Table of Contents

Introduction

1.1 Problem Background

This research is motivated by the challenges associated with manually reviewing and interpreting documents, especially resumes. Manual resume reviews require reading through lengthy, dense passages of text, often spanning several pages, which can be mentally exhausting and time-consuming. Furthermore, inconsistent language and formatting make it more difficult to precisely extract relevant details. Moreover, human error is common and can result in the oversight of important information or a misinterpretation of a candidate's qualifications. As a result, the manual resume evaluation process reduces the effectiveness of candidate selection procedures and may lead to unsuitable hiring choices or missed opportunities (Muludi et al., 2024).

1.2 Objectives/Aims

1. **To measure the similarity of responses generated by different embedding and question-answering models compared to the provided PDF content**
    - This objective aims to quantify how closely the responses generated by the models match the content of the resume. Human evaluation and similarity metrics such as Levenshtein Distance will be employed to quantify the resemblance between the generated responses and the original content.

2. **To evaluate the response time of different embedding and question-answering models based on different queries' length**
    - This objective focuses on measuring the time taken by different embedding and question-answering models to generate responses to user queries. Response time can be measured from the moment a user sends a query to the moment the application displays a response. This metric helps assess the efficiency and responsiveness of the models in real-time interactions.

1.3 Motivation

- **Enhanced recruitment workflow efficiency**: By providing recruiters with a faster and more efficient way to analyze resumes and match candidates to job descriptions, this project could streamline recruitment processes for businesses, leading to cost and time savings and improved quality of hiring outcomes. (Alexandria, 2023)

- **Recruitment industry integration**: Companies are integrating new generative AI large language models (LLMs) into their talent intelligence platforms to revolutionize the recruitment industry. This project, which utilizes similar technology, could be commercialized as a complementary tool for recruiters and job seekers, offering enhanced resume analysis and interaction capabilities. (Alexandria, 2023)

1.4 Timeline/Milestone

					PHASE ONE																									PHASE TWO																																		
TASK NO	TASK TITLE	START DATE	END DATE	DAYS USED	WEEK 2					WEEK 3					WEEK 4					WEEK 5					WEEK 6					WEEK 7					WEEK 8					WEEK 9					WEEK 10					WEEK 11					WEEK 12					WEEK 13				
					M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F	M	T	W	T	F
1	Brainstorm potential project topics and discussion on their feasibility	3/1/24	3/8/24	8																																																												
2	Explore relevant academic papers, research articles, and industry projects on the chosen topics	3/11/24	3/22/24	12																																																												
3	Discussion and brainstorming on gathered information	3/25/24	3/27/24	3																																																												
4	Define the specific objectives and scope of the project	3/28/24	3/29/24	2																																																												
5	Design the user interface for the interactive chat application	4/1/24	4/2/24	2																																																												
6	Implement functionality for uploading and parsing PDF resumes	4/3/24	4/5/24	3																																																												
7	Develop the backend infrastructure, research and testing	4/8/24	5/7/24	30																																																												
11	Finalize project and prepare documentation	5/8/24	5/10/24	3																																																												
12	Submission of project	5/13/24	5/13/24	1																																																												

Research Background

2.1 Background of the applications

2.1.1 Natural Language Processing (NLP)

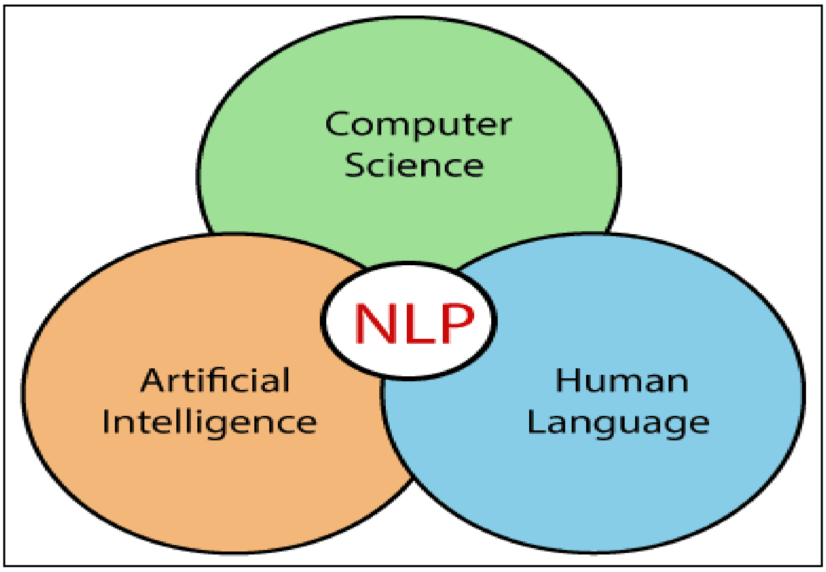


Figure 2.1.1: Fields related to NLP

- Natural Language Processing (NLP) involves utilizing computers to process and analyze human language, acting as a bridge between computer science, artificial intelligence, and human linguistics.

- Two main components of NLP are natural language understanding (NLU), and natural language generation (NLG). NLU focuses on breaking down text data and interpreting its meaning while NLG focuses on producing structured text data that humans can easily understand.

- NLP forms the core of this project, enabling understanding and generation of human language. In this project, NLP techniques are crucial for interpreting user queries, extracting meaningful information from resumes, and generating coherent responses.

- NLP techniques like text cleaning, tokenization, and chunking facilitate the understanding of user intents and the context of document content, enabling effective communication between users and the chat application.

2.1.2 Portable Document Format (PDF)

Figure 2.1.2: Adobe PDF Logo

- PDF is a file format created by Adobe, this format provides a simple, dependable means of sharing and presenting documents to others regardless of the software, hardware, or operating system used by the recipients.

- PDFs are widely used digital document formats known for their ability to maintain the original layout and formatting across different platforms and devices. PDFs can be viewed and interacted using different software applications, including Adobe Acrobat Reader, web browsers, and mobile apps, making them accessible across different platforms and devices.

- PDFs can contain a diverse range of content, including text, images, vector graphics, interactive elements, hyperlinks, annotations, and multimedia elements like audio and video.

- In this project, PDFs serve as the input source for the interactive chat application, allowing users to upload resumes in PDF for discussion and analysis within the application interface.

2.1.3 Chatbot Development

- Chatbot development involves creating conversational agents that can interact with users in a natural language format, simulating human-like conversations.

- Chatbots are broadly categorized into rule-based chatbots, which rely on predefined scripts or rules for responses, and artificially intelligent chatbots, which utilize NLP to understand and respond to nuanced human conversation, learning from interactions to provide improved responses over time.

- In this project, chatbot development is crucial as it enables the interactive chat application to engage with users in a conversational manner regarding the content of resumes.

- It allows users to ask questions, seek clarifications, or request insights related to the resume's content, thereby facilitating a more engaging and informative experience within the application.

2.2 Analysis of selected tool/algorithm with any other relevant tools/algorithms

2.2.1 Portable Document Format (PDF) Extractor

Tools comparison	**pdfminer.six (selected)**	**PyPDF**	**PyMuPDF**
Type of license and open source license	MIT License (MIT/X) (License)	BSD License (License)	open-source AGPL and commercial license agreements from Artifex. (License)
Year founded	2014 (First traceable official release of pdfminer.six)	2005 (First traceable official release of PyPDF)	2016 (First traceable official release of PyMuPDF)
Founding company	Yusuke Shinyama (Founder)	Mathieu Fenniak (Founder)	Jorj X. McKie. (Founder)
License Pricing	Free	Free	Free (open source), Starting from: $549.00 (One time payment)
Supported features	Text, elements, images extraction from PDF (Official documentation)	Text, images, attachments from PDF, Post-Processing of Text Extraction, Encryption and Decryption of PDF, etc. (Official documentation)	Text, images, attachments from different document types, OCR, etc (Official documentation)
Common applications	Data extraction and mining for document processing and analysis	Data extraction and mining for document processing and analysis, Document Manipulation	Data extraction and mining for document processing and analysis, Document Manipulation
Advantages	Open Source, Python Compatibility, Comprehensive Features	Open Source, Python Compatibility, Comprehensive Features	Rich Feature Set, Python Compatibility, High Performance Rendering
Limitations	Limited Support for Interactive Content	Limited Support for Interactive Content	License Restrictions

2.2.2 Word Embedding Models

Tools comparison	All MPNet Base V2	GTE Base EN V1.5	MiniLM-L6-v2
Type of license and open source license	Apache 2.0 License. (License)	Apache 2.0 License. (License)	Apache 2.0 License. (License)
Year founded	2021 (First traceable official release of MPNet)	2023 (First traceable official release of GTE)	2021 (First traceable official release of MPNet)
Founding company	Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu (Founder)	Alibaba Group (Founder)	Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, Ming Zhou (Founder)
License Pricing	Free	Free	Free
Supported features	Encode sentences and short paragraphs (Official documentation)	Text embedding (Official documentation)	used as a sentence and short paragraph encoder. Given an input text, it outputs a vector which captures the semantic information (Official documentation)
Common applications	semantic search, text clustering, recommendation system, multi-modal retrieval (Reference)	text embeddings, including information retrieval, semantic textual similarity, text reranking (Reference)	Information retrieval, clustering or sentence similarity tasks (Reference)
Advantages	Open Source, Integration with Libraries	Open source, high scalability	Open source, efficient mapping
Limitations	High computational cost	exclusively caters to English texts, and any lengthy texts will be truncated (Reference)	The sequence length was limited to 128 tokens. This means that while it technically can process longer sequences up to 256 tokens, the quality of the embeddings may not be as reliable for tokens beyond the 128th position due to the training constraints (Reference)

2.2.3 Question Answering (QA) Models

Tools comparison	RoBERTa Base SQuAD 2	mDeBERTa V3 Base SQuAD 2	DistilBERT Base Cased Distilled SQuAD	BERT Large Uncased Whole Word Masking SQuAD 2	Longformer Base 4096 Finetuned SQuAD V2
Type of license and open source license	cc-by-4.0 (License)	mit (License)	Apache 2.0 License. (License)	cc-by-4.0 (License)	Apache 2.0 License. (License)
Year founded	2019 (First traceable official release of RoBERTa)	2021 (First traceable official release of DeBERTa)	2019 (First traceable official release of DistilBERT)	2019 (Uncased Whole Word Masking SQuAD 2)	2020 (First traceable official release of Longformer)
Founding company	Deepset (Founder)	Deepset (Founder)	Hugging Face (Founder)	Deepset (Founder)	the Allen Institute for Artificial Intelligence (AI2) (Founder)
License Pricing	Free	Free	Free	Free	Free
Supported features	Text classification, Question answering	Named entity recognition,Question answering	Language understanding, Question answering	Question answering, Text generation	Text classification, Long-form question answering
Common applications	Question answering system (Reference)	Question answering system (Reference)	Question answering system (Reference)	Question answering system (Reference)	Question answering system (Reference)
Advantages	Open source, Excels in understanding contextual information in text data through bidirectional encoding.	Open source, Understand and process text data efficiently while being computationally lightweight	Open source, maintaining performance while reducing computational requirements.	Open source, Suitable for complex NLP tasks requiring a deep understanding of context and semantics	Open source, Suitable for tasks involving lengthy passages or documents
Limitations	Requires significant computational resources and time for pre-training due to its size and complexity.	While optimized for edge devices, it may not achieve the same level of performance as larger models due to its smaller size and parameter count.	Limited in capturing fine-grained linguistic nuances compared to larger models, which may affect performance on certain tasks.	Demands significant computational resources and time for both pre-training and inference due to its size, limiting its scalability.	The increased sequence length comes with computational overhead, potentially limiting its scalability for extremely long documents.

2.3 Justification on selected tools/algorithms

2.3.1 Portable Document Format (PDF) Extractor

For Portable Document Format (PDF) Extractor, **pdfminer.six** is chosen to perform the extraction.Pdfminer.six's versatility and strong capabilities make it an excellent option for extracting text from PDFs in a range of applications. Because of its proficiency at effectively parsing and extracting text from PDF documents, any PDF reader program can reliably retrieve the content required to respond to user inquiries. To further enhance the depth of information that systems can offer users, pdfminer.six provides a plethora of features for navigating through PDF files, such as accessing text, images, and metadata. Its ability to work seamlessly in a variety of Python environments guarantees a smooth integration into projects, allowing for smooth functionality without a large overhead. All things considered, pdfminer.six's dependability and adaptability make it a perfect tool for extracting PDF content in a variety of applications, allowing systems to efficiently respond to user queries with pertinent information.

2.3.2 Word Embedding Models

For Word Embedding Models, All MPNet Base V2 and GTE Base EN V1.5 are chosen to perform the word embedding.

All MPNet Base V2:
The MPNet Base V2 model is a suitable choice for word embeddings in a project that involves creating a PDF reader with question-answering capabilities. This model is a pre-trained masked language model that leverages the power of transformer architectures to capture rich contextual information from text. One of its key strengths lies in its ability to handle long-range dependencies within text, allowing it to effectively capture semantic relationships across sentences and paragraphs. This feature is particularly valuable when dealing with PDF documents, which often contain complex and intricate text structures. Additionally, MPNet Base V2 has been pre-trained on a vast corpus of data, providing it with a robust understanding of natural language across various domains. This broad knowledge base can be advantageous when addressing questions that span a wide range of topics, as is often the case with PDF documents.

GTE Base EN V1.5:
The GTE Base EN V1.5 model, developed by Google, is another suitable option for word embeddings in a PDF reader project with question-answering capabilities. This model is based on the Transformer architecture and has been pre-trained on a massive corpus of English text, allowing it to develop a deep understanding of the language. One of its key strengths lies in its ability to capture nuanced semantic relationships and handle polysemy, which is particularly valuable when dealing with the diverse vocabulary and context found in PDF documents. Additionally, GTE Base EN V1.5 has been specifically optimized for efficiency, making it well-suited for deployment in resource-constrained environments or applications that require real-time performance. This efficiency can be beneficial for a PDF reader system that needs to process and analyze documents promptly, ensuring a smooth user experience.

2.3.3 Question Answering (QA) Models

For Question Answering (QA) Models, all models are chosen to perform question answering.

RoBERTa Base SQuAD 2:
The RoBERTa Base SQuAD 2 model is a robust choice for question answering in a PDF reader project. RoBERTa (Robustly Optimized BERT Pre-training Approach) is a variant of the popular BERT model, pre-trained on a larger dataset and with improved training techniques. The SQuAD 2 version of RoBERTa has been fine-tuned specifically on the Stanford Question Answering Dataset (SQuAD) 2.0, which includes unanswerable questions. This makes RoBERTa Base SQuAD 2 well-suited for handling complex and diverse questions that may arise from PDF documents, including those that do not have an explicit answer in the text. Its ability to identify when a question is unanswerable can enhance the user experience by providing informative responses instead of incorrect or irrelevant information.

mDeBERTa V3 Base SQuAD 2:
The mDeBERTa V3 Base SQuAD 2 model is another compelling option for question answering in a PDF reader project. mDeBERTa (meta-DeBERTa) is a meta-learning approach that builds upon the DeBERTa (Decoding-enhanced BERT with Disentangled Attention) model. This model has been fine-tuned on the SQuAD 2.0 dataset, making it adept at handling unanswerable questions. One of the key strengths of mDeBERTa is its ability to learn from diverse data sources, potentially giving it a broader understanding of language and context. This characteristic can be particularly beneficial when dealing with PDF documents covering a wide range of topics and domains. Additionally, mDeBERTa has been optimized for improved performance on long-range dependencies, which can be advantageous when tackling complex and lengthy passages found in PDF documents.

DistilBERT Base Cased Distilled SQuAD:
The DistilBERT Base Cased Distilled SQuAD model is a lightweight and efficient option for question answering in a PDF reader project. DistilBERT is a distilled version of the BERT model, meaning it has been compressed to reduce its size and computational requirements while retaining much of its performance. The Base Cased Distilled SQuAD version has been fine-tuned on the SQuAD dataset, making it adept at extracting relevant answers from text passages. While smaller in size, DistilBERT still maintains a strong understanding of natural language, making it a viable choice for question answering tasks. Its reduced computational demands can be advantageous in resource-constrained environments or for applications that require fast inference times, ensuring a responsive user experience in the PDF reader system.

BERT Large Uncased Whole Word Masking SQuAD 2:
The BERT Large Uncased Whole Word Masking SQuAD 2 model is a powerful option for question answering in a PDF reader project. BERT (Bidirectional Encoder Representations from Transformers) is a widely-used and highly effective language model that has been pre-trained on a massive corpus of text. The Large version of BERT offers increased capacity and depth compared to its smaller counterparts, potentially enhancing its ability to

handle complex language tasks. Additionally, the Whole Word Masking technique used during pre-training helps BERT better understand word boundaries and contexts. This model has also been fine-tuned on the SQuAD 2.0 dataset, making it adept at handling unanswerable questions. While computationally more demanding than some other options, the BERT Large Uncased Whole Word Masking SQuAD 2 model can provide superior performance in question answering tasks, particularly when dealing with intricate and challenging text passages found in PDF documents.

Longformer Base 4096 Finetuned SQuAD V2:
The Longformer Base 4096 Finetuned SQuAD V2 model is a specialized choice for question answering in a PDF reader project. Longformer is an extension of the Transformer architecture designed to handle long sequences of text efficiently. The Base 4096 version of Longformer can process input sequences up to 4,096 tokens in length, making it well-suited for tackling lengthy passages and documents often found in PDFs. This model has been fine-tuned on the SQuAD V2 dataset, ensuring its ability to handle unanswerable questions effectively. By combining the strengths of the Transformer architecture with the capability to process long sequences, the Longformer Base 4096 Finetuned SQuAD V2 model can be a valuable asset in a PDF reader system. It can accurately extract relevant information from extensive text passages, while also providing informative responses when a question cannot be answered based on the given content.

Methodology

3.1 Description of dataset

The sample resumes used for evaluating our project are manually sourced and downloaded from ResumeViking. ResumeViking is a platform offering a wide range of downloadable professionally crafted resume templates that cater to varying industry standards and career trajectories.

Section	Field	Description	Example
Personal Details	Name	Applicant's full name	Jason Miller
	Contact Information	Phone number, Email address, Address	3868683442, email@email.com, 1515 Pacific Ave, Los Angeles, CA 90291, United States
Professional Experience	Job Title	Title of the position held	Amazon Warehouse Associate
	Company Name	Name of the employer/company	Amazon
	Dates of Employment	Start and end dates of employment	January 2021 — July 2022
	Job Description	Description of roles, responsibilities	Performed warehouse laborer duties,
	Achievements	Recognitions or accomplishments achieved	Saved 12% on UPS orders by staying on top of special deals. Awarded "Employee of the month" due to consistent attendance, punctuality, and performance.
Education	Degree	Degree attained	Associates Degree in Logistics and Supply Chain Fundamentals
	Institution	Name of the educational institution	Atlanta Technical College
	Graduation Year	Year of graduation	July 2022
	Major	Majors pursued during the degree program	Warehousing Operations, Logistics and Distribution Practices
	Minor	Minors pursued during the degree program	Inventory Systems, Supply Chain Principles
	Course	Course or certification pursued	Online Graduate Certificate in Warehousing & Supply Chain

			Management
	Course Institution	Institution offering the course or certification	Southern New Hampshire University (SNHU)
	Course Duration	Duration of the course	July 2022 — July 2022
Additional Information	Languages Spoken	Languages fluently spoken	English, Spanish
	Place of Birth	Place of birth	San Antonio
	Driving License	Type of driving license held	Full
	Links	Links to professional profiles	LinkedIn
	Skills	Technical or soft skills possessed	Cleaning Equipment, Mathematics, Deep Sanitation Practices
	Hobbies	Interests or recreational activities	Action Cricket, Rugby, Athletics

Table 3.1: Data dictionary of resumes from ResumeViking

Jason Miller

Amazon Associate

Profile

Experienced Amazon Associate with five years' tenure in a shipping yard setting, maintaining an average picking/packing speed of 98%. Holds a zero error% score in adhering to packing specs and 97% error-free ratio on packing records. Completed a certificate in Warehouse Sanitation and has a valid commercial driver's license.

Employment History

Amazon Warehouse Associate at Amazon, Miami Gardens

January 2021 — July 2022

Performed all warehouse laborer duties such as packing, picking, counting, record keeping, and maintaining a clean area.

- Consistently maintained picking/packing speeds in the 98th percentile.
- Picked all orders with 100% accuracy despite high speeds.
- Maintained a clean work area, meeting 97.5% of the inspection requirements.

Laboratory Inventory Assistant at Dunrea Laboratories, Orlando

January 2019 — December 2020

Full-time lab assistant in a small, regional laboratory tasked with · participating in Kaizen Events, Gemba walks, and 5S to remove barriers and improve productivity.

- Filled the warehouse helper job description, which involved picking, packing, shipping, inventory management, and cleaning equipment.
- Saved 12% on UPS orders by staying on top of special deals.
- Cut down storage waste by 23% by switching to a Kanban system.

Education

Associates Degree in Logistics and Supply Chain Fundamentals, Atlanta Technical College, Atlanta

January 2021 — July 2022

- *Majors: Warehousing Operations, Logistics and Distribution Practices*
- *Minors: Inventory Systems, Supply Chain Principles*

Courses

Online Graduate Certificate in Warehousing & Supply Chain Management, Southern New Hampshire University (SNHU), NH.

July 2022 — July 2022

Details

1515 Pacific Ave
Los Angeles, CA 90291
United States
3868683442
email@email.com

Place of birth

San Antonio

Driving license

Full

Links

LinkedIn
Pinterest
Resume Templates
Build this template

Skills

Cleaning Equipment

Cleaning Equipment

Mathematics

Cleaning Equipment

Deep Sanitation Practices

Hobbies

Action Cricket, Rugby, Athletics

Languages

English

Spanish

Warehousing, Operations, and Disposal Course, Graduate School USA, Washington DC.
January 2021 — May 2021

Achievements

- Decrease the errors rate and QC several more orders that we ship per day.
- Awarded "Employee of the month" due to consistent attendance, punctuality, and performance.
- Managed workflow of associates for the FC.

Figure 3.1: Sample resume from ResumeViking

3.2 Applications of the algorithm(s)

3.2.1 RecursiveCharacterTextSplitter

Application:
The RecursiveCharacterTextSplitter is used to divide the resume text into smaller, manageable chunks. This segmentation is essential for detailed analysis and processing, as it allows the embedding models to focus on smaller sections of text, which improves the quality of the embeddings and the efficiency of information retrieval.

Technical Details:
The RecursiveCharacterTextSplitter divides text into smaller, more manageable pieces without losing the context necessary for effective embedding and retrieval. This splitter is designed to cut at logical points that minimize the disruption of semantic meaning, typically at sentence or paragraph boundaries. It often includes functionality to handle overlaps, ensuring that contextual linkage isn't lost between chunks. This tool is especially useful in processing long documents where direct application of NLP models might be computationally expensive or technically impractical.

3.2.2 Embedding Models (All MPNet Base V2 and GTE Base EN V1.5)

Application:
Embedding models are utilized to convert textual data from resumes into numerical form that machines can understand. Specifically, the GTE Base EN V1.5 and All MPNet Base V2 models are applied to generate dense vector representations of text chunks extracted from the resumes. These embeddings capture semantic meanings of the text, enabling the efficient retrieval of information relevant to user queries.

Technical Details:
Embedding models like GTE Base EN V1.5 and All MPNet Base V2 are critical for converting text into high-dimensional vectors that capture semantic meanings. These models use advanced neural network architectures to process text data. GTE Base EN V1.5, for example, might leverage transformer architectures that use self-attention mechanisms to understand the context within sentences. MPNet (Masked and Permuted Pre-training for Language Understanding) is a novel pre-training method that integrates the strengths of masked language modeling and permuted language modeling in a single framework. This allows for a more robust understanding of language context and syntax, crucial for generating meaningful embeddings.

3.2.3 FAISS (Facebook AI Similarity Search)

Application:
Once embeddings are generated, FAISS is used to store these vectors efficiently and to perform fast similarity searches. This library is designed to help locate the nearest neighbors

of any given vector quickly, even within very large datasets, which is perfect for the use case of searching through large numbers of resume chunks.

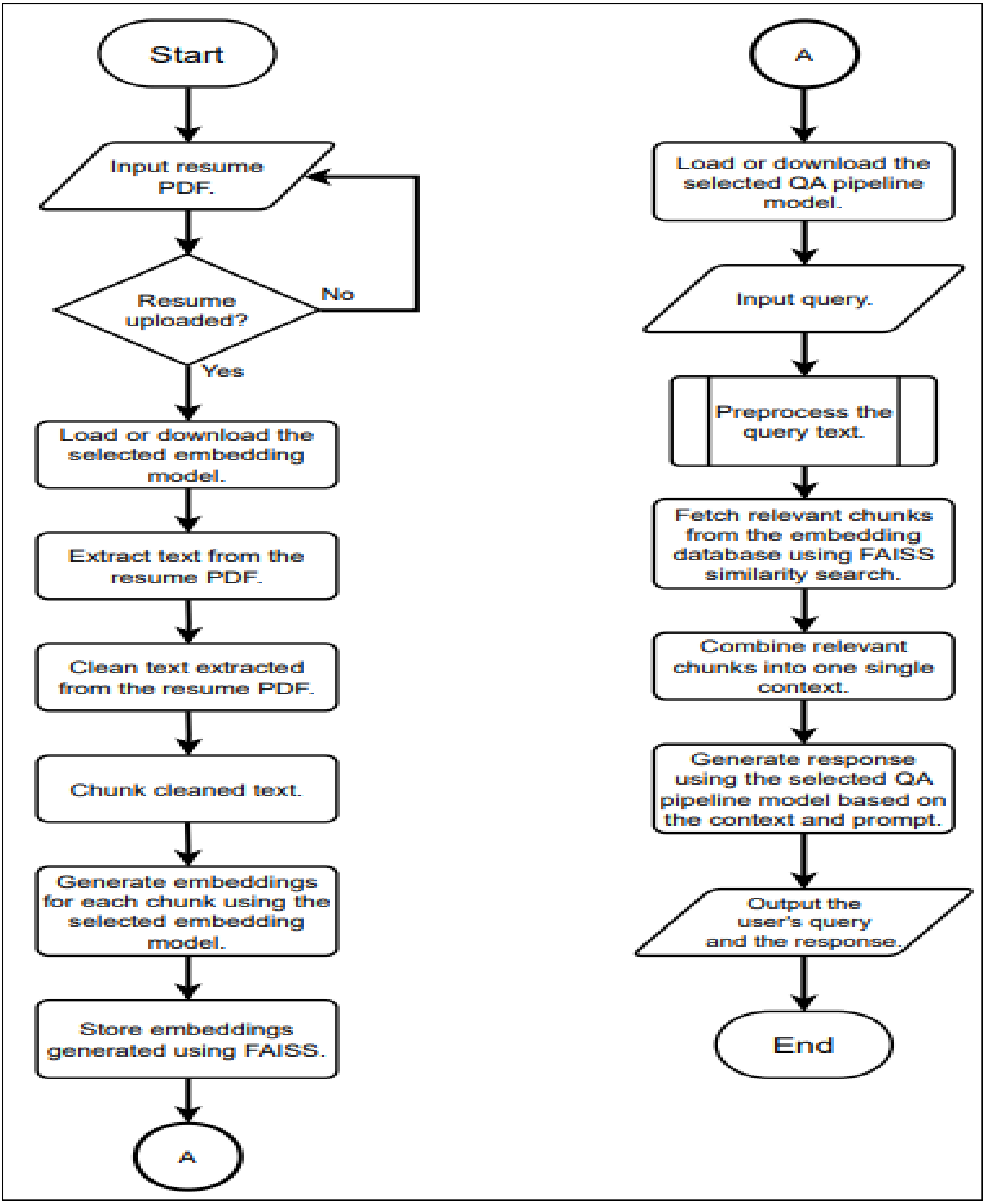
Technical Details:
FAISS is an efficient similarity search and clustering library specifically designed for dense vectors. It contains algorithms that are optimized for various hardware architectures to speed up the search for nearest neighbors in high-dimensional spaces. FAISS uses techniques such as quantization and inverted indexing to achieve high speed and accuracy in vector searches, which significantly enhances the retrieval phase in the application by quickly identifying the best text chunks to use as context for the QA models.

3.2.4 Question Answering Models (BERT Large Uncased Whole Word Masking SQuAD 2, DistilBERT Base Cased Distilled SQuAD, etc.)

Application:
These QA models take the context provided by the closest text chunks (identified via FAISS) and the user's query to generate a specific answer. The models are trained on a wide range of internet texts and fine-tuned on question-answering datasets, which allows them to understand the query and the context and to generate a coherent and contextually appropriate response.

Technical Details:
These QA models are built on top of BERT-like transformer architectures, which are pre-trained on a large corpus of text and then fine-tuned for specific tasks like question answering. BERT Large Uncased Whole Word Masking employs a technique that masks whole words instead of subwords or characters during the pre-training phase, enhancing its understanding of word-level semantics. DistilBERT is a smaller, faster, cheaper variant of BERT that retains most of its predictive power. These models understand context by examining the text on either side of a word within a sentence (bidirectional context), which is essential for generating accurate answers. They are fine-tuned on datasets like SQuAD (Stanford Question Answering Dataset), which trains them to predict the answer spans within the context provided.

3.3 System flowchart



3.3.1 Module 1 - Preprocess the query text

The preprocessing of the query text involves several steps aimed at preparing the user's input for effective question-answering. Steps involved are as follows:

- Normalize whitespaces within the text to ensure consistent formatting and remove any extra whitespace.

- Utilize the Google Translator API to translate the text to English if it's not already in English, ensuring uniform language processing.

- Expand contractions to their full forms, converting expressions like "isn't" to "is not," improving text comprehensibility.

- Convert the text to lowercase, standardizing the text and making it case-insensitive for consistent processing.

- Remove non-alphanumeric characters from the text, eliminating extraneous symbols and focusing solely on the textual content.

- Apply part-of-speech tagging to label each word with its grammatical category, with the purpose of only lemmatizing the verbs.

- Lemmatize verbs to their base forms, ensuring consistency in word representation and reducing lexical variation.

- Filter out stopwords, common words that do not contribute significantly to the text's meaning, enhancing the relevance of the processed text for question-answering tasks.

3.4 Proposed test plan/hypothesis

Test Plan 1: Evaluating Response Quality of Different Embedding and QA Model Combinations

- The test covers two embedding models and five QA models, resulting in ten unique model combinations. Each combination will be tested against 5 predefined questions.

- The test data consists of:
    - **Questions**: Queries that are designed to assess the models' ability to understand and respond appropriately.
    - **Expected Answers**: Benchmarked answers for these questions, which serve as a standard for evaluating the correctness and relevance of the models' responses.

- The evaluation will be based on the following metrics:
    - **Levenshtein Distance**: Measures the number of single-character edits required to change the generated answer into the expected answer, providing a quantitative measure of response accuracy.
    - **Acceptance (Human-Like Performance, HLP)**: A binary metric assessing whether the generated response meets the quality standard of a reasonable human answer.

Test Plan 2: Evaluating Response Time of Different Embedding and QA Model Combinations Based on Queries' Length

- For this test, the same set of resume will be used as the focus is more on how queries' length affect response time. The sample resume used to create queries: Stockholm-Resume-Template-Simple (3).pdf

- A total of 5 queries will be asked, each query will be asked 5 times to the same model.

- The response time taken for each run will be noted down, and the average response time will be used to compare and find out the best model.

Result

4.1 Results

4.1.1 Result of Response Quality

All MPNet Base V2 Embedding Model					
Question	**Expected Answer**	**QA Model**	**Generated Answer**	**Levenshtein Distance**	**Acceptance (HLP)**
Question 1: What is Jason Miller's current job title?	Amazon Associate	RoBERTa Base SQuAD 2	warehouse helper	14	✖
		mDeBERTa V3 Base SQuAD 2	Full-time lab assistant in a small, regional laboratory	47	✖
		DistilBERT Base Cased Distilled SQuAD	Employee of the month	17	✖
		BERT Large Uncased Whole Word Masking SQuAD 2	warehouse helper	14	✖
		Longformer Base 4096 Finetuned SQuAD V2	warehouse helper job description	26	✖

Question 2: For which qualities was Jason awarded "Employee of the month"?	Consistent attendance, punctuality, and performance	RoBERTa Base SQuAD 2		51	✖
		mDeBERTa V3 Base SQuAD 2	consistent attendance, punctuality, and performance.	2	✔
		DistilBERT Base Cased Distilled SQuAD	Employee of the month	44	✖
		BERT Large Uncased Whole Word Masking SQuAD 2	Awarded "Employee of the month	42	✖
		Longformer Base 4096 Finetuned SQuAD V2	Awarded "Employee of the month	42	✖
Question 3: What certification does he hold?	Certificate in Warehouse Sanitation	RoBERTa Base SQuAD 2	Warehousing Operations, Logistics and Distribution Practices	46	✖
		mDeBERTa V3 Base SQuAD 2	Amazon Warehouse Associate	21	✖
		DistilBERT Base Cased Distilled SQuAD	Warehouse Sanitation	15	✔
		BERT Large Uncased Whole Word Masking SQuAD 2	Completed a certificate in Warehouse Sanitation	12	✔

		Longformer Base 4096 Finetuned SQuAD V2	Online Graduate Certificate in Warehousing & Supply Chain Management	39	✔
Question 4: What were Jason Miller's responsibilities as a Warehouse Associate at Amazon?	Performed all warehouse laborer duties such as packing, picking, counting, record keeping, and maintaining a clean area	RoBERTa Base SQuAD 2	Experienced Amazon Associate with five years tenure in a shipping yard setting	89	✖
		mDeBERTa V3 Base SQuAD 2	January 2021 July 2022 Performed all warehouse laborer duties	99	✔
		DistilBERT Base Cased Distilled SQuAD	Managed workflow of associates for the FC	98	✖
		BERT Large Uncased Whole Word Masking SQuAD 2	Performed all warehouse laborer duties	81	✔
		Longformer Base 4096 Finetuned SQuAD V2	Warehousing Operations, Logistics and Distribution Practices	90	✖
Question 5: What are some skills listed?	Cleaning equipment, Mathematics, deep sanitation practices	RoBERTa Base SQuAD 2	Warehousing Operations, Logistics and Distribution Practices	35	✖
		mDeBERTa V3 Base SQuAD 2	Laboratory Inventory Assistant	48	✖
		DistilBERT Base Cased Distilled SQuAD	Online Graduate Certificate in Warehousing & Supply Chain Management	54	✖

		BERT Large Uncased Whole Word Masking SQuAD 2	Cleaning Equipment Cleaning Equipment	33	✔
		Longformer Base 4096 Finetuned SQuAD V2	English Spanish Courses Online Graduate Certificate in Warehousing & Supply Chain Management	72	✖
				Average-Minimum Levenshtein Distance: 28.4 (142/5)	**Acceptance Level: 28% (7/25)**

Table 4.1.1.1: Test Results for "All MPNet Base V2" Embedding Model

GTE Base EN V1.5 Embedding Model					
Question	**Expected Answer**	**QA Model**	**Generated Answer**	**Levenshtein Distance**	**Acceptance (HLP)**
Question 1: What is Jason Miller's current job title?	Amazon Associate	RoBERTa Base SQuAD 2	Amazon Associate	0	✔
		mDeBERTa V3 Base SQuAD 2	Amazon Warehouse Associate	10	✔
		DistilBERT Base Cased Distilled SQuAD	Employee of the month	17	✖
		BERT Large Uncased Whole Word Masking SQuAD 2	Amazon Associate	0	✔
		Longformer Base 4096 Finetuned SQuAD V2	Amazon Associate Profile	8	✔
Question 2: For which qualities was Jason awarded "Employee of the month"?	Consistent attendance, punctuality, and performance	RoBERTa Base SQuAD 2		51	✖
		mDeBERTa V3 Base SQuAD 2	consistent attendance, punctuality, and performance.	2	✔
		DistilBERT Base Cased Distilled SQuAD	Employee of the month	44	✖

		BERT Large Uncased Whole Word Masking SQuAD 2	Awarded "Employee of the month	42	✖
		Longformer Base 4096 Finetuned SQuAD V2	Awarded "Employee of the month	42	✖
Question 3: What certification does he hold?	Certificate in Warehouse Sanitation	RoBERTa Base SQuAD 2	Warehousing Operations, Logistics and Distribution Practices	46	✖
		mDeBERTa V3 Base SQuAD 2	Amazon Warehouse Associate	21	✖
		DistilBERT Base Cased Distilled SQuAD	Warehouse Sanitation	15	✔
		BERT Large Uncased Whole Word Masking SQuAD 2	Completed a certificate in Warehouse Sanitation	12	✔
		Longformer Base 4096 Finetuned SQuAD V2	Online Graduate Certificate in Warehousing & Supply Chain Management	39	✔
Question 4: What were Jason Miller's responsibilities as a Warehouse Associate at	Performed all warehouse laborer duties such as packing, picking, counting, record keeping, and	RoBERTa Base SQuAD 2	Experienced Amazon Associate with five years tenure in a shipping yard setting	89	✖
		mDeBERTa V3 Base SQuAD 2	January 2021 July 2022	110	✖

Amazon?	maintaining a clean area	DistilBERT Base Cased Distilled SQuAD	Jason Miller Amazon	107	✖
		BERT Large Uncased Whole Word Masking SQuAD 2	Amazon Associate	107	✖
		Longformer Base 4096 Finetuned SQuAD V2	Associate with five years tenure in a shipping yard setting	91	✖
Question 5: What are some skills listed?	Cleaning equipment, Mathematics, deep sanitation practices	RoBERTa Base SQuAD 2	Warehousing Operations, Logistics and Distribution Practices	35	✖
		mDeBERTa V3 Base SQuAD 2	Cleaning Equipment Cleaning Equipment Mathematics Cleaning Equipment Deep Sanitation	48	✔
		DistilBERT Base Cased Distilled SQuAD	Skills Cleaning Equipment Cleaning Equipment Mathematics Cleaning Equipment	47	✔
		BERT Large Uncased Whole Word Masking SQuAD 2	Cleaning Equipment Cleaning Equipment Mathematics Cleaning Equipment Deep Sanitation	48	✔
		Longformer Base 4096 Finetuned SQuAD V2	Skills Cleaning Equipment Cleaning Equipment Mathematics Cleaning Equipment	47	✔

				Average-Minimum Levenshtein Distance: 27.6 (138/5)	**Acceptance Level: 48% (12/25)**

Table 4.1.1.2: Test Results for "GTE Base EN V1.5" Embedding Model

4.1.2 Result of Response Time

All MPNet Base V2 Embedding Model							
Constant	**Description**	**Response time (seconds)**					
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What is Jason Miller's current job title?	**Run 1:**	2.7112	3.5089	2.2024	4.6257	3.5109
		Run 2:	1.4989	3.0194	0.9919	3.6373	3.3800
Question length (word count)	7	**Run 3:**	1.6074	3.0661	1.0196	3.5314	3.3501
		Run 4:	1.5405	3.1484	1.0899	3.4879	3.3356
		Run 5:	1.5375	2.9126	1.0147	3.6068	3.6398
Average response time			**1.7791**	**3.1311**	**1.2637**	**3.7778**	**3.4433**

Table 4.1.2.1: Test results using first question and "All MPNet Base V2" embedding model

All MPNet Base V2 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	For which qualities was Jason awarded "Employee of the month"?	**Run 1:**	1.5811	3.5811	1.0569	4.1429	3.5754
		Run 2:	1.5072	2.9572	1.0893	3.7346	3.4371
Question length (word count)	10	**Run 3:**	1.5000	3.0603	1.0564	3.5722	3.3897
		Run 4:	1.4705	2.8071	1.0724	3.5816	3.3686
		Run 5:	1.5499	2.7442	1.0357	3.6613	3.7161
Average response time			**1.5217**	**3.0300**	**1.0621**	**3.7385**	**3.4974**

Table 4.1.2.2: Test results using second question and "All MPNet Base V2" embedding model

All MPNet Base V2 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What certification does he hold?	**Run 1:**	1.5698	3.5304	0.9903	3.8230	3.4595
		Run 2:	1.4693	2.7431	0.9868	3.9119	3.3822
Question length (word count)	5	**Run 3:**	1.4784	2.7506	1.0104	3.3979	3.3514
		Run 4:	1.4804	2.6963	0.9775	3.3893	3.4191
		Run 5:	1.4886	2.7341	0.9745	3.8710	3.5527
Average response time			**1.4973**	**2.8909**	**0.9879**	**3.6786**	**3.4330**

Table 4.1.2.3: Test results using third question and "All MPNet Base V2" embedding model

All MPNet Base V2 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What were Jason Miller's responsibilities as a Warehouse Associate at Amazon?	**Run 1:**	2.6635	3.4626	1.0489	3.9353	3.5332
		Run 2:	1.6671	2.9586	1.0733	3.7013	3.4638
Question length (word count)	11	**Run 3:**	1.5891	2.8888	0.9815	3.8610	3.3901
		Run 4:	1.5629	2.8336	1.0453	3.5268	3.3118
		Run 5:	1.7312	2.9620	1.0147	3.5119	3.6748
Average response time			**1.8428**	**3.0211**	**1.0327**	**3.7073**	**3.4747**

Table 4.1.2.4: Test results using fourth question and "All MPNet Base V2" embedding model

All MPNet Base V2 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What are some skills listed?	**Run 1:**	1.4537	3.0850	1.0380	4.0269	3.6704
		Run 2:	1.5512	2.8105	1.0405	3.6048	3.3099
Question length (word count)	5	**Run 3:**	1.5226	2.7124	1.0148	3.3722	3.3243
		Run 4:	1.5120	2.7113	0.9924	3.3381	3.2912
		Run 5:	1.5382	2.8016	1.0045	3.6385	3.7144
Average response time			**1.5155**	**2.8242**	**1.0180**	**3.5961**	**3.4620**

Table 4.1.2.5: Test results using fifth question and "All MPNet Base V2" embedding model

GTE Base EN V1.5 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What is Jason Miller's current job title?	**Run 1:**	2.8836	3.2556	1.9680	6.3612	3.9558
		Run 2:	1.6920	2.9465	1.1638	3.6133	3.6277
Question length (word count)	7	**Run 3:**	1.6888	2.9710	1.1526	3.5861	3.5036
		Run 4:	1.6463	2.8447	1.1279	3.5372	3.4681
		Run 5:	1.7622	3.2096	1.1780	3.6541	3.6055
Average response time			**1.9346**	**3.0455**	**1.3181**	**4.1504**	**3.6321**

Table 4.1.2.6: Test results using first question and "GTE Base EN V1.5" embedding model

GTE Base EN V1.5 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	For which qualities was Jason awarded "Employee of the month"?	**Run 1:**	1.8265	3.1881	1.1719	8.5252	3.6797
		Run 2:	1.6956	2.8611	1.1463	4.1350	3.6409
Question length (word count)	10	**Run 3:**	1.6449	2.8898	1.1911	3.5730	3.5276
		Run 4:	1.8069	2.7618	1.0805	3.5322	4.3245
		Run 5:	1.7247	3.1050	1.1723	5.2875	3.5659
Average response time			**1.7397**	**2.9612**	**1.1524**	**5.0106**	**3.7477**

Table 4.1.2.7: Test results using second question and "GTE Base EN V1.5" embedding model

GTE Base EN V1.5 Embedding Model							
Constant	**Description**	**Response time (seconds)**					
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What certification does he hold?	**Run 1:**	1.6739	3.2173	1.0893	3.8113	3.6155
		Run 2:	1.6638	2.7635	1.0394	3.5267	3.5020
Question length (word count)	5	**Run 3:**	1.5748	2.6428	0.9901	3.4337	3.4719
		Run 4:	1.5280	2.6719	1.0063	3.3545	3.5732
		Run 5:	1.6040	2.7619	1.0179	4.5860	3.5170
Average response time			**1.6089**	**2.8115**	**1.0286**	**3.7424**	**3.5359**

Table 4.1.2.8: Test results using third question and "GTE Base EN V1.5" embedding model

GTE Base EN V1.5 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What were Jason Miller's responsibilities as a Warehouse Associate at Amazon?	**Run 1:**	1.7689	3.0948	1.1966	3.9894	3.5953
		Run 2:	1.5847	3.0588	1.0733	3.5620	3.4838
Question length (word count)	11	**Run 3:**	1.6485	2.8922	1.1283	3.6706	3.4485
		Run 4:	1.5759	2.7481	1.0478	3.6789	3.4262
		Run 5:	1.6819	2.9337	1.2040	3.7474	3.5273
Average response time			**1.6520**	**2.9455**	**1.1300**	**3.7297**	**3.4962**

Table 4.1.2.9: Test results using fourth question and "GTE Base EN V1.5" embedding model

GTE Base EN V1.5 Embedding Model								
Constant	**Description**	**Response time (seconds)**						
		Run #	**RoBERTa Base SQuAD 2**	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
Question	What are some skills listed?	**Run 1:**	1.6421	3.4582	2.1589	4.3223	3.5389
		Run 2:	1.5177	2.7773	0.9474	3.7363	3.5182
Question length (word count)	5	**Run 3:**	1.5011	2.7049	0.9438	3.4598	3.5292
		Run 4:	1.5067	2.6809	0.9979	3.3177	3.3537
		Run 5:	1.6679	2.8550	1.0687	3.7650	3.4976
Average response time			**1.5671**	**2.8953**	**1.2233**	**3.7202**	**3.4875**

Table 4.1.2.10: Test results using fifth question and "GTE Base EN V1.5" embedding model

4.2 Discussion/Interpretation

4.2.1 Discussion/Interpretation on Response Quality

Embedding Model	**Average-Minimum Levenshtein Distance**	**Acceptance Level (%)**
All MPNet Base V2	28.4 (142/5)	28% (7/25)
GTE Base EN V1.5	27.6 (138/5)	48% (12/25)

Table 4.2.1.1: Response Quality Comparison of Different Embedding Models

- From the table above, the "GTE Base EN V1.5" embedding model not only has a lower average-minimum Levenshtein distance, indicating that responses from this model are generally closer to the expected answers and require fewer corrections to achieve accuracy, but it also demonstrates a higher acceptance level.

- Specifically, 48% of the responses from the "GTE Base EN V1.5" embedding model met the human-like performance criteria, compared to only 28% for the "All MPNet Base V2" model. This higher acceptance rate suggests that the "GTE Base EN V1.5" embedding model is more effective in producing responses that are deemed more coherent and contextually appropriate for human-like interaction.

Minimum Levenshtein Distance/Accepted Count					
	RoBERTa Base SQuAD 2	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**
All MPNet Base V2	1	3	1	7	1
GTE Base EN V1.5	4	4	2	5	3
Total	**5**	**7**	**3**	**12**	**4**
% (/31)	**16.13**	**22.58**	**9.68**	**38.71**	**12.90**

Table 4.2.1.2: Response Quality Comparison of Different QA Models

- The numbers in the table are counted for each question and are determined by:

- Adding 1 to the model if it achieves the minimum Levenshtein distance compared to the other models. If multiple models share the same minimum distance, each model is counted.
- Adding 1 to the model if its response is accepted based on human evaluation.

- The “BERT Large Uncased Whole Word Masking SQuAD 2” QA model shows the highest number of total instances (12), suggesting this model often produced responses that were closest to the expected answers.

- The “BERT Large Uncased Whole Word Masking SQuAD 2” QA model leads when paired with the “All MPNet Base V2” embedding model, with 7 counts.

- Conversely, the “DistilBERT Base Cased Distilled SQuAD” model has the fewest total instances, indicating the poorest performing model.

By considering both of these information from the 2 tables above, the “BERT Large Uncased Whole Word Masking SQuAD 2” QA model might perform the best when combined with the “All MPNet Base V2” embedding model. However, the “GTE Base EN V1.5” embedding model generally performs the best with all QA models.

4.2.2 Discussion/Interpretation on Response Time

All MPNet Base V2 Embedding Model					
RoBERTa Base SQuAD 2	**mDeBERTa V3 Base SQuAD 2**	**DistilBERT Base Cased Distilled SQuAD**	**BERT Large Uncased Whole Word Masking SQuAD 2**	**Longformer Base 4096 Finetuned SQuAD V2**	**REMARK**
1.7791	3.1311	1.2637	3.7778	3.4433	**Q1 AVERAGE**
1.5217	3.03	1.0621	3.7385	3.4974	**Q2 AVERAGE**
1.4973	2.8909	0.9879	3.6786	3.433	**Q3 AVERAGE**
1.8428	3.0211	1.0327	3.7073	3.4747	**Q4 AVERAGE**
1.5155	2.8242	1.018	3.5961	3.462	**Q5 AVERAGE**
1.63128	**2.97946**	**1.07288**	**3.69966**	**3.46208**	**TOTAL AVERAGE**
GTE Base EN V1.5 Embedding Model					
1.9346	3.0455	1.3181	4.1504	3.6321	**Q1 AVERAGE**
1.7397	2.9612	1.1524	5.0106	3.7477	**Q2 AVERAGE**
1.6089	2.8115	1.0286	3.7424	3.5359	**Q3 AVERAGE**
1.652	2.9455	1.13	3.7297	3.4962	**Q4 AVERAGE**
1.5671	2.8953	1.2233	3.7202	3.4875	**Q5 AVERAGE**
1.70046	**2.9318**	**1.17048**	**4.07066**	**3.57988**	**TOTAL AVERAGE**

Table 4.2.2.1 Average response time for each QA model

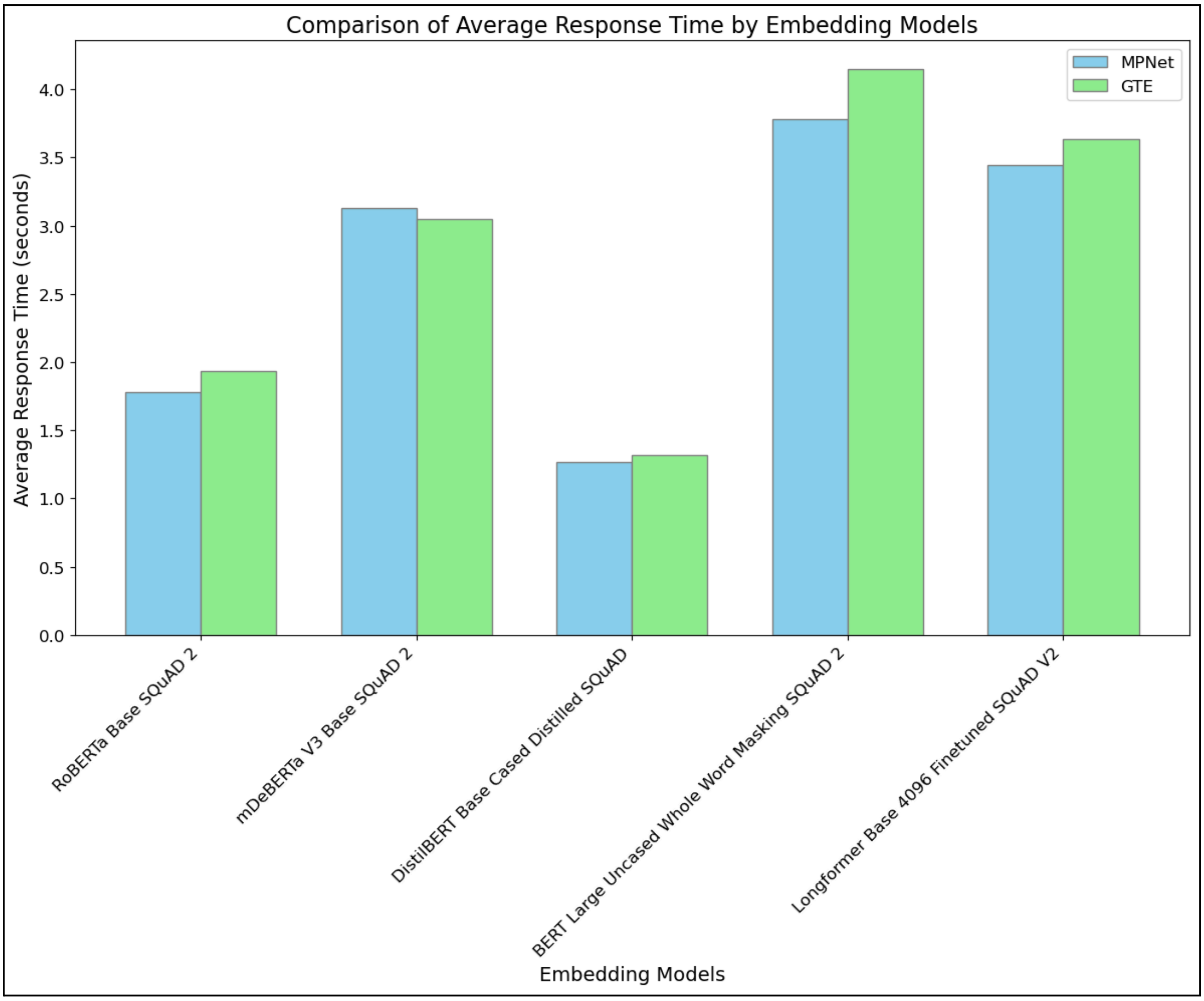


Figure 4.2.2.1 Average response time for each QA model

From the figure and table above, several discussions are made:

- "DistilBERT Base Cased Distilled SQuAD" emerged as the most efficient QA model within both embedding models by exhibiting the lowest average response time, suggesting its effectiveness and efficiency in handling diverse question types.

- "BERT Large Uncased Whole Word Masking SQuAD 2" and "Longformer Base 4096 Finetuned SQuAD V2" consistently have the highest response times, suggesting they might be the most computationally expensive or require more processing for complex tasks. While potentially offering high accuracy, they might not be ideal for applications requiring real-time responses.

- The "All MPNet Base V2" embedding model generally outperformed the "GTE Base EN V1.5" Embedding Model in terms of average response times.

- The result confirms a consistent trend: As question length increases, so does response time for all models and embedding models. Longer questions require the model to process more text and potentially access more information from the document to find the answer.

Discussion and Conclusion

5.1 Achievements

The project has achieved several key functionalities related to building a Resume Question-Answering (QA) application using Streamlit, leveraging various natural language processing (NLP) tools and techniques. Below are the major accomplishments and an assessment of how well the project met its stated objectives:

1) User Interaction:

- Users can upload resumes in PDF format via the Streamlit interface.
- Selection of different embedding models and QA pipeline models is facilitated through interactive radio buttons in the sidebar.
- After uploading a resume and selecting models, users can start question-answering sessions to extract information from the resume content.

2) Model Handling:

- The application handles the loading or downloading of the chosen embedding and QA models, ensuring they are available for use during chat sessions.
- Model loading operations are optimized through Streamlit's caching mechanism, enhancing performance and user experience.

3) Text Preprocessing:

- The application implements comprehensive text preprocessing routines that normalize, translate, expand contractions, convert text to lowercase, remove non-alphanumeric characters, perform part-of-speech tagging, lemmatize, and filter out stopwords in user queries. These steps are crucial for preparing the text for further processing and analysis.

4) Document Processing:

- Resume text is segmented into chunks using the RecursiveCharacterTextSplitter.
- Embeddings of these text chunks are computed using the selected embedding model and are efficiently managed using the FAISS vector store.

5) Question-Answering:

- The application retrieves relevant document chunks based on user queries, forming a context for question-answering.
- These contexts and user questions are fed into the selected QA model to generate accurate and relevant responses, which are then displayed within the chat interface.

Evaluation of Objectives:

Objective 1: To measure the similarity of responses generated by different embedding and question-answering models compared to the provided PDF content

- The project integrates multiple embedding and QA models, producing responses closely aligned with the original resume content. This objective was achieved through both human evaluation and quantitative assessments, identifying "GTE Base EN V1.5" and "BERT Large Uncased Whole Word Masking SQuAD 2" as the optimal models for embedding and QA, respectively, based on their performance in Levenshtein distance and human-like response quality.

Objective 2: To evaluate the response time of different embedding and question-answering models based on different queries' length

- The application tracks and analyzes response times from query initiation to answer delivery. This systematic evaluation revealed that "All MPNet Base V2" provides the fastest response times among embedding models, and "DistilBERT Base Cased Distilled SQuAD" excels among QA models, particularly with shorter query lengths. These insights are supported by data collected across various operational scenarios, ensuring robust performance evaluation.

In conclusion, the project has successfully met its objectives in developing a robust Resume QA application. Through comprehensive model integration, efficient processing, and accurate question-answering capabilities, the application effectively assists users in extracting valuable information from their resumes.

5.2 Limitations and Future Works

Limitations	Future works
Single-Language Processing ● The system is designed to process queries in a single language effectively, such as English or Chinese. It assumes that the entire query will be written in one language. Queries that mix multiple languages together, commonly known as "rojak," present a challenge for the system. It may struggle to accurately interpret and understand the meaning of such queries due to the complexity introduced by multiple languages.	This enhancement will involve developing language identification algorithms and integrating multilingual models to accurately interpret and respond to rojak queries. Additionally, language-specific preprocessing techniques will be incorporated to improve the chatbot's understanding of multilingual nuances.
Inadequate Handling of Dialects and Slang ● The chatbot may struggle to understand queries containing dialectal variations or informal slang expressions not present in its training data. It may misinterpret or fail to recognize dialect-specific vocabulary and colloquialisms.	To overcome this limitation, the chatbot's training data will be augmented with dialectal variations and colloquial expressions to enhance its understanding of regional language nuances. Furthermore, dialect identification mechanisms and slang dictionaries will be integrated to accurately interpret and respond to dialect-specific queries. Collaboration with linguists and language experts will be sought to enhance the chatbot's proficiency and adaptability in handling dialectal variations.
Audio Input Processing ● Currently, the chatbot is only capable of processing textual queries, thereby excluding users who prefer or require audio input methods. Without support for audio input processing, users are constrained to interact with the chatbot exclusively through text-based interfaces, hindering accessibility and usability in scenarios where voice-based interaction is preferred or necessary.	To address this limitation, the integration of speech-to-text (STT) technology is crucial. By implementing STT capabilities, the chatbot can transcribe spoken language into text, enabling users to communicate with the system via voice commands or queries. Additionally, thorough testing and refinement of the STT functionality will be conducted to ensure accurate transcription and seamless integration with the chatbot's existing features. Further exploration of advancements in natural language understanding (NLU) and voice recognition algorithms will be pursued to enhance the chatbot's ability to interpret diverse speech patterns and accents, thereby improving the overall user experience for audio input processing.
Short Answers without Contextual Detail ● Currently, the project provides succinct answers devoid of contextual detail. For instance, when prompted about specific details like age, the response is a bare numerical value without accompanying context. This limitation results in responses that lack depth and fail to provide users with a complete understanding of the information being conveyed. Without contextual detail, users may find it challenging to interpret the responses	To enhance user experience, future work could focus on providing more informative and contextually rich responses. This may involve developing algorithms to analyze query context and resume information, enabling the chatbot to generate responses with appropriate detail and descriptive language. Improvements in natural language understanding and integration of response variation techniques could further tailor responses to individual preferences. Additionally, implementing feedback mechanisms for users to

accurately and may desire more informative answers to their queries. Consequently, this limitation hampers the user experience and diminishes the overall utility of the project.	provide input on response quality could drive continuous improvement.

Reference & Source

Alibaba-NLP/GTE-base-en-V1.5 · Hugging face. (n.d.). https://huggingface.co/Alibaba-NLP/gte-base-en-v1.5

All Mpnet Base V2 by Sentence Transformers | AI model details. (n.d.). https://www.aimodels.fyi/models/huggingFace/all-mpnet-base-v2-sentence-transformers

allenai/longformer-base-4096 · Hugging Face. (n.d.). https://huggingface.co/allenai/longformer-base-4096

Artifex. (n.d.). PyMuPDF 1.24.3 documentation. https://pymupdf.readthedocs.io/en/latest/index.html

Beltagy, I., Peters, M. E., & Cohan, A. (2020, April 10). Longformer: The Long-Document Transformer. arXiv.org. https://arxiv.org/abs/2004.05150

Consensus/All-MiniLM-L6-V2 · Hugging face. (2001, January 18). https://huggingface.co/Consensus/all-MiniLM-L6-v2

deepset/bert-large-uncased-whole-word-masking-squad2 · Hugging Face. (2001, March 4). https://huggingface.co/deepset/bert-large-uncased-whole-word-masking-squad2

deepset/deberta-v3-base-squad2 · Hugging Face. (2001, March 4). https://huggingface.co/deepset/deberta-v3-base-squad2

deepset/roberta-base-squad2 · Hugging Face. (2001, March 4). https://huggingface.co/deepset/roberta-base-squad2

distilbert/distilbert-base-cased-distilled-squad · Hugging Face. (2001, March 4). https://huggingface.co/distilbert/distilbert-base-cased-distilled-squad

Frequently asked questions — pdfminer.six 20221105.post22+git.9cc4d1dd documentation. (n.d.). https://pdfminersix.readthedocs.io/en/latest/faq.html

google-bert/bert-large-uncased-whole-word-masking · Hugging Face. (2001, March 11). https://huggingface.co/google-bert/bert-large-uncased-whole-word-masking

He, P., Gao, J., & Chen, W. (2021, November 18). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv.org. https://arxiv.org/abs/2111.09543

History of pypdf — pypdf 4.2.0 documentation. (n.d.). https://pypdf.readthedocs.io/en/stable/meta/history.html

How LLMs Are Supercharging AI-Powered Talent Intelligence. (2023). HRNews, https://www.proquest.com/trade-journals/how-llms-are-supercharging-ai-powered-talent/docview/2846240554/se-2

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). ROBERTA: A robustly optimized BERT pretraining approach. arXiv.org. https://arxiv.org/abs/1907.11692

mrm8488/longformer-base-4096-finetuned-squadv2 · Hugging Face. (2001, March 4). https://huggingface.co/mrm8488/longformer-base-4096-finetuned-squadv2

Muludi et al. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. ProQuest. https://doi.org/10.14569/IJACSA.2024.0150379

Paper page - Towards General Text Embeddings with Multi-stage Contrastive Learning. (2001, March 10). https://huggingface.co/papers/2308.03281

Pdfminer.six. (2023, December 28). PyPI. https://pypi.org/project/pdfminer.six/

PyMuPDF. (2024, May 9). PyPI. https://pypi.org/project/PyMuPDF/

PyMuPDF - Artifex. (n.d.). https://artifex.com/products/pymupdf/

pypdf. (2024, April 7). PyPI. https://pypi.org/project/pypdf/4.2.0/

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019, October 2). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv.org. https://arxiv.org/abs/1910.01108

sentence-transformers/all-MiniLM-L6-v2 · Hugging Face. (2001, January 18). https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

sentence-transformers/all-mpnet-base-v2 · Hugging Face. (2001b, January 18). https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Shien, L. Y. (2022, January 7). What is Natural Language Processing: The Definitive Guide. Speak Ai. https://speakai.co/what-is-natural-language-processing-the-definitive-guide/

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. (2020, April 20). MPNET: Masked and Permuted pre-training for language understanding. arXiv.org. https://arxiv.org/abs/2004.09297

thenlper/gte-large - Demo - DeepInfra. (n.d.). https://deepinfra.com/thenlper/gte-large

timpal0l/mdeberta-v3-base-squad2 · Hugging Face. (2001, March 4).

https://huggingface.co/timpal0l/mdeberta-v3-base-squad2

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020, February 25). MINILM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv.org. https://arxiv.org/abs/2002.10957

Welcome to pdfminer.six's documentation! — pdfminer.six 20221105.post22+git.9cc4d1dd documentation. (n.d.). https://pdfminersix.readthedocs.io/en/latest/

Welcome to pypdf — pypdf 4.2.0 documentation. (n.d.). https://pypdf.readthedocs.io/en/latest/

What are some algorithms for comparing how similar two strings are? (n.d.). Stack Overflow. https://stackoverflow.com/questions/15303631/what-are-some-algorithms-for-comparing-how-similar-two-strings-are