEATTLE--(BUSINESS WIRE) April 27, 2023--Amazon.com, Inc. (NASDAQ: AMZN) today announced financial results
for the first quarter of 2023.

```

Figure 11: Implementation of QA with curated dataset

8 Curated dataset creation

Datasets are created manually for each domain by running the dataset creation notebook.



```
})  
# Save the dataset to a CSV file  
dataset.to_csv('question_answer_dataset_financial.csv', index=False)  
# Display the dataset  
dataset
```

	Question	Context	GroundTruth
0	What were Amazon's net sales in the first quar...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Net sales increased 9% to \$127.4 billion in th...
1	How much did net sales increase compared to th...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Excluding the \$2.4 billion unfavorable impact ...
2	What was the impact of foreign exchange rates ...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Excluding the \$2.4 billion unfavorable impact ...
3	How did North America segment sales change yea...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	North America segment sales increased 11% year...
4	What was the percentage increase in AWS segmen...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	AWS segment sales increased 16% year-over-year...
5	What was the operating income in the first qua...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Operating income increased to \$4.8 billion in ...
6	How did North America segment operating income...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	North America segment operating income was \$0...
7	What was the operating income for AWS segment?	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	AWS segment operating income was \$5.1 billion,...
8	What was Amazon's net income in the first quar...	\nSEATTLE—(BUSINESS WIRE) April 27, 2023—Amazo...	Net income was \$3.2 billion in the first quart...

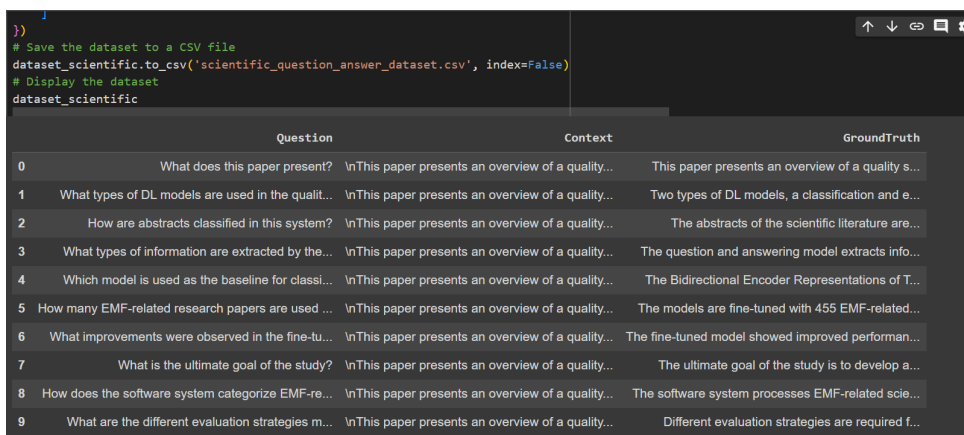
Figure 12: Financial dataset creation



```
# Save the dataset to a CSV file  
dataset_covid.to_csv('covid_question_answer_dataset.csv', index=False)  
# Display the dataset  
dataset_covid
```

	Question	Context	GroundTruth
0	What vulnerabilities has the COVID-19 pandemic...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic has revealed vulnerabili...
1	How has the COVID-19 pandemic affected the bio...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic has affected the biomed...
2	What challenges and lessons learned are discus...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The paper discusses the current status of rese...
3	How did the COVID-19 pandemic highlight the ne...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The COVID-19 pandemic highlighted the necessit...
4	What is the current status of research describ...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The paper describes the current status of rese...
5	How did public-private collaborations contribut...	\nIntroduction\nThe COVID-19 pandemic, a publi...	Public-private collaborations during the COVID...
6	What strategies did the National Institutes of...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The NIH adopted strategies such as investing i...
7	What priorities are outlined in The Science Ag...	\nIntroduction\nThe COVID-19 pandemic, a publi...	The Science Agenda for COVID-19 by the CDC out...
8	How did research funders, including non-profit...	\nIntroduction\nThe COVID-19 pandemic, a publi...	Research funders, including non-profit entitie...
9	What efforts did the Patient-Centered Outcomes...	\nIntroduction\nThe COVID-19 pandemic, a publi...	PCORI undertook efforts focused on adaptations...

Figure 13: Biological dataset creation



```
})  
# Save the dataset to a CSV file  
dataset_scientific.to_csv('scientific_question_answer_dataset.csv', index=False)  
# Display the dataset  
dataset_scientific
```

	Question	Context	GroundTruth
0	What does this paper present?	\nThis paper presents an overview of a quality...	This paper presents an overview of a quality s...
1	What types of DL models are used in the qualiti...	\nThis paper presents an overview of a quality...	Two types of DL models, a classification and e...
2	How are abstracts classified in this system?	\nThis paper presents an overview of a quality...	The abstracts of the scientific literature are...
3	What types of information are extracted by the...	\nThis paper presents an overview of a quality...	The question and answering model extracts info...
4	Which model is used as the baseline for classi...	\nThis paper presents an overview of a quality...	The Bidirectional Encoder Representations of T...
5	How many EMF-related research papers are used ...	\nThis paper presents an overview of a quality...	The models are fine-tuned with 455 EMF-related...
6	What improvements were observed in the fine-tu...	\nThis paper presents an overview of a quality...	The fine-tuned model showed improved performan...
7	What is the ultimate goal of the study?	\nThis paper presents an overview of a quality...	The ultimate goal of the study is to develop a...
8	How does the software system categorize EMF-re...	\nThis paper presents an overview of a quality...	The software system processes EMF-related scie...
9	What are the different evaluation strategies m...	\nThis paper presents an overview of a quality...	Different evaluation strategies are required f...

Figure 14: Scientific dataset creation

9 Model Fine tuning

This step implements fine-tuning of DistillBERT on the SQAuD dataset where hyper-parameters are set and the model is trained with trainer class. This tuned model is evaluated with a Financial dataset

```
[ ] from datasets import load_dataset

squad = load_dataset("squad", split="train[:5000]")

[ ] squad = squad.train_test_split(test_size=0.2)

[ ] squad["train"][0]

{'id': '57340aae4776f4190066178f',
 'title': 'Genocide',
 'context': 'Slobodan Milošević, as the former President of Serbia and of Yugoslavia, was the most senior political figure to stand trial at the ICTY. He died on 11 March 2006 during his trial where he was accused of genocide or complicity in genocide in territories within Bosnia and Herzegovina, no verdict was returned. In 1995, the ICTY issued a warrant for the arrest of Bosnian Serbs Radovan Karadžić and Ratko Mladić on several charges including genocide. On 21 July 2008, Karadžić was arrested in Belgrade, and he is currently in The Hague on trial accused of genocide among other crimes. Ratko Mladić was arrested on 26 May 2011 by Serbian special police in Lazarevo, Serbia. Karadžić was convicted of ten of the eleven charges laid against him and sentenced to 40 years in prison on March 24 2016.',
 'question': 'Which former president was by far the most senior politician to be accused of genocidal crimes by the ICTY?',
 'answers': {'text': ['Slobodan Milošević'], 'answer_start': [0]}}

[ ] squad["test"][0]

{'id': '56cfe179234ae51400d9c133',
 'title': 'Ratko Mladić',
 'context': 'Ratko Mladić was a Bosnian Serb military leader and politician who was the commander of the Army of Republika Srpska during the Bosnian War. He was indicted by the International Criminal Tribunal for the former Yugoslavia (ICTY) in 1997 for genocide, war crimes and crimes against humanity. He was found guilty of genocide and war crimes by the ICTY in 2001 and sentenced to 11 years in prison. He was released in 2002 and returned to Bosnia and Herzegovina. He was re-arrested in 2011 and sentenced to 11 years in prison by the ICTY in 2012. He was released in 2017 and returned to Bosnia and Herzegovina. He was re-arrested in 2018 and sentenced to 11 years in prison by the ICTY in 2019. He was released in 2020 and returned to Bosnia and Herzegovina. He was re-arrested in 2021 and sentenced to 11 years in prison by the ICTY in 2022. He was released in 2023 and returned to Bosnia and Herzegovina. He was re-arrested in 2024 and sentenced to 11 years in prison by the ICTY in 2025. He was released in 2026 and returned to Bosnia and Herzegovina. He was re-arrested in 2027 and sentenced to 11 years in prison by the ICTY in 2028. He was released in 2029 and returned to Bosnia and Herzegovina. He was re-arrested in 2030 and sentenced to 11 years in prison by the ICTY in 2031. He was released in 2032 and returned to Bosnia and Herzegovina. He was re-arrested in 2033 and sentenced to 11 years in prison by the ICTY in 2034. He was released in 2035 and returned to Bosnia and Herzegovina. He was re-arrested in 2036 and sentenced to 11 years in prison by the ICTY in 2037. He was released in 2038 and returned to Bosnia and Herzegovina. He was re-arrested in 2039 and sentenced to 11 years in prison by the ICTY in 2040. He was released in 2041 and returned to Bosnia and Herzegovina. He was re-arrested in 2042 and sentenced to 11 years in prison by the ICTY in 2043. He was released in 2044 and returned to Bosnia and Herzegovina. He was re-arrested in 2045 and sentenced to 11 years in prison by the ICTY in 2046. He was released in 2047 and returned to Bosnia and Herzegovina. He was re-arrested in 2048 and sentenced to 11 years in prison by the ICTY in 2049. He was released in 2050 and returned to Bosnia and Herzegovina. He was re-arrested in 2051 and sentenced to 11 years in prison by the ICTY in 2052. He was released in 2053 and returned to Bosnia and Herzegovina. He was re-arrested in 2054 and sentenced to 11 years in prison by the ICTY in 2055. He was released in 2056 and returned to Bosnia and Herzegovina. He was re-arrested in 2057 and sentenced to 11 years in prison by the ICTY in 2058. He was released in 2059 and returned to Bosnia and Herzegovina. He was re-arrested in 2060 and sentenced to 11 years in prison by the ICTY in 2061. He was released in 2062 and returned to Bosnia and Herzegovina. He was re-arrested in 2063 and sentenced to 11 years in prison by the ICTY in 2064. He was released in 2065 and returned to Bosnia and Herzegovina. He was re-arrested in 2066 and sentenced to 11 years in prison by the ICTY in 2067. He was released in 2068 and returned to Bosnia and Herzegovina. He was re-arrested in 2069 and sentenced to 11 years in prison by the ICTY in 2070. He was released in 2071 and returned to Bosnia and Herzegovina. He was re-arrested in 2072 and sentenced to 11 years in prison by the ICTY in 2073. He was released in 2074 and returned to Bosnia and Herzegovina. He was re-arrested in 2075 and sentenced to 11 years in prison by the ICTY in 2076. He was released in 2077 and returned to Bosnia and Herzegovina. He was re-arrested in 2078 and sentenced to 11 years in prison by the ICTY in 2079. He was released in 2080 and returned to Bosnia and Herzegovina. He was re-arrested in 2081 and sentenced to 11 years in prison by the ICTY in 2082. He was released in 2083 and returned to Bosnia and Herzegovina. He was re-arrested in 2084 and sentenced to 11 years in prison by the ICTY in 2085. He was released in 2086 and returned to Bosnia and Herzegovina. He was re-arrested in 2087 and sentenced to 11 years in prison by the ICTY in 2088. He was released in 2089 and returned to Bosnia and Herzegovina. He was re-arrested in 2090 and sentenced to 11 years in prison by the ICTY in 2091. He was released in 2092 and returned to Bosnia and Herzegovina. He was re-arrested in 2093 and sentenced to 11 years in prison by the ICTY in 2094. He was released in 2095 and returned to Bosnia and Herzegovina. He was re-arrested in 2096 and sentenced to 11 years in prison by the ICTY in 2097. He was released in 2098 and returned to Bosnia and Herzegovina. He was re-arrested in 2099 and sentenced to 11 years in prison by the ICTY in 2100. He was released in 2101 and returned to Bosnia and Herzegovina. He was re-arrested in 2102 and sentenced to 11 years in prison by the ICTY in 2103. He was released in 2104 and returned to Bosnia and Herzegovina. He was re-arrested in 2105 and sentenced to 11 years in prison by the ICTY in 2106. He was released in 2107 and returned to Bosnia and Herzegovina. He was re-arrested in 2108 and sentenced to 11 years in prison by the ICTY in 2109. He was released in 2110 and returned to Bosnia and Herzegovina. He was re-arrested in 2111 and sentenced to 11 years in prison by the ICTY in 2112. He was released in 2113 and returned to Bosnia and Herzegovina. He was re-arrested in 2114 and sentenced to 11 years in prison by the ICTY in 2115. He was released in 2116 and returned to Bosnia and Herzegovina. He was re-arrested in 2117 and sentenced to 11 years in prison by the ICTY in 2118. He was released in 2119 and returned to Bosnia and Herzegovina. He was re-arrested in 2120 and sentenced to 11 years in prison by the ICTY in 2121. He was released in 2122 and returned to Bosnia and Herzegovina. He was re-arrested in 2123 and sentenced to 11 years in prison by the ICTY in 2124. He was released in 2125 and returned to Bosnia and Herzegovina. He was re-arrested in 2126 and sentenced to 11 years in prison by the ICTY in 2127. He was released in 2128 and returned to Bosnia and Herzegovina. He was re-arrested in 2129 and sentenced to 11 years in prison by the ICTY in 2130. He was released in 2131 and returned to Bosnia and Herzegovina. He was re-arrested in 2132 and sentenced to 11 years in prison by the ICTY in 2133. He was released in 2134 and returned to Bosnia and Herzegovina. He was re-arrested in 2135 and sentenced to 11 years in prison by the ICTY in 2136. He was released in 2137 and returned to Bosnia and Herzegovina. He was re-arrested in 2138 and sentenced to 11 years in prison by the ICTY in 2139. He was released in 2140 and returned to Bosnia and Herzegovina. He was re-arrested in 2141 and sentenced to 11 years in prison by the ICTY in 2142. He was released in 2143 and returned to Bosnia and Herzegovina. He was re-arrested in 2144 and sentenced to 11 years in prison by the ICTY in 2145. He was released in 2146 and returned to Bosnia and Herzegovina. He was re-arrested in 2147 and sentenced to 11 years in prison by the ICTY in 2148. He was released in 2149 and returned to Bosnia and Herzegovina. He was re-arrested in 2150 and sentenced to 11 years in prison by the ICTY in 2151. He was released in 2152 and returned to Bosnia and Herzegovina. He was re-arrested in 2153 and sentenced to 11 years in prison by the ICTY in 2154. He was released in 2155 and returned to Bosnia and Herzegovina. He was re-arrested in 2156 and sentenced to 11 years in prison by the ICTY in 2157. He was released in 2158 and returned to Bosnia and Herzegovina. He was re-arrested in 2159 and sentenced to 11 years in prison by the ICTY in 2160. He was released in 2161 and returned to Bosnia and Herzegovina. He was re-arrested in 2162 and sentenced to 11 years in prison by the ICTY in 2163. He was released in 2164 and returned to Bosnia and Herzegovina. He was re-arrested in 2165 and sentenced to 11 years in prison by the ICTY in 2166. He was released in 2167 and returned to Bosnia and Herzegovina. He was re-arrested in 2168 and sentenced to 11 years in prison by the ICTY in 2169. He was released in 2170 and returned to Bosnia and Herzegovina. He was re-arrested in 2171 and sentenced to 11 years in prison by the ICTY in 2172. He was released in 2173 and returned to Bosnia and Herzegovina. He was re-arrested in 2174 and sentenced to 11 years in prison by the ICTY in 2175. He was released in 2176 and returned to Bosnia and Herzegovina. He was re-arrested in 2177 and sentenced to 11 years in prison by the ICTY in 2178. He was released in 2179 and returned to Bosnia and Herzegovina. He was re-arrested in 2180 and sentenced to 11 years in prison by the ICTY in 2181. He was released in 2182 and returned to Bosnia and Herzegovina. He was re-arrested in 2183 and sentenced to 11 years in prison by the ICTY in 2184. He was released in 2185 and returned to Bosnia and Herzegovina. He was re-arrested in 2186 and sentenced to 11 years in prison by the ICTY in 2187. He was released in 2188 and returned to Bosnia and Herzegovina. He was re-arrested in 2189 and sentenced to 11 years in prison by the ICTY in 2190. He was released in 2191 and returned to Bosnia and Herzegovina. He was re-arrested in 2192 and sentenced to 11 years in prison by the ICTY in 2193. He was released in 2194 and returned to Bosnia and Herzegovina. He was re-arrested in 2195 and sentenced to 11 years in prison by the ICTY in 2196. He was released in 2197 and returned to Bosnia and Herzegovina. He was re-arrested in 2198 and sentenced to 11 years in prison by the ICTY in 2199. He was released in 2200 and returned to Bosnia and Herzegovina. He was re-arrested in 2201 and sentenced to 11 years in prison by the ICTY in 2202. He was released in 2203 and returned to Bosnia and Herzegovina. He was re-arrested in 2204 and sentenced to 11 years in prison by the ICTY in 2205. He was released in 2206 and returned to Bosnia and Herzegovina. He was re-arrested in 2207 and sentenced to 11 years in prison by the ICTY in 2208. He was released in 2209 and returned to Bosnia and Herzegovina. He was re-arrested in 2210 and sentenced to 11 years in prison by the ICTY in 2211. He was released in 2212 and returned to Bosnia and Herzegovina. He was re-arrested in 2213 and sentenced to 11 years in prison by the ICTY in 2214. He was released in 2215 and returned to Bosnia and Herzegovina. He was re-arrested in 2216 and sentenced to 11 years in prison by the ICTY in 2217. He was released in 2218 and returned to Bosnia and Herzegovina. He was re-arrested in 2219 and sentenced to 11 years in prison by the ICTY in 2220. He was released in 2221 and returned to Bosnia and Herzegovina. He was re-arrested in 2222 and sentenced to 11 years in prison by the ICTY in 2223. He was released in 2224 and returned to Bosnia and Herzegovina. He was re-arrested in 2225 and sentenced to 11 years in prison by the ICTY in 2226. He was released in 2227 and returned to Bosnia and Herzegovina. He was re-arrested in 2228 and sentenced to 11
```

Figure 15: fine tuning with SQAud dataset

```
[ ] from transformers import AutoModelForQuestionAnswering, TrainingArguments, Trainer

model = AutoModelForQuestionAnswering.from_pretrained("distilbert-base-uncased")

Some weights of DistilBertForQuestionAnswering were not initialized from the model checkpoint at distilbert-base-uncased and are newly initialized: ['
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

[ ] training_args = TrainingArguments(
    output_dir="my_awesome_qa_model",
    evaluation_strategy="epoch",
    learning_rate=1e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    push_to_hub=False,
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_squad["train"],
```

Figure 16: Hyperparameter tuning

```
[ ] trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_squad["train"],
    eval_dataset=tokenized_squad["test"],
    tokenizer=tokenizer,
    data_collator=data_collator,
)
trainer.train()
```

[1500/1500 10.15, Epoch 3/3]

Epoch	Training Loss	Validation Loss
1	3.728600	2.778563
2	2.300600	2.126101
3	1.840000	2.011740

```
TrainOutput(global_step=1500, training_loss=2.623054972330729, metrics={'train_runtime': 620.3921, 'train_samples_per_second': 19.343,
'train_steps_per_second': 2.418, 'total_flos': 117587790288000.0, 'train_loss': 2.623054972330729, 'epoch': 3.0})
```

+ Code + Text

Figure 17: Fine-tuning of DistilBERT calling trainer class

```
[ ] 2 2.300600 2.126101
    3 1.840000 2.011740

TrainOutput(global_step=1500, training_loss=2.623054972330729, metrics={'train_runtime': 620.3921, 'train_samples_per_second': 19.343,
'train_steps_per_second': 2.418, 'total_flos': 117587790028000.0, 'train_loss': 2.623054972330729, 'epoch': 3.0})

[ ] from transformers import TrainingArguments

[ ] question = "what are different search engines?"
context = "BLOOM has 176 billion parameters and can generate text in 46 languages natural languages and 13 programming languages."

[ ] pwd

'/content/gdrive/MyDrive/Colab Notebooks'

[ ] from transformers import pipeline
question_answerer = pipeline("question-answering", model="/content/gdrive/MyDrive/Colab Notebooks/my_awesome_qa_model/checkpoint-500")
question_answerer(question=question, context=context)

{'score': 0.25022581219673157, 'start': 10, 'end': 21, 'answer': '176 billion'}
```

Figure 18: Evaluating financial dataset on fine-tuned model

10 QA system implementation for BioMedical dataset

The BioMedical dataset was implemented using pre-trained bioBERT

```
# tokenize and preprocess the predicted answer
predicted_answer_processed = re.sub(r'##', '', predicted_answer)
# Ensure that ground truth and predicted answer have the same length
max_length = max(len(ground_truth_processed), len(predicted_answer_processed))
ground_truth_processed = ground_truth_processed.ljust(max_length)
predicted_answer_processed = predicted_answer_processed.ljust(max_length)
# Calculate f1 score for ground truth and predicted answer
f1_ground_truth_predicted = f1_score(list(ground_truth_processed), list(predicted_answer_processed), average='weighted', zero_division=1)
return predicted_answer, f1_ground_truth_predicted

[ ] import pandas as pd
# Loop through rows and call the function
for index, row in df.iterrows():
    user_question = row['Question']
    context = row['Context']
    ground_truth = row['GroundTruth']
    # Call answer_question function
    ans, f1_value = answer_question4(user_question, context, ground_truth)
    # Print the result
    print(f"Question: {user_question}")
    print(f"Context: {context}")
    print(f"Answer: {ans}")
    print(f"F1 Score: {f1_value}")
    print("=" * 50)
```

Figure 19: QA system with BioClinical BERT for biomedical dataset

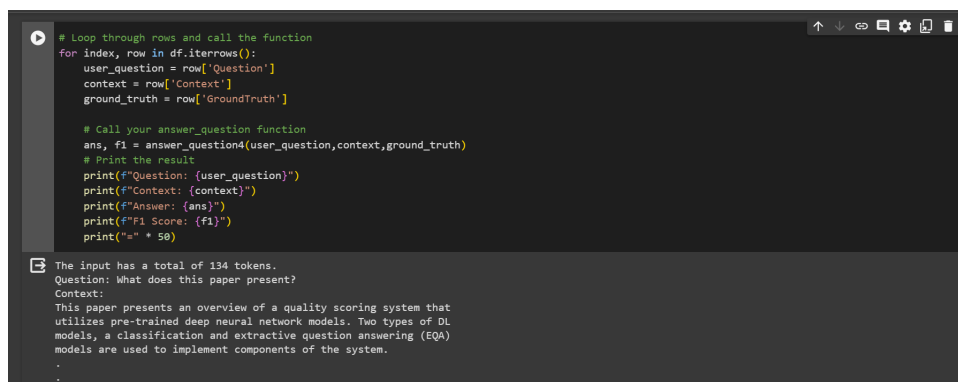
```
As the nation's largest public funder of biomedical research, the National Institutes of Health (NIH) leveraged existing infrastructure to establish
1. Invest in NIH and NIH-funded researchers to increase fundamental and foundational knowledge of SARS-CoV-2 and COVID-19.
2. Speed innovation in COVID-19 testing technologies through NIH's recently launched Rapid Acceleration of Diagnostics (RADx) initiative, which aims
3. Participate in public-private partnerships, such as NIH's Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) partnership, and f
4. Support studies on preventative treatments and behavioral and community prevention practices to identify and implement effective approaches for p
5. Ensure that diagnosis, treatment, and prevention options are accessible and available for underserved and vulnerable populations that have been a
Similarly, the Centers for Disease Control and Prevention (CDC), at the forefront of the public health response to the COVID-19 pandemic, establishe
1. COVID-19 disease detection, burden, and impact, especially as it relates to understanding disproportionate impacts on people at increased risk fo
2. transmission of SARS-CoV-2;
3. natural history of SARS-CoV-2 infection;
4. protection in health care and non-health care work settings;
5. prevention, mitigation, and intervention strategies; and
6. social, behavioral, and communication science.
Other research funders, including non-profit entities, created research agendas focused on their unique missions and opportunities to contribute to
Medical specialty societies, such as the Infectious Disease Society of America, identified priorities for COVID-19 research more broadly and funded

Answer: yet also serves as a remarkable learning opportunity for transformational changes . effects of the covid - 19 pandemic touch every aspect of
F1 Score: 0.15767441524915934
=====
The input has a total of 512 tokens.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So
Question: What challenges and lessons learned are discussed in the paper?
Context:
Introduction
The COVID-19 pandemic, a public health emergency of unprecedented scale and consequences, has revealed vulnerabilities in our health care system and
Despite the rapid innovation occurring during the COVID-19 pandemic, longstanding problems remain. The disproportionate burden of COVID-19 cases and
This paper describes the current status of research and the challenges, lessons learned, and the potential, if the challenges are overcome, for a lo
```

Figure 20: Predicted answers for Bioclinical BERT

11 QA system implementation for Scientific dataset

The Scientific dataset was implemented using pre-trained sciBERT

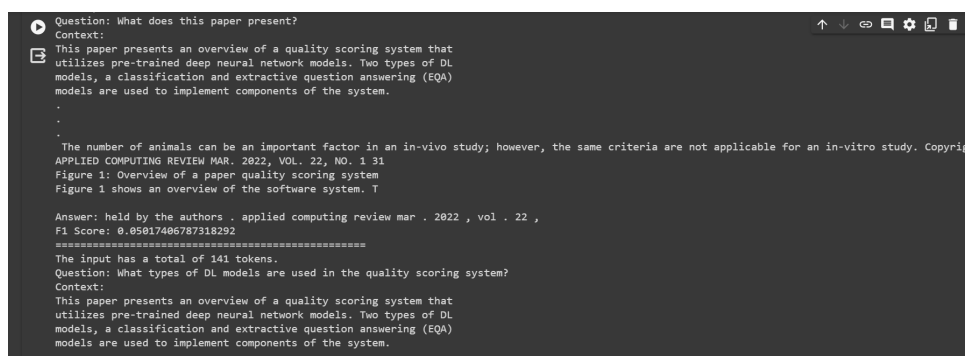


```
# Loop through rows and call the function
for index, row in df.iterrows():
    user_question = row['Question']
    context = row['Context']
    ground_truth = row['GroundTruth']

    # Call your answer_question function
    ans, f1 = answer_question4(user_question, context, ground_truth)
    # Print the result
    print(f"Question: {user_question}")
    print(f"Context: {context}")
    print(f"Answer: {ans}")
    print(f"F1 Score: {f1}")
    print(f"=" * 50)
```

The input has a total of 134 tokens.
Question: What does this paper present?
Context:
This paper presents an overview of a quality scoring system that utilizes pre-trained deep neural network models. Two types of DL models, a classification and extractive question answering (EQA) models are used to implement components of the system.
.
.

Figure 21: QA system with Scientific BERT for Scientific dataset



```
Question: What does this paper present?
Context:
This paper presents an overview of a quality scoring system that utilizes pre-trained deep neural network models. Two types of DL models, a classification and extractive question answering (EQA) models are used to implement components of the system.
.
.
The number of animals can be an important factor in an in-vivo study; however, the same criteria are not applicable for an in-vitro study. Copyright
APPLIED COMPUTING REVIEW MAR. 2022, VOL. 22, NO. 1 31
Figure 1: Overview of a paper quality scoring system
Figure 1 shows an overview of the software system. T

Answer: held by the authors . applied computing review mar . 2022 , vol . 22 ,
F1 Score: 0.05017406787318292
=====
The input has a total of 141 tokens.
Question: What types of DL models are used in the quality scoring system?
Context:
This paper presents an overview of a quality scoring system that utilizes pre-trained deep neural network models. Two types of DL models, a classification and extractive question answering (EQA) models are used to implement components of the system.
.
.
```

Figure 22: Predicted answers for Scientific BERT

12 Cross domain Evaluation

Biomedical and scientific datasets were evaluated on the BERT large model to check the QA system's generalizability.

```
[ ] ground_truth_processed = ground_truth_processed.ljust(max_length)
    predicted_answer_processed = predicted_answer_processed.ljust(max_length)

    # Calculate F1 score for ground truth and predicted answer
    f1_ground_truth_predicted = f1_score(list(ground_truth_processed), list(predicted_answer_processed), average='weighted', zero_division=1)
    return predicted_answer, f1_ground_truth_predicted

[ ] # Loop through rows and call the function
for index, row in df_BioMedical.iterrows():
    user_question = row['Question']
    context = row['Context']
    ground_truth = row['GroundTruth']
    # Call answer_question function
    ans, f1_value = answer_question4(user_question, context, ground_truth)
    # Print the result
    print(f"Question: {user_question}")
    print(f"Context: {context}")
    print(f"Answer: {ans}")
    print(f"F1 Score: {f1_value}")
    print("=" * 50)

Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So
The input has a total of 512 tokens.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So
```

Figure 23: Cross-domain analysis of BiobERT on BERT large

```
[ ] print(
5. Ensure that diagnosis, treatment, and prevention options are accessible and available for underserved and vulnerable populations that have been a
Similarly, the Centers for Disease Control and Prevention (CDC), at the forefront of the public health response to the COVID-19 pandemic, establishe
1. COVID-19 disease detection, burden, and impact, especially as it relates to understanding disproportionate impacts on people at increased risk fo
2. transmission of SARS-CoV-2;
3. natural history of SARS-CoV-2 infection;
4. protection in health care and non-health care work settings;
5. prevention, mitigation, and intervention strategies; and
6. social, behavioral, and communication science.
Other research funders, including non-profit entities, created research agendas focused on their unique missions and opportunities to contribute to
Medical specialty societies, such as the Infectious Disease Society of America, identified priorities for COVID-19 research more broadly and funded

Answer: vulnerabilities
F1 Score: 0.02802335279394993
=====
The input has a total of 512 tokens.
Be aware, overflowing tokens are not returned for the setting you have chosen, i.e. sequence pairs with the 'longest_first' truncation strategy. So
Question: How has the COVID-19 pandemic affected the biomedical and health research enterprises?
Context:
Introduction
The COVID-19 pandemic, a public health emergency of unprecedented scale and consequences, has revealed vulnerabilities in our health care system and
Despite the rapid innovation occurring during the COVID-19 pandemic, longstanding problems remain. The disproportionate burden of COVID-19 cases and
This paper describes the current status of research and the challenges, lessons learned, and the potential, if the challenges are overcome, for a lo
These lessons learned can be applied to help advance the rapid translation of research into practice (from basic science to clinical and popula
Overview of the Research Landscape
https://www.fda.gov/oc/identify-a-research-landscape-known-as-sars-cov-2-and-its-disease-manifestation-covid-19-institutions-researchers-public-repo
```

Figure 24: predicted answers for Cross-domain analysis of BiobERT on BERT large

```
+ Code + Text Connect ^
[ ] max_length = max(len(ground_truth_processed), len(predicted_answer_processed))
    ground_truth_processed = ground_truth_processed.ljust(max_length)
    predicted_answer_processed = predicted_answer_processed.ljust(max_length)

    # Calculate F1 score for ground truth and predicted answer
    f1_ground_truth_predicted = f1_score(list(ground_truth_processed), list(predicted_answer_processed), average='weighted', zero_division=1)
    return predicted_answer, f1_ground_truth_predicted

for index, row in df_Scientific.iterrows():
    user_question = row['Question']
    context = row['Context']
    ground_truth = row['GroundTruth']
    # Call your answer_question function
    ans, f1 = answer_question4(user_question, context, ground_truth)

    # Print the result
    print(f"Question: {user_question}")
    print(f"Context: {context}")
    print(f"Answer: {ans}")
    print(f"F1 Score: {f1}")
    print("=" * 50)

The input has a total of 512 tokens.
```

Figure 25: Cross-domain analysis of sciBERT on BERT large

```
models, a classification and extractive question answering (EQA)
models are used to implement components of the system.
.
.
.
The number of animals can be an important factor in an in-vivo study; however, the same criteria are not applicable for an in-vitro study. Copyright
APPLIED COMPUTING REVIEW MAR. 2022, VOL. 22, NO. 1 31
Figure 1: Overview of a paper quality scoring system
Figure 1 shows an overview of the software system. T

Answer: an overview of a quality scoring system
F1 Score: 0.026548872566371678
=====
The input has a total of 141 tokens.
Question: What types of DL models are used in the quality scoring system?
Context:
This paper presents an overview of a quality scoring system that
utilizes pre-trained deep neural network models. Two types of DL
models, a classification and extractive question answering (EQA)
models are used to implement components of the system.
.
.
.
The number of animals can be an important factor in an in-vivo study; however, the same criteria are not applicable for an in-vitro study. Copyright
APPLIED COMPUTING REVIEW MAR. 2022, VOL. 22, NO. 1 31
Figure 1: Overview of a paper quality scoring system
Figure 1 shows an overview of the software system. T
```

Figure 26: predicted answers for Cross-domain analysis of sciBERT on BERT large

References

- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T. and McDermott, M. (2019). Publicly available clinical bert embeddings, *arXiv preprint arXiv:1904.03323*.
- Beltagy, I., Lo, K. and Cohan, A. (2019). Scibert: A pretrained language model for scientific text, *arXiv preprint arXiv:1903.10676*.
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*.
- Pearce, K., Zhan, T., Komanduri, A. and Zhan, J. (2021). A comparative study of transformer-based language models on extractive question answering, *arXiv preprint arXiv:2110.03142*.