EXPLORING HYBRID SEARCH FOR TEXT SUMMARIZATION: A STUDY OF RAG BASED METHODS WITH ACCURACY EVALUATION

SUSHANT SUR

Final Thesis Report

APRIL 2025

Dedication

Text summarization has made significant progress in the past with the emergence from the basic deep learning models to advancement of hybrid search strategies, which combine the retrieval and generation approaches back to back in order to enhance the quality of the text generated. This study explores the effectiveness of the hybrid search techniques specifically focusing on Retrieval Augmented Generation based methods, for improving text summarization accuracy. I want to propose a novel framework that integrates RAG based retrieval mechanisms with start of the art generative models to produce concise and informative summaries. The research evaluates the proposed methods against standard benchmarks using various kind of accuracy metrics, including ROGUE score, BLUE score etc in order to access their performance. Our findings demonstrate that the hybrid approach significantly outperforms the traditional methods in terms of relevance and coherence. The results would highlight the potential of combining retrieval and generative methods to advance the field of text summarization. This comprehensive analysis dives deep into the latest advancements, revealing game-changing insights that could reshape the future. From innovative techniques to practical implications, every finding holds the power to transform how we approach complex challenges.

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to all those who have supported and guided me throughout the course of this research. First and foremost, I am profoundly grateful to my supervisor, for their invaluable guidance, expertise, and encouragement. Their insightful feedback and unwavering support were instrumental in shaping this study on hybrid search methods for text summarization.

Special thanks to my colleagues and fellow researchers in the Ayush Kiran Pujari, Dinesh Gaonkar for their collaboration and for providing a stimulating research environment. Their discussions and shared knowledge were crucial in refining the methodologies used in this study.

I am grateful to LJMU for providing the resources and facilities necessary for conducting this research. The availability of computational resources and access to relevant datasets was essential for the success of this project. Lastly, I would like to acknowledge the support of my family and friends, whose encouragement and understanding have been a constant source of motivation throughout this journey.

Thank you all for your contributions and support.

ABSTRACT

In the recent years, advancement in text summarization has been significantly influenced by various kind of hybrid search approaches, which was a mix of retrieval based and generation based methods. This study takes a look with studying the application of Retrieval-Augmented Generation (RAG) strategies to improve text summarization strategy. I am trying to bring up a singular hybrid framework RAG primary based retrieval along with generative models in order to bring up more coherent and contextually relevant text or summaries. Our methodology involves evaluating this hybrid approach against the established benchmarks using accuracy metrics such as ROGUE, BLUE, Uptrain and RAGAS scores. Our methodology involves meticulously evaluating this innovative hybrid approach against established industry benchmarks, meticulously analyzing key accuracy metrics such as the widely-recognized ROUGE scoring system. With various metric scores obtained based on the experiments conducted with Gemini and GPT 40-mimni model we come to the conclusion that although both of them give pretty good results still however the inference time for Gemini was faster as compared to GPT 4o-mini. The text generated was able to withstand with respect to the answer completeness or answer conciseness as metric parameters from RAGAS. The experimental results we've obtained reveal that the proposed method substantially elevates the quality of summarized/generated content, significantly outperforming traditional summarization/generation techniques in terms of relevance and informativeness. This pioneering research makes a valuable contribution to the field, providing researchers and practitioners with a powerful new tool to enhance the efficiency and effectiveness of their summarization efforts.

TABLE OF CONTENTS

DEDICATION 2
ACKNOWLEDGEMENTS 3
ABSTRACT 4
LIST OF TABLES 7
LIST OF FIGURES 8
LIST OF ABBREVIATIONS 10
CHAPTER 1: INTRODUCTION 11
1.1 Background of the Study 11
1.2 Problem Statement 15
1.3 Aim and Objectives 16
1.4 Research Questions (IF ANY) 18
1.5 Scope of the Study 18
1.6 Significance of the Study 19
1.7 Structure of the Study 21
CHAPTER 2: LITERATURE REVIEW 22
2.1 Introduction 22
2.2 Origin and Necessity of Text Summarization 23
2.3 Advancements in Summarization Techniques 29
2.4 Recent Innovations: The Role of Retrieval-Augmented Generation (RAG) 32
2.5 Discussion 35
2.6 Summary 36

CHAPTER 3: RESEARCH METHODOLOGY 30
3.1 Introduction 30
3.2 Research Methodology 30
3.2.1 Data Selection 32
3.2.2 Data Pre-processing 33
3.2.3 Model Selection 34

3.2.4 Hybrid Search Implementation 34
3.2.5 Experimental Setup 36
3.2.6 Acccuracy Measurement 37
3.2.7 Result Analysis 38
3.3 Summary 41
CHAPTER 4: ANALYSIS 43
4.1 Introduction 43
4.2 Dataset Description 43
4.3 Data Preparation 44
4.3.1 Basic Statistics 44
4.4 Model Hyperparameters 50
4.5 Model Implementation 52
4.5.1 Model Selection and Integration 52
4.5.2 Prompt Design 52
4.5.3 Hybrid Search Implementation 53
4.6 Summary 53
CHAPTER 5: RESULTS AND DISCUSSIONS 55
5.1 Introduction 55
5.2 Findings and Discussion 56
5.3 Interpretation of Results 56
5.4 Summary 57
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS 58
6.1 Introduction 58
6.2 Conclusion 58
6.3 Contribution to knowledge 59
6.4 Future Recommendations 60
REFERENCES 62
APPENDIX A: RESEARCH PROPOSAL 167

LIST OF TABLES

Figure 3.2.4.1 Architecture Design …….40
Figure 5.2.1 GPT Scores before using a RAG approach ……....55
Figure 5.2.2 Gemini Scores before using a RAG approach ……56
Figure 5.2.3 Scores after using a RAG approach ……56
Figure 5.2.4 Scores after using a RAG approach …… 57

LIST OF FIGURES

Figure 3.2.1.1 CNN Daily Mail Dataset Format …….. 37

Figure 3.2.1.2 RAG Mini Wikipedia Dataset Format …….. 37

Figure 3.2.1.3 RAG mini Bioasq Dataset Format …… 38

Figure 4.3.1 Overall length of the dataset …… 45

Figure 4.3.2 Word Cloud …….. 46

Figure 4.3.3 Statistics(RAG Mini Wiki) …….. 48

Figure 4.3.4 Document Length(RAG Mini Wiki) …… 48

Figure 4.3.5 Statistics(RAG Mini Bioasq) …… 50

Figure 4.3.6 Document length(RAG Mini Bioasq) …….. 50

LIST OF ABBREVIATIONS

1. RAG - Retrieval-Augmented Generation
2. NLP - Natural Language Processing
3. ML - Machine Learning
4. AI - Artificial Intelligence
5. IR - Information Retrieval
6. TF-IDF - Term Frequency-Inverse Document Frequency
7. BERT - Bidirectional Encoder Representations from Transformers
8. GPT - Generative Pre-trained Transformer
9. ROUGE - Recall-Oriented Understudy for Gisting Evaluation
10. BLEU - Bilingual Evaluation Understudy
11. LM - Language Model
12. RNN - Recurrent Neural Network
13. SOTA - State of the Art
14. F1 - F1 Score
15. QA - Question Answering
16. ETC - Entity, Topic, and Context
17. Qdrant – Open Source Advanced Vector Search
18. PR - Precision-Recall
19. SR - Summarization Research
20. TS - Text Summarization
21. IR - Information Retrieval
22. PS - Precision Score
23. RC - Retrieval Component
24. FC - Fine-tuning Component

CHAPTER 1

INTRODUCTION

1.1 Background of the study

The purpose of the summary is to condense a long document into a short summary while retaining key information. Traditional methods often rely on statistical methods or extrapolation techniques, which may be limited in their ability to produce coherent and informative data sets. After the evolution with the traditional methods Retrieval-augmented generation (RAG) has emerged as a promising method for summarization. Here RAG combines various information retrieval and text generation methods through which it allows to get more accurate and contextual relevant summaries or texts.

However there were some limitations on the keyword based search techniques. Some of them are they fail to uncover various conversions or forms of a single word. Additionally if there is a word that is misspelled or a spelling not updated correctly then basic keyword search will miss to treat these issues in the context. Some of them can be treated through the use of wildcard operations which will be helpful to find our different variations of a single word. One of the way to use the best wildcard operations are “*” or “!”which will be helpful to find different forms of a word. For example lets assume I was working on a legal case where I had to find out documents and email related to a particular topic in order to determine that if an organization had knowledge of the potential breach of code of conduct. I could try to search for words such as misconduct or malconduct in pace of conduct, I could search for unethical or unethically in place of ethical, I could also search for fraudulent or fraudster in place of words called as fraud. Additionally we could also combine the search terms as "code of conduct AND breach" or "conflict of interest OR bribery" NEAR "ethics" or "unethical conduct NOT training" which will effectively create a better

search for documents or emails related to potential breach of code of conduct in order to helping out to identify any evidence and patterns of misconduct. Although wildcard operations help to find out different variations of a word but still they have some limitations as well. If in case a word is not spelled correctly or if there is a abbreviation then there are chances that the wildcard operation will miss them out.

There can be additional limitations and risks associated with relying solely on the keyword searches for document and privilege assessments in legal proceedings. Let me highlight some of the mistakes from the past cases like Enron and Victor Stanley. Enron's legal team had heavily relied on the keyword based searches in order to identify and review potential privileged documents. The keyword lists used were criticized for being incomplete and inaccurate finally being failed to capture many relevant documents. Various privileged documents including the attorney-client communications as well as work product were inadvertently produced due to low limitations of the keyword based search. In fact, improper identification of documents as non-privileged resulted in serious legal repercussions like exposure of privileged information and potential lawyer fines. Some of the lessons we learn out of the Enron case here is keywords searches are supplemented by an indepth manual analysis, particularly for more complex or sensitive documents.

There is another concept of Boolean search which adds as additional extension to keyword based search allowing us to multiple keywords together or exclusive of each other with a certain distance from each other. Boolean searching for text would be mainly based out of some underlying logic which states a articular condition to be a true or false statement and would use some of the standard operators in order to interlink the search terms. Some of the standard operators can be used as and, or, not, within, near or more. Boolean search is a techniques used more for information retrieval which allow users to refine and filter search results on based f a criteria. It leverages Boolean logic—derived from Boolean algebra, a branch of mathematics formulated by George Boole in the 19th century. This is mainly used in case of database search, search engines and information retrieval system in order to enhance the precision and relevance of research results. This helps to improve the relevancy of the documents identified by the search methodology. The OR operator is used to find documents containing one term or another. For instance, searching for vacation OR holiday will return any document that contains either the word "vacation" or "holiday." This

operator is useful for capturing synonyms or related concepts in your search results. Conversely, the NOT operator excludes documents containing a specified term. For example, apple NOT orange would return documents that mention "apple" but do not include the word "orange." This is helpful for filtering out irrelevant results that might otherwise clutter your search results.

The WITHIN or NEAR operator allows you to search for terms that appear within a certain distance from each other, either in terms of words or characters. For example, climate w/10 change would return documents where "climate" appears within ten words of "change." Similarly, health w/p medicine would find documents where "health" and "medicine" appear within the same sentence, and technology w/s innovation would return results where both terms are in the same sentence.

Some of the other useful Boolean search operators include:

a) **Phonic Searching**: This can find words that sound similar. For example, searching for Fleming with a phonic operator might also return results for Flemming, capturing variations in spelling based on pronunciation.

b) **Stemming**: This helps find variations of word endings. For instance, searching for run with stemming enabled would also include results for "running," "runner," and "ran." This is useful for capturing different forms of a word in search results.

Another search technique can be considered is a Semantic Search which is a data searching technique in the information retrieval world which can understand, explore and conceptualize user intent of its context; with it has an understanding that doesn't treat all instances as mere words enabling highly accurate results. While semantic search is primarily used in web searches like the one google does, there are also some other areas using it for content management systems internal corporate chatbots-commerce platforms.

Standard word search mechanisms, Lexical Search only searches for the precise categories of words within a query used by someone doing a web search. This works in unearthing the direct matching terms, but it does not cover the comprehending possibilities like homonyms, synonyms and context determined senses.

By comparison, a search that is semantically-driven operates with the intention to determine what it thinks its user's real goal was and looks for results based on the context regardless of whether those words appear in all their complexity within some highly original query. In short, semantic search algorithms interpret what users want to find rather than just their literal words. This is where semantic search algorithms come in, using external sources for this information such as knowledge graph databases, ontologies (a list of terms relevant to a particular topic or field), and corpora. It also includes user context, like location and search history.

While semantic search algorithms offer a significant edge over traditional keyword-based searches due to their ability to understand context and meaning, they come with their own set of challenges. Their complex architectures make them more powerful but also much harder to design, build, and maintain. These systems often require regular updates and fine-tuning to stay effective, which means they need a solid understanding of machine learning and specialized tools. This level of expertise and resource commitment can be a barrier for smaller organizations and individual researchers who might not have access to such advanced skills or infrastructure. Even though being so powerful, they come with hefty demands in terms of computational resources. Their size and complexity require substantial processing power and memory to operate effectively. As the volume of data being analyzed grows, so do these resource needs. This means setting up, running, and managing the necessary computational infrastructure can be quite expensive and energy-intensive. Moreover, this high energy consumption raises concerns about environmental sustainability, making it a significant consideration for anyone implementing these advanced systems.

One of the key strengths of semantic search algorithms is their ability to grasp the specific context of a user's search, which helps in delivering more relevant results. However, achieving this level of understanding requires tracking and analyzing various pieces of user data, like location, browsing habits, and search history. While this enhances the search experience, it also brings up significant privacy concerns. People may feel uncomfortable with their personal data being monitored, and there could be regulatory issues, especially in places with strict data protection laws like the European Union's General Data Protection Regulation (GDPR).

Hybrid search is all about combining different search methods to get better, more accurate results. Here search strategies can be anything keyword based search or semantic based search etc. Instead

of relying solely on traditional keyword searches, it blends these with advanced techniques like natural language processing (NLP), semantic search, and machine learning.

This approach has found practical uses in various areas. For instance, enterprise search engines that use hybrid search help employees quickly locate the right information within a company's knowledge base. E-commerce sites are also incorporating hybrid search to enhance their search features, making it easier for customers to find exactly what they're looking for, even if they don't know the precise name of the product. Even traditional web search engines are adopting hybrid search to deliver more relevant and precise results to users.

1.2 Problem Statement

In spite of multiple advancements happening in search technologies, traditional methods often fall short when it comes to providing accurate and contextually relevant text summaries. Current approaches can struggle with understanding the nuances of language and the context in which information is presented. Additionally, while semantic search and large language models (LLMs) offer improvements, they come with high computational costs and complex implementation challenges.

This research aims to address these issues by exploring a hybrid search approach that combines multiple search techniques with Retrieval-Augmented Generation (RAG). The ultimate goal is to improve the quality and relevance of text summaries or text generation while managing the practical challenges associated with computational resources and data privacy. By investigating how different search strategies can be effectively integrated and evaluating the impact of reranking, this study seeks to enhance both the accuracy and efficiency of text summarization, making these advanced technologies more accessible and effective for various applications.

In addition to discussing the current problems I see that as per the paper where the researchers tested on the CLERC dataset containing legal documents, the researchers tried to evaluate models on their ability to find corresponding citations for a given piece of legal analysis and compile the text of these citations (as well as previous context) into a cogent analysis that supports a reasoning goal. While doing a benchmarking on the current state of the art models on CLERC dataset it was realized that current approaches still face challenges :- GPT4o is able to generate context or

analysis with the highest ROGUE F scores but at the same time hallucinates most of the time, additionally zero shot Information Retrieval models were only able to achieve 48.3% recall @ 1000. (Hou et al., 2024)

Another problem found out in the paper with chain of agents was current LLM Models was not able to effectively process long context which was a critical issue. Some of the key strategies emerged was either to reduce the input length which involves breaking down information into smaller, relevant chunks using Retrieval-Augmented Generation (RAG) to make it easier to process and summarize effectively. Another strategy was to focus on increasing the amount of context that large language models (LLMs) can handle at once, allowing them to better understand and generate summaries from larger pieces of text. But it was soon realized that the above strategies had their drawbacks as well. Reducing the input length can sometimes miss important information, while expanding the context window of large language models (LLMs) often struggles with keeping focus on the most relevant details. To address these issues, a new approach was proposed called Chain-of-Agents (CoA). (Zhang et al., 2024)

CoA is an innovative framework that uses a team of specialized agents working together through natural language to handle long and complex tasks. It features several worker agents that manage different parts of the text in sequence. After these workers have processed their segments, a manager agent brings all their contributions together to produce a clear and cohesive final summary. This method aims to effectively aggregate information and reason across extended contexts, improving the quality of text summarization.

1.3 Aims and Objectives

The main goal of this research is to develop an improved hybrid search strategy using RAG (Retrieval-Augmented Generation). Additionally, I plan to explore whether reranking can enhance the overall quality of the search results. To achieve this, I will investigate various search techniques and experiment with the available dataset to see how these approaches can be optimized for better performance.

The research objectives for this study are centred around our main aim and include:-

- **To Implement Various Hybrid Search Techniques with Retrieval-Augmented Generation (RAG)**:- This involves integrating and designing several hybrid search strategies within the RAG framework. The goal is to combine different search methods to improve both the retrieval and generation capabilities of the models we test. By doing so, the study hopes to enhance the effectiveness of the search results and the quality of generated content.
- **To Test Different Weight Assignments for Search Techniques**:- This objective is all about exploring how different weightings for various search techniques affect the hybrid model's performance. The study will experiment with adjusting these weights to find the best combination that improves both performance and accuracy.
- **To Evaluate and Compare Hybrid Search Methods**:- To properly assess and compare the effectiveness of different hybrid search strategies, the study will use specific evaluation metrics. This comparative analysis will help determine which hybrid approach works best for tasks like text summarization and other applications, providing insights into the most effective methods.
- **To Develop Scripts to Automate Testing and Evaluation**:- I plan to create scripts to automate the testing and evaluation processes. This will be crucial for ensuring the experiments are efficient and consistent. Automation will streamline the workflow and reduce the need for manual intervention, making the entire process smoother.
- **To Create Logging Tools to Track Performance Metrics**:- I will also develop logging tools to systematically record and manage performance metrics. These tools are essential for monitoring how well the hybrid search techniques perform over time and for analysing their effectiveness in different scenarios.

1.4 Research Questions

Below are the research questions based on the study done:

- ✓ How do hybrid search techniques enhance the quality of text summarization or other use cases compared to traditional methods? This question explores whether using hybrid search methods leads to better text summaries than conventional approaches.

- ✓ What's the best mix of keyword-based and other search techniques for achieving optimal text generation results? This aims to find out which combination of search techniques works best for creating effective summaries.

- ✓ Does reranking make the summaries produced by hybrid search methods more accurate? Here, we'll investigate whether adjusting the ranking of results improves the accuracy of summaries generated through hybrid search.

- ✓ How do the new hybrid search and reranking methods stack up against existing text summarization techniques in terms of performance and accuracy? This question seeks to compare the performance and accuracy of the proposed methods with those of current text summarization techniques.

- ✓ Can the new hybrid search methods be effectively applied to different types of documents, such as news articles, research papers, and large text corpora? This explores whether the proposed hybrid search methods work well for summarizing various types of documents.

1.5 Scope of the study

In case of in scope I'll be working with multiple documents as the source of data for my project. First, I'll index these documents so they're ready for searching. When I enter a query, I'll use embeddings to represent it, which will then be used in a similarity-based hybrid search to find the most relevant documents. A reranking mechanism will help refine this search, bringing the top 2 or 3 most relevant documents to the forefront. I'll then use these documents to extract an answer using a large language model (LLM). Finally, I'll measure how accurate the answers are to evaluate the performance of the hybrid search approach.

Considering out of scope I would start by fine-tuning the model with multiple documents to enhance its performance. Next, I would use these documents as data sources for indexing. After that, when I input a query, I'd generate its embedding and use it for a similarity-based hybrid search. The reranking process would then help identify the top documents to focus on. The final

step involves sending the input query to a large language model (LLM) to generate an answer. To assess the performance, I would compare this answer with a human-generated response. This comparison will help us understand how well the model performs.

1.6 Significance of the Study

Here I am focussing on the exploring the hybrid search for text summarization or other use cases with a RAG based approach and accuracy evaluation. However there are some of the key points to explain here. By combining multiple search strategies or creating a hybrid search technique with Retrieval-Augmented Generation (RAG), my study aims to enhance how well we can summarize text or generate text. The goal is to make summaries more relevant and accurate, which can be incredibly useful for various users. Whether it's researcher's need of concise information, journalists summarizing news, or businesses extracting key points from documents, better text summarization can significantly improve the usefulness and clarity of the information provided.

This study also explores newer ways to combine different search methods to find out the best approach for summarizing text. By experimenting with various hybrid search techniques, we aim to discover which combinations work best to enhance search results. This innovative approach not only improves how we find information but also offers fresh insights into optimizing search methods, making it easier to get the most relevant and useful results from a variety of data sources. Additionally I am also looking at how reranking can improve the accuracy of text summaries. Reranking involves adjusting the order of search results to better match the query, which could lead to more precise and relevant summaries. By evaluating how this process affects the quality of the summaries, we aim to find out if and how reranking can make a difference in getting the most accurate and useful information from the search results.

This study isn't just about improving text summarization in theory; it's about seeing how well these new methods work in real-world scenarios. By testing the hybrid search techniques on different types of documents, like news articles, research papers, and large text corpora, we can understand their practical value. This means that our findings could be beneficial across various fields, from journalism and academia to business and beyond. In other words, we're looking to see how these advancements can make a real impact in everyday applications where effective summarization is

key. Another aspect of the study is developing tools to help evaluate and track how well our models perform. By creating automated scripts and logging tools, we'll be able to systematically test and monitor the hybrid search techniques and their effectiveness. This ensures that our evaluations are thorough and consistent, helping us understand how well the methods work and where improvements might be needed. Essentially, we're setting up a way to accurately measure and compare performance, which is crucial for refining and advancing these search techniques.

I will also be discussing the real-world issues of managing the computational resources needed for advanced search methods. By focusing on optimizing performance and finding practical solutions, we aim to make these sophisticated techniques more accessible and efficient. This is important for organizations and researchers who might not have unlimited resources, as it helps them handle complex algorithms without overwhelming their systems or budgets. In short, we're working on making these powerful tools more practical and feasible for everyday use. There is still a lot of importance of handling data responsibly and adhering to privacy regulations. By carefully managing how data is used and ensuring compliance with laws like the GDPR, we aim to set a standard for ethical practices in search technology. This focus will help to ensure that our methods not only perform well but also respect users' privacy and meet legal requirements.

1.7 Structure of the Study

Chapter 1: Introduction and Context

In the first chapter, we dive into the current advancements in text summarization. We explore the latest techniques and technologies that are shaping this field, along with the challenges that come with them. This chapter covers various search methods and their limitations, and discusses how these methods can be improved. We also address the shortcomings of current large language models (LLMs), such as their tendency to produce inaccurate information (hallucinations) and the difficulties in effectively implementing search strategies. The goal here is to set the stage by understanding where the field stands and the issues we need to tackle.

Chapter 2: Literature Review

Here we take a deep dive into the world of text summarization, focusing on the rich landscape of hybrid search methods. I begin by tracing the evolution of summarization techniques, from

traditional extractive methods that pull sentences directly from the source to more advanced abstractive methods that generate new sentences. A significant part of the discussion revolves around retrieval-augmented generation (RAG) models, which combine the strengths of both retrieval and generation to enhance summarization quality. We analyse several key studies that highlight the pros and cons of existing methods, emphasizing their ability to maintain coherence while delivering concise summaries.

Chapter 3: Research Methodology

In this part I will provide the detailed methodology employed in this study, aiming to explore RAG-based methods for text summarization or other use cases. Here, I adopt a mixed-methods approach, blending qualitative insights with quantitative analysis to achieve a well-rounded understanding of the topic. This research is expected to unfolds three key phases: first, I have focussed on data collection, assembling a diverse set of texts, including articles and reports. Next, I have outlined the development of our algorithms, integrating retrieval techniques with generative processes to create effective summarization strategies. Finally, I have discussed the evaluation methods, combining human assessments with automated metrics to determine the accuracy and effectiveness of the generated summaries. This multi-faceted approach not only provides a robust framework for analysis but also helps us address the complexities inherent in text summarization or generation.

Chapter 4: Analysis

This chapter establishes the foundational framework for exploring hybrid search strategies in text summarization and generation using RAG-based methods. It begins with an in-depth overview of datasets, analyzing their structure, content, and relevance to the research goals. Exploratory Data Analysis (EDA) highlights key patterns, distributions, and potential anomalies. The discussion emphasizes tailoring model configurations to suit dataset-specific requirements, with detailed exploration of models like Gemini and GPT-4omini, integrated into the hybrid search framework to address complex challenges. This chapter underscores the iterative process of combining data analysis, model fine-tuning, and innovative retrieval techniques to enhance outcomes in text summarization and generation.

Chapter 5: Results and Discussions

Here the experiments with Gemini and GPT-4omini models reveal the potential of hybrid search strategies in enhancing text summarization and generation through Retrieval-Augmented Generation (RAG) methods. By balancing sparse and dense vector representations, these approaches significantly improve text accuracy and relevance. A key insight is the importance of weight optimization and adaptive tuning, as the interplay between retrieval methods varies with dataset characteristics and task requirements. Comparative analysis highlights the impact of subtle changes in search techniques and configurations, reinforcing the need for continuous experimentation. The findings affirm the effectiveness of hybrid search with advanced models, contributing valuable insights to refining RAG-based methods for real-world text processing challenges.

Chapter 6: Conclusions and Recommendations

I understand that Future research can enhance hybrid search strategies in several key areas :-

a) **Prompt Design**: Improving the scalability and adaptability of prompts to larger, more complex knowledge bases is crucial. Dynamic and context-aware prompt engineering, incorporating techniques like few-shot learning, chain-of-thought prompting, or contextual embeddings, can make retrieval and generation processes more robust and relevant.

b) **Qdrant Filtering**: Leveraging and optimizing Qdrant's built-in filtering capabilities, such as condition-based keyword filtering, can refine the retrieval process. This would allow for narrowing results to the most relevant contexts while minimizing noise, particularly in extensive and diverse datasets.

c) **Hyperparameter Optimization**: Exploring and fine-tuning Qdrant's configurable parameters, including vector quantization and indexing strategies, could significantly enhance search efficiency and accuracy. Future studies could analyze their impact across different scenarios and employ automated optimization techniques to discover the best configurations.

By addressing these areas, future research can further refine hybrid search workflows, improving precision, relevance, and performance in handling large-scale knowledge repositories.

CHAPTER 2

LITERATURE REVIEW

Literature Review is one of the crucial part for the study completed, as it offers a comprehensive overview of the existing research related to the topic at hand.

The literature review maps out the current state of knowledge on a subject, highlighting significant findings, theories, and gaps. It identifies key works, debates, and trends within the field, providing a foundation upon which new research can be built. By synthesizing and analysing previous studies, a literature review helps to position a new study within the broader scholarly context, illustrating how it contributes to, challenges, or expands upon existing knowledge.

2.1 Introduction

This chapter provides a brief overview of the evolution of text summarization applications, ranging from lightweight models like LSTM to advanced techniques such as Retrieval-Augmented Generation (RAG). It is organized into three main sections. The first section outlines the foundational developments in early text summarization, focusing on how simpler models addressed the challenges of this task. Next we examine the rise of large language models (LLMs) and their significant impact on summarization methods, highlighting the advantages they offer compared to traditional approaches. Finally, the last section discusses recent advancements in the field that hold potential for enhancing current research. It identifies existing gaps that warrant further exploration, setting the stage for future inquiries into effective summarization strategies.

2.2 Origins and Necessity of Text Summarization

Text summarization is about pulling out key parts of a document—like sentences or paragraphs—to create a concise summary. Understanding extractive text summarization involves selecting and

extracting key sentences, phrases, or segments directly from the original text to create a summary. The goal is to piece together a coherent summary using parts of the text itself.

Text mining and text summarization are closely related fields, both focused on extracting useful information from text data, but they approach the task from different angles. Mining involves analysing large volumes of text to uncover patterns, trends, and relationships. It uses techniques like natural language processing to identify key information, categorize text, and extract meaningful insights.

Automatic Text Summarization is a rapidly expanding field designed to save readers time and effort by automatically generating summaries from large volumes of text. In recent years, there have been significant advancements, but challenges remain, prompting extensive research in this area. With the explosion of textual data, interest in ATS has grown considerably, and some of the surveys study thoroughly examines the topic. Researchers have been refining ATS techniques since the 1950s, primarily classifying them into three main types: extractive, abstractive, and hybrid approaches. The extractive method involves selecting key sentences from the original documents to create a summary. In contrast, the abstractive approach generates summaries that may differ from the original text by using an intermediary representation of the input documents. Hybrid methods combine aspects of both extractive and abstractive techniques. Despite various methodologies being suggested, the summaries produced still show noticeable differences from those created by humans. Some studies have done a comprehensive exploration of ATS, addressing its challenges, types, classifications, approaches, applications, methods, implementations, processing and preprocessing techniques, linguistic analysis, datasets, and evaluation measures. It aims to meet the needs of researchers in the field and enhance understanding of this important area of study.(Khan et al., 2023)

Some of the early notable challenges one was the lack of qualified linguistic experts working alongside programmers. This gap can lead to summaries that don't fully capture the subtleties of language. Having a linguistic expert on board could definitely help address this problem. Another challenge is that many summarization systems struggle with grammar, especially when it comes to applying linguistic theories. This could be improved by revising and updating our language theories to better align with how these systems work. However, creating machines that truly understand and reason like humans is still a tough nut to crack. Machines have a hard time grasping

context and nuances in the way people do. On the flip side, when dealing with huge volumes of data, using machines to summarize documents can be a real advantage. Machines can process and shift through information much faster than humans, making it easier and more efficient to get the gist of large amounts of text.(2014 Iranian Conference on Intelligent Systems (ICIS) : 4-6 February 2014 : Bam, Iran, 2014)

Start of text summarizer happened with Punjabi extractive text summarizer developed using an unsupervised machine learning approach. The methodology includes several key steps: tokenizing the text, removing stopwords, creating a similarity matrix, and generating the final summary. Cosine similarity was used here to develop the similarity matrix. Finally a graph was created from the similarity matrix, where the sentences of the input text served as the nodes and the similarity weights represented the edges. The next step involved ranking these nodes using a PageRank algorithm. The sentences were then sorted based on their ranks, and the summarization was generated by selecting the top N sentences, where N is the number of sentences specified by the user. At the end evaluation was done using ROUGE scores, and the results showed that the top three scores were ROUGE-L at 0.56, ROUGE-S at 0.56, and ROUGE-1 at 0.71. (Garg et al., 2021)

Since text categorization along with summarization had some relevance hence there was some find outs with relevance to this. Both of the involve assigning predefined class labels to incoming, unclassified documents. These labels are based on examples from a training set of pre-classified documents. Development was done with an approach for automatic text categorization and summarization that analyses the structure of input text. A text analyser was created that uses a rule reduction technique, which works in three stages: Token Creation, Feature Identification, and Categorization and Summarization. The analyser was tested with sample texts and produced impressive results. Extensive experiments confirmed the effectiveness of our parameter choices and the overall approach for text classification. This work has potential applications in various areas, such as document indexing and retrieval, organizing large web catalogs, automatically extracting metadata, and word sense disambiguation.(Ieee, 2012)

It was time to look forward towards designing a flexible summarization system made up of independent linguistic processing tools that can be easily configured and adjusted. The summarization process would starts with segmenting the text at points where a topic change is likely to occur. These segments would be then classified to identify those that strongly align with

the topic indicated by the user's query. Ultimately, summary sentences are extracted from the most relevant segments instead of the entire text. As a next phase it involved with implementing the first version of the text summarization system. It was required to evaluate its performance and efficiency using established summary evaluation criteria and methods to determine the optimal configurations. (Copeck et al., 1998)

An in-depth study was done on creating and analysing summaries of complex research papers. This is achieved using a powerful sequential model (seq2seq) based on the Transformer architecture. A brief comparison was done with extractive and abstractive summarization approaches to highlight the importance of using abstractive methods for summarizing research papers. Extractive techniques often miss the complexity and context of research articles, leading to a lack of coherence. In contrast, abstractive summarization generates unique, concise, and cohesive summaries by rephrasing and reorganizing information. This shift was crucial for effectively condensing the vast amount of information in research papers while maintaining readability and informativeness. It was soon realised that Hardware and software limitations can restrict the effective training of resource-intensive Transformer models, which need significant computational power. Addressing these issues could improve the models' performance, leading to more accurate summaries of complex scientific research.(Mehta et al., 2024)

There has been an exponential growth in data mainly in terms of product review, lengthy and large number of reviews available. A solution was in need here in order to automatically create a brief summary of those containing the main idea behind it. Abstractive text summarization was used to create summaries by understanding the context and generating new sentences instead of just rephrasing the original text. Here focus was on using a beam search decoder during the inference phase of an encoder-decoder sequence-to-sequence model, along with linear normalization and LSTM. An attention mechanism is employed to enhance processing speed by concentrating on specific parts of the review sentence. The results indicate that summaries produced with the beam search decoder are more accurate than those generated by the commonly used greedy decoder.(Patel and Goswami, 2021)

In the recent past, automatic text summarization (ATS) heavily relied on supervised statistical machine learning, which often resulted in summaries that lacked accuracy and coherence due to its dependence on text features. Apart from this the computational demands and performance of these methods didn't meet modern requirements. A new ATS model which was based on a

Sequence-to-Sequence (Seq2Seq) architecture, enhanced with an attention-based bidirectional Long Short-Term Memory (LSTM). These changes aimed to improve the relevance of the generated summaries to the original text, address issues with out-of-vocabulary (OOV) words, reduce repeated phrases, and minimize cumulative errors in the summaries. After conducting experiments on two public datasets demonstrated that the proposed models outperform baseline methods and several leading models in the field. (Jiang et al., 2021)

Even though the prior models achieved significant performance still they often struggle to capture complex features in long, intricate sentences and the dependencies between sentences. IN order to tackle the issue a stacked Long Short-Term Memory (LSTM) model was proposed to create a more sophisticated feature representation. By stacking layers, we develop a hierarchical structure with attention mechanisms, which enhances the model's ability to process and summarize input text effectively. This approach predicts target sequences for generating text summaries, and results show that it outperforms existing state-of-the-art phrase-based systems on the Gigaword dataset. This framework demonstrated strong performance across various ROUGE scores. (Siddhartha et al., 2021)

There was a moment when attention mechanism became widely accepted in order to enhance the neural machine translation. This led to the comparison of the local and global attention in a Long Short-Term Memory (LSTM) model for generating abstractive text summarization. The Amazon Fine Food Reviews dataset was chosen for experimentation purpose and it was evaluated with Glove, finally where it was realised that the global attention produces better ROUGE-1 scores, indicating it generates more words from the actual summary. In contrast, local attention yields higher ROUGE-2 scores, as it effectively captures more pairs of words by focusing on subsets of input words rather than the entire input. (Hanunggul, n.d.)

Another level of development for an abstractive text summarization model called the multi-layered attentional peephole convolutional LSTM (MAPCoL). This model aimed to automatically generates summaries from long texts, optimizing its parameters using central composite design (CCD) combined with response surface methodology (RSM) to achieve the highest accuracy in summary generation. Evaluation was done with MAPCoL on the CNN/DailyMail dataset and conducted a comparative analysis against state-of-the-art models under various experimental conditions. Finally the results showed hat MAPCoL not only surpasses traditional LSTM-based

models but also excels in semantic coherence in the generated summaries.(Rahman and Siddiqui, 2019)

Since the volume of documents in electric power systems continued to grow, it became increasingly important for managers in this field to quickly analyse these materials and make informed decisions. Text summarization techniques offer a practical solution by efficiently extracting the main points from documents, allowing for faster comprehension and decision-making. This was the time when Hierarchical Bidirectional Long Short-Term Memory (BiLSTM) model was introduced for extractive text summarization specifically designed for electric power systems. his model features a four-layer architecture, including embedding, word, sentence, and classification layers, arranged hierarchically. We conducted experiments using a dataset of over 2,000 electrical papers and compared our approach to existing methods based on Conditional Random Fields, Convolutional Neural Networks, and Recurrent Neural Networks. Finally the results demonstrated that the current model outperforms these existing methods across key performance metrics, including ROUGE-1, ROUGE-2, and ROUGE-L, confirming its effectiveness in summarizing electric power documents.(Jiang et al., 2020)

Overload of data or text with the rise of internet usage among the rising population created the continuous need of keeping on exploring various different ways to bring about a new solution for a better automatic text summarization process. Although this challenge was faced in the various industry, there was also a shortage of high-quality labelled data for text summarization. Researchers tried to solve this issue taking Chinese dataset and create a Chinese text summarization model that would use Long Short-Term Memory networks along with an attention mechanism. The LSTM would capture semantic features, while the attention mechanism enhances contextual understanding. Experimental results proved that this model significantly improves the F1 score compared to other models, effectively would generate summaries and also address the problem of low-quality summarization in certain areas. (Ji et al., 2021)

Literature and texts in Indian regional languages can be challenging to understand because there are often no summaries that effectively convey their main ideas. This issue is compounded by a significant lack of regional datasets, making it difficult for researchers to work in this area. In order to address the shortage of datasets for Indian regional languages like Hindi and Marathi, two deep learning architectures were developed for abstractive text summarization. These models utilize

Attention-based and Stacked LSTM-based Sequence to Sequence neural networks. A list of stop-words and rare words was created specific to Hindi and Marathi for preprocessing. This approach allowed the models to process text in these languages and generate clear, concise summaries that effectively capture the essence of the original content.(Karmakar et al., 2021)

The development of the product or service has highlighted the challenge of understanding diverse customer reviews across various apps and websites. With users providing a wide range of opinions, it had been overwhelming for individuals to sift through all the information. Therefore, implementing a text summarization model could significantly streamline this process, helping users quickly grasp the essential points of reviews without needing to read every detail. Here the idea was to create a text summarizer that automatically generates concise summaries from food reviews using LSTM technology. The process involves breaking down the input sentences and converting them into vectors, allowing for a reduction of lengthy texts while maintaining their original context. Initially the aim was to develop a model that takes food reviews as input and outputs a clear, easily readable summary. Additionally this tool is particularly beneficial for individuals ordering food, as it helps them quickly understand the key points of the reviews they are interested in. (Saraswathi et al., 2022)

Another use case where Abstractive summarization caught importance which involved a deeper comprehension of sentences. Expectation was that this approach generates summaries that may not directly mirror the input text, ultimately improving the reader's understanding of the content. This time a variety of basic methods were used including Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Bayesian Networks. Even Bert, along with a Sequence to Sequence (Seq2Seq) model, using the CNN Daily Mail dataset for evaluation. Here Bert was used for feature extraction, and its results were passed to Seq2Seq for summarization. Evaluation of the models performance was done with the ROGUE score comparing the generated summaries with those created by humans. The results indicated that the accuracy of summaries for shorter articles was not significantly different from that for longer articles.(Widowati et al., 2023)

Another Technology advancement led to examination of how different parameters of Long Short-Term Memory networks affect automated text summarization in Korean. It focuses on three key factors: word embedding size, sentence length, and encoding depth. The current findings revealed that the size of word embeddings significantly influenced the performance, with two-dimensional representations enhancing summarization accuracy. Additionally, increasing sentence length

improved the results, with the best performance achieved using three times the embedding length. However, the encoding depth had a minimal impact, showing only slight improvements with double and triple encodings. It was concluded as optimal performance for Korean text summarization was achieved by combining two-dimensional embeddings with longer sentence lengths and single encoding depth. (Naaz et al., 2024)

There was an increased demand for automatic text summarization for efficient data processing tailored to user needs. Researchers did a thorough review of neural network models for abstractive summarization highlights five critical components in their design: the encoder-decoder architecture, various mechanisms, training strategies and optimization algorithms, dataset selection, and evaluation metrics. Each of these elements plays a vital role in improving the summarization process. Focus was on both single-document and multiple-document summarization techniques, with a special emphasis on empirical methods and extractive approaches. Innovative strategies that address specific complexities of summarization tasks was also highlighted. A significant portion of the study is devoted to the automatic evaluation of summarization systems, underscoring its importance in shaping future research in the field.(Kundan Chaudhari et al., 2024)

2.3 Advancements in Summarization Techniques

After the emergence of smaller models, a new trend developed around large language model (LLM) applications across various natural language processing (NLP) use cases, gaining significant popularity. One notable application of this trend is in the financial domain. Financial narrative summarization was one of the important task to focus which involved created concise summaries within limited words for the various annual reports. Prior to this there was innovation around the DiSum framework which evolved due to the nature of being able to automatically pinpoint important narrative sections in financial reports and measure their significance by quantifying their weighted contributions. Witnessing some of the popular investigation led to the contribution where it was found out that large language models (LLMs) significantly improve the quality of financial report summarization when guided by the DiSum framework. This was the first time LLMs have been applied to the financial narrative summarization (FNS) task, presenting a new approach for enhancing how financial reports are summarized. (Shukla et al., 2023)

Another challenge with LLMs was not being able to effectively process long contexts. Due to this 2 common strategies came into picture. (1) Reducing Input Length: This involves retrieving relevant chunks of information using Retrieval-Augmented Generation (RAG). (2) Expanding Context Window: This aims to increase the amount of text the model can consider at once. But finally both the approaches also had the downsides as well. Reducing input length doesn't guarantee that all necessary information is included, while expanding the context window could make it difficult for the model to focus on the most relevant details needed to complete a task. Addressing to the limitations, it was time to introduce Chain-of-Agents (CoA) as a new concept or framework which utilizes multi-agent collaboration through natural language to enhance information aggregation and context reasoning for long-context tasks across various Large Language Models (LLMs). CoA involves multiple worker agents that communicate sequentially to handle different segments of the text, while a manager agent synthesizes their contributions into a coherent final output. This framework processes the entire input by alternating between reading and reasoning, which helps overcome issues related to focusing on long contexts by assigning each agent a shorter context. After conducting a through evaluation of the CoA with respect to question answering, summarizing etc tasks it was realized that results showed significant improvements of up to 10% compared to strong baseline methods like RAG, Full-Context, and other multi-agent LLMs. (Zhang et al., 2024)

LLMs have been changing as to how the retrieval of information happens by summarizing vast amounts of knowledge through natural language conversations. However focus was mainly on the most common information from their training data while neglecting rarer insights. In biomedical research, recent discoveries are crucial for both academia and industry, but they can be hidden within the growing volume of literature, leading to information overload. Somehow this makes it challenging for LLMs to uncover new connections between biomedical entities, such as drugs, genes, and diseases, as they struggle to capture the less common knowledge within the field. Addressing the above challenge RAG proposed to enhance prompts with context from external datasets. However, RAG methods often miss a significant amount of relevant information because they tend to focus on clusters of overrepresented concepts in biomedical literature. A new information-retrieval method was introduced that utilizes a knowledge graph to downsample these clusters, helping to alleviate the issue of information overload. This method demonstrates approximately double the retrieval performance compared to traditional embedding similarity

approaches in terms of both precision and recall. Additionally it was showing that combining embedding similarity and knowledge graph retrieval into a hybrid model can outperform both methods individually, potentially improving biomedical question-answering systems.(Delile et al., 2024)

In the healthcare domain clinicians and researchers are facing an overwhelming increase in information, including literature and clinical notes. Text summarization is a vital task that aims to extract key information from this complexity and present it in a more concise format. By utilizing automatic text summarization, healthcare professionals can efficiently access important information, helping them avoid the pitfalls of information overload. There were a total of 5 pretrained language models compared for single document summarization. In case of multi document summarization various models like BART, Pegasus, PRIMERA was evaluated. Their performance was assessed using ROUGE scores on the MEDQA-AnS dataset, which contains 156 examples. Interestingly, while TextRank achieved higher ROUGE scores than most generative models, this does not necessarily mean it performed better overall. ROUGE scores measure the overlap between generated content and reference summaries, which can be misleading. Overall, while TextRank scores high in ROUGE metrics, its extractive nature may not always reflect superior summarization quality compared to generative models.(Yang and Chew, n.d.)

2.4 Recent Innovations: The Role of Retrieval-Augmented Generation (RAG)

Moving towards the latest advancements we have been watching the concept if RAG Retrieval-Augmented Generation being implemented giving us good results. enhance the capabilities of Large Language Models (LLMs) for creating effective Q&A systems. It works by combining these models with a private set of documents to generate accurate answers. However, as the number of documents grows, keeping the system accurate becomes more challenging. This is because the performance of RAG heavily depends on how well the system's "retrievers" can pull out the most relevant documents. The retrievers need to effectively find and provide the right context to the LLM, which is crucial for maintaining accuracy in the responses. Some of the recent papers have discussed the approach of Blended RAG combines different semantic search techniques, such as Dense Vector indexes and Sparse Encoder indexes, with hybrid query strategies. Latest research shows research shows that this approach improves retrieval results and establishes new

benchmarks for information retrieval which is tested on some of the sample datasets as available. Most of the current methods used in Retrieval-Augmented Generation (RAG) rely on keyword searches and similarity-based techniques, which can limit the overall accuracy of these systems. Figuring out the best search method for RAG is still a developing field. Recent studies have put a step forward towards improving the accuracy of retrieval and RAG systems by using Semantic Search-Based Retrievers and Hybrid Search Queries. Evaluating different search techniques using three main types of indices became a key part of the research. These included BM25 for keyword-based searches, KNN for vector-based searches, and Elastic Learned Sparse Encoder (ELSER) for sparse encoder-based semantic searches. The Blended RAG pipeline has proved to be highly effective across various datasets, even though it wasn't specifically trained on them. This method stands out because it doesn’t need example prompts for fine-tuning, which are usually necessary in few-shot learning. This indicates that Blended RAG can generalize well in zero-shot scenarios. The overall accuracy of the RAG pipeline improved significantly, with the Blended RAG approach setting new records. This shows that Blended RAG outperforms previous methods even with smaller language models. The findings highlight that focusing on advanced retrieval techniques can greatly enhance RAG systems, paving the way for smarter and more context-aware Generative Q&A systems.(Sawarkar et al., 2024)

RAG was introduced as a method common limitations of large language models (LLMs), such as providing outdated information and the tendency to create inaccurate or "hallucinated" content. Still assessing the performance of RAG systems has its own set of challenges as well. Some of the current benchmarks only focusses on question-answering tasks or applications and at the same time overlooks the larger scope of scenarios where RAG could be useful. The influence of retrieval components and external knowledge construction has been completely neglected. RAG already incorporates external knowledge which is capable of generating responses which can improve the accuracy and realism. In order to tackle the challenges researchers focussed on 3 key approaches. Pre-Retrieval Processing which involves refining data to standardize text, ensure accuracy, optimize index structures, adjust block sizes, and improve query formulation. Retrieval Model Optimization which step includes fine-tuning models specific to certain domains and using dynamic embedding techniques to enhance performance. Finally Post-Retrieval Processing where after the retrieving information process, this approach aims to shorten context length by reranking and compressing data. The goal is to highlight key information, reduce noise, and improve how

the generator integrates and uses the retrieved content. CRUD actions was helpful here that describe interactions between users and knowledge bases, and also categorize the range of RAG applications into four distinct types—Create, Read, Update, and Delete. Developing different datasets according to various scenarios performance of the RAG was evaluated, Additionally effect of different components of RAG as retriever, context length, knowledge base construction and LLM was analysed.(Lyu et al., 2024)

Next it was time to analyse for query focussed task. It was realised that RAG failed when questions was asked on a global level from an entire corpus of text. Some of the prior methodologies struggle to scaled when it was needed to handle the large amounts of text that are usually indexed by typical RAG systems. There was a need to harness the strengths of different existing methods, Graph RAG was a new approach proposed for question answering task on private text corpus. This approach is designed to scale effectively with both the variety of user questions and the volume of source text being indexed. The method utilizes a large language model (LLM) to create a graph-based text index in two steps: first, it builds an entity knowledge graph from the source documents, and then it generates community summaries for groups of closely related entities. When a question is posed, each community summary helps produce a partial response, which is then combined into a final answer for the user. For a specific type of global sensemaking questions, especially with datasets containing around one million tokens, the Graph RAG approach shows significant improvements over a basic RAG model, both in terms of the comprehensiveness and diversity of the answers generated.(Edge et al., 2024)

From the legal domain side the professionals were in a need to write up some analysis that refer to past situations i.e. previous case decisions. Some of the Intelligence systems designed to assist them in drafting these documents can be extremely beneficial, but creating such systems is still challenging. Those systems must be able to effectively help users locate, summarize, and reason over key precedents to be truly useful. In order to support these tasks it was needed to collaborate with legal professionals to transform a large open-source legal corpus into a dataset. This dataset will support two essential tasks: information retrieval (IR) and retrieval-augmented generation (RAG). CLERC dataset was designed to train and evaluate models on their ability to (1) find relevant citations for a given piece of legal analysis and (2) compile these citations, along with the necessary context, into a coherent analysis that supports reasoning. State-of-the-art models was benchmarked using CLERC and it was realised that current approaches still face challenges. For

example, while GPT-4o generates analyses with the highest ROUGE F-scores, it also tends to hallucinate the most. Additionally, zero-shot information retrieval models only achieve a recall rate of 48.3% at 1000. Current high end models even Llama also faces difficulty meeting citation metrics. However, we demonstrate that scoring high on these metrics doesn't necessarily mean the generated content is of good quality.(Hou et al., 2024)

At a certain point it was realised that Retrieval-Augmented Generation (RAG) improves Large Language Models (LLMs) by retrieving relevant memories from an external database. However, current RAG methods often store all memories in a single database, which can limit the model's focus on the most important memories and introduce unnecessary noise. Finally there was a introduction to multiple partition paradigm for Retrieval-Augmented Generation (RAG), called M-RAG, where each database partition functions as a basic unit for RAG execution. Building on this approach, it was proposed as a novel framework that uses Large Language Models (LLMs) combined with Multi-Agent Reinforcement Learning to optimize various language generation tasks. Through extensive experiments across seven datasets and three different language generation tasks, it was found that M-RAG consistently outperforms several baseline methods. Specifically, it achieved improvements of 11% for text summarization, 8% for machine translation, and 12% for dialogue generation.(Wang et al., 2024)

Researchers again wanted to focus on medical domain with biomedical texts. There was another framework introduced which was called as RAG-RLRCLaySum. This aims to make complex biomedical research more understandable for laypeople using advanced Natural Language Processing (NLP) techniques. The RAG solution enhanced by a reranking method, draws on multiple knowledge sources to ensure that lay summaries are accurate and relevant. Additionally, our Reinforcement Learning for Readability Control (RLRC) strategy enhances readability, making scientific content accessible to non-specialists. When evaluations was done no publicly available datasets like eLife, it demonstrated that current methods outperform the Plain Gemini model. This resulted in achieving 20% increase in readability scores, a 15% improvement in ROUGE-2 relevance scores, and a 10% enhancement in factual accuracy. The RAG-RLRC-LaySum framework effectively democratizes scientific knowledge, fostering greater public engagement with biomedical discoveries.(Ji et al., 2024)

2.5 Discussion

In reviewing the existing literature on text summarization, several notable gaps emerge, particularly concerning hybrid search techniques :-

- ✓ **Exploration of Hybrid Search Techniques**: While the integration of various search strategies is acknowledged, there remains insufficient exploration of how different hybrid search techniques can enhance summarization outcomes. Specifically, the comparative effectiveness of dense search, sparse search, and text filter search needs further investigation. Understanding how these approaches can be synergistically combined may yield significant improvements in the relevance and coherence of generated summaries.
- ✓ **Impact of Reranking Methods**: Another gap lies in the application of reranking strategies within the summarization process. While reranking has been shown to refine results in other contexts, its potential benefits in text summarization are not well-documented. Future research could assess whether incorporating reranking techniques can enhance the quality of the summaries produced by hybrid search models.
- ✓ **Advanced RAG Strategies**: The literature also lacks comprehensive studies on advanced Retrieval-Augmented Generation (RAG) strategies. Although RAG has gained traction, its implementation in conjunction with various hybrid search methods remains largely unexplored. Investigating how these advanced strategies can be utilized to optimize summarization performance presents an exciting opportunity for future research.
- ✓ **Limited Contextual Understanding**: While many models have demonstrated impressive performance in generating summaries, there is a noticeable lack of focus on contextual understanding. Existing studies often emphasize surface-level extraction without adequately addressing how nuanced meanings can be preserved. Future research could explore methods that enhance models' ability to interpret and summarize context more effectively.

2.6 Summary

The literature review conducted for this research highlights the ongoing advancements in text summarization and generation techniques. As smarter and more efficient methodologies emerge, the potential for these innovations to positively impact various fields becomes increasingly

evident. From healthcare and education to finance and legal sectors, improved text summarization can enhance information accessibility and decision-making processes.

Moreover, since the integration of technologies like Retrieval-Augmented Generation (RAG) showcases a promising direction for the future. These developments not only would facilitate the extraction of relevant information from vast datasets but also improve the quality of generated summaries, making them more coherent and contextually accurate.

Researchers would continue to explore and refine these techniques, the evolution of text summarization or text generation would likely lead to more user-friendly applications, allowing individuals and organizations to navigate and digest information more effectively. Ultimately, this progress would contribute significantly to our ability to harness knowledge, fostering greater understanding and innovation across diverse domains.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In today's innovation driven world, the sheer volume of text generated across all the possible domains ---ranging from news articles, social media to research articles – it poses a significant challenge for individuals and organizations alike. At this influx of data being continued to keep growing over period of time, the ability to efficiently distill essential information becomes increasingly crucial. Text summarization, which aims to automatically generate concise summaries that capture the main ideas of larger texts, has emerged as a vital solution to this problem.

My current research would focus on exploring hybrid search methods for text summarization, specifically leveraging Retrieval-Augmented Generation approaches. By combining traditional retrieval techniques with advanced generative models, I aim to enhance the quality and relevance of generated summaries. RAG-based methods hold promise for improving summarization accuracy by utilizing external knowledge sources to provide context, thereby addressing the limitations of conventional summarization models.

I would evaluate the effectiveness of various RAG-based methods through comprehensive accuracy measurements, utilizing established metrics such as ROUGE scores to assess the performance of the generated summaries. Through a systematic approach, including dataset selection, preprocessing, model selection, and hybrid search implementation, this research aims to contribute valuable insights into the evolving landscape of text summarization techniques. Ultimately, our goal is to develop a robust summarization framework that not only improves the efficiency of information processing but also enhances the user experience by delivering clear and coherent summaries.

3.2 Research Methodology

3.2.1 Dataset Selection

For this research, the datasets selected were: CNN/Daily Mail, RAG Mini Wikipedia, and RAG Mini BioASQ, all sourced from Hugging Face. This selection aligns with established guidelines and ensures a comprehensive approach to text summarization. Figure 3.2.1.1 shows the overall structure of the CNN DailyMail Dataset which is little different as compared to the RAG datasets Figure 3.2.1.2 and 3.2.1.3 for Rag Mini Wiki and Rag mini Bioasq respectively.. By incorporating datasets from diverse domains, I aimed to minimize potential bias and enhance the robustness of the evaluation.

Additionally, implementing a Retrieval-Augmented Generation (RAG) solution necessitates an understanding of how effectively it can retrieve accurate information from a corpus, even when faced with multiple documents. This is particularly important for ensuring that the model can efficiently find the right answers in response to specific queries. Overall, the chosen datasets are instrumental in evaluating the performance of the RAG-based methods in various contexts.

```
DatasetDict({
    train: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 287113
    })
    validation: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 13368
    })
    test: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 11490
    })
})
```

Figure 3.2.1.1 CNN Daily Mail Dataset Format

Dataset link :- https://huggingface.co/datasets/abisee/cnn_dailymail

```
DatasetDict({
    passages: Dataset({
        features: ['passage', 'id'],
        num_rows: 3200
    })
})
<class 'datasets.dataset_dict.DatasetDict'>
```

```
DatasetDict({
    test: Dataset({
        features: ['question', 'answer', 'id'],
        num_rows: 918
    })
})
```

Figure 3.2.1.2 RAG Mini Wikipedia Dataset Format

Dataset link :- https://huggingface.co/datasets/rag-datasets/rag-mini-wikipedia

```
DatasetDict({
    test: Dataset({
        features: ['question', 'answer', 'id'],
        num_rows: 918
    })
})
```

```
Available configurations: ['text-corpus', 'question-answer-passages']
DatasetDict({
    passages: Dataset({
        features: ['passage', 'id'],
        num_rows: 40221
    })
})
```

Figure 3.2.1.3 RAG mini Bioasq Dataset Format

Dataset link :- https://huggingface.co/datasets/rag-datasets/rag-mini-bioasq

3.2.2 Dataset Preprocessing

Here we have opted not to not to apply traditional data preprocessing techniques, before inputting it into our models. This decision stems from our recognition of the capabilities of large language models (LLMs), which are inherently designed to process raw, unstructured data effectively.

LLMs possess advanced mechanisms for understanding context, handling noise, and identifying relevant patterns within the data. By allowing the models to work with the original dataset, we preserve its natural structure and complexity, which can provide richer insights during the summarization process.

3.2.3 Model Selection

Here I have selected the GPT-4o-mini and Gemini 1.5 Flash models for experimentation due to their complementary strengths in handling text generation and retrieval tasks.

GPT-4o-mini is described as a fast and affordable model for very focussed tasks. It is considered to be ideal fine tuning tasks and the output of this model can be compared to be similar to that of the bigger models performance making it low cost and low latency. It has a higher context window of 128k tokens and max output token to be as 16.3k. This made it ideal for iterative testing and integration within the hybrid search framework.

On the other hand Gemini 1.5 flash model is considered to be as lightweight, optimized for fast performance and efficiency. This model is considered to be very useful in case of processing queries or for real-time or large-scale search tasks. The model architecture is aligned with rapid information retrieval and contextual understanding, which aligned well with the need for efficient and accurate summarization in the hybrid search system.

This strategic selection ensured a comprehensive evaluation of hybrid search performance across different scenarios, balancing accuracy, efficiency, and scalability.

3.2.4 Hybrid Search Implementation

Under the Hybrid Search Implementation I am utilizing a combination of dense vector search, sparse search, and text filter search methods. This multi-faceted approach allows us to leverage the strengths of each method, enhancing the overall effectiveness and accuracy of our text summarization process.

a) **Dense Vector Search**: This method enables us to capture semantic similarities between texts by representing them as high-dimensional vectors. By utilizing dense embeddings, we can retrieve relevant documents based on their contextual meaning rather than just keyword matching. This is mainly beneficial as it helps us identify and extract information that may not be explicitly stated but is contextually relevant.

b) **Sparse Search**: In contrast, Sparse Search method focus on retrieving documents based on exact term matches, which can be crucial for identifying specific pieces of information. This technique ensures that we do not overlook important keywords or phrases that may play a significant role in the summarization. By combining sparse search with dense vector search, we create a robust retrieval system that captures both nuanced meanings and precise content.

c) **Text Filter Search**: This method adds an additional layer of refinement by allowing us to apply filters based on specific criteria, such as date, relevance, or topic as per the condition used internally. Text filter search enhances the retrieval process by ensuring that the documents selected for summarization align closely with our research objectives. This targeted approach not only improves the quality of the summaries generated but also helps us maintain focus on the most pertinent information.

Here we will be using Qdrant vector database to store the embeddings of the corpus and apply the above searches. Internally it applied cosine similarity as a similarity search mechanism. When we apply the text filter internally we actually apply rules in the form of must have certain value/Should have a certain value/must not have certain value, multiple combination of these kind of clauses is possible to apply. Finally the hybrid search will be responsible to retrieve the correct document based on the user query asked from the knowledge base. One of the best time to apply a similar RAG pipeline like we have done here is when we have a series of enterprise level documents and also the supporting infrastructure in order to improve the information retrieval and answer construction. Here one of the objective of implementation a basic level of RAG pipeline is also to understand if the answer is also bring able to generate from the correct document. Our RAG pipeline should be able to extract the answer from the right document source.

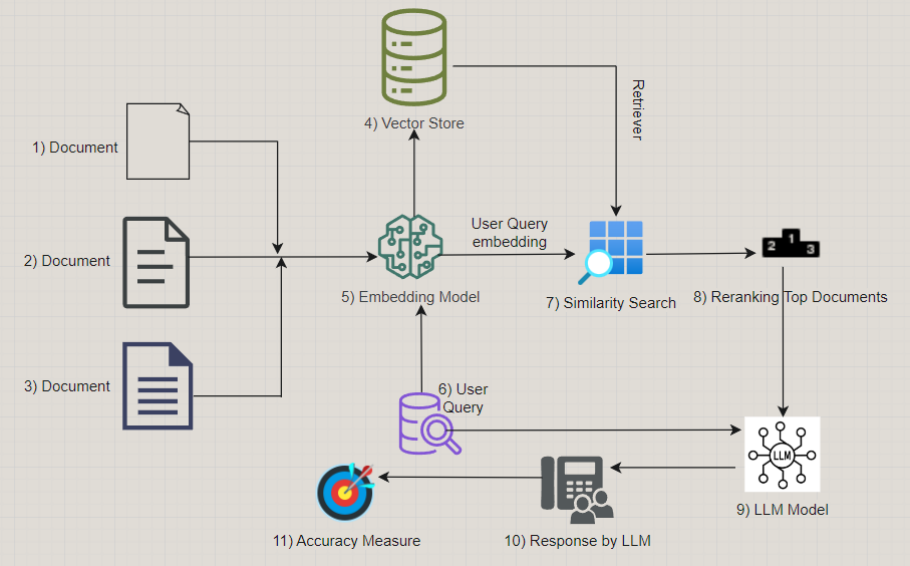


Figure 3.2.4.1 Architecture Design

3.2.5 Experimental Setup

In our experimental setup, As per the above figure 3.2.4.1 we conduct some experiments to evaluate the effectiveness of our hybrid search methods with two different models for text summarization/generation.

1) **Experiment 1: Dense and Sparse Search Methods**

In the first case we are using only Dense and Sparse search methods which will allows us to assess the individual contributions of each method in retrieving relevant documents. This combination provides a solid foundation for understanding how these methods work together to enhance summarization quality.

This experiment also serves as a baseline, helping us identify the strengths and weaknesses of the two methods in isolation. It allows us to analyse how well they perform in tandem, particularly in terms of retrieval accuracy and the quality of the resulting summaries.

2) **Experiment 2: All Search Methods**

Here we are integrating all the search methods(dense, sparse) where this comprehensive approach will aim to maximize the retrieval potential by leveraging the strengths of each method. I can achieve a more nuanced retrieval process that captures both the semantic depth of the text and the precision of specific terms.

I expect to see improved performance in the quality and relevance of the summaries generated. The inclusion of text filter search allows for additional refinement, ensuring that only the most pertinent documents are selected for summarization.

Once we have the relevant document retrieved by the vector database the user query will be converted into a prompt which will be source of input to the LLM. Here I am using Gemini and GPT 40-mini LLM which will generate the appropriate response based on the user input. I have chosen this model since this is specifically designed for instruction-following tasks, making it particularly adept at understanding and responding to user inputs in a coherent and contextually appropriate manner.

3.2.6 Accuracy Measurement

Finally I need to evaluate the accuracy and effectiveness of our text summarization and response generation, I can utilize several key metrics as ROGUE which is considered to the be the most common tool for assessment. Some of the other commonly used variants include ROUGE-N, which evaluates n-gram overlap, and ROUGE-L, which assesses the longest common subsequence, providing insights into both content coverage and fluency. I can also use some of the other evaluation metrics to gain a comprehensive understanding of our model's performance if tested on different tasks other than summarization:

a) **BLEU Score**: This was originally developed for machine translation, BLEU measures the overlap of n-grams between generated text and reference text, helping us assess the precision of the generated summaries.

b) **Uptrain**: This is a combination of NLP models and LLMs to do LLM evaluations. To ensure reliable scores, it was built with a dedicated pipeline for each of the pre-defined checks which is much more complex than just prompting the LLM to act as an evaluator.

c) **RAGAS**: This is a set of evaluation metrics that can be used to measure the performance of your LLM application. These metrics are designed to help you objectively measure the performance of your application. Metrics are available for different applications and tasks, such as RAG and Agentic workflows.

3.2.7 Results Analysis

Now I have a detailed examination of the outcomes from the above experiments, focusing on the effectiveness of the hybrid search methods and the quality of the generated summaries. After analysing the performance metrics—such as ROUGE, BLEU, RAGAS and Uptrain—we assess the impact of integrating text filter search on the overall summarization quality. Additionally, I would also be able to analyse the precision and recall metrics to understand the trade-offs between capturing relevant content and minimizing extraneous information.

Beyond the quantitative metrics, I would also like to delve delve into qualitative analyses based on human evaluations. This includes assessing the coherence, fluency, and overall relevance of the generated summaries. Further more I may observe that certain types of documents yield better summaries with specific search methods, indicating a need for tailored approaches based on content type. This analysis helps us refine our models and suggests areas for further research.

3.2.8 Summary

Depending on the overall result that I will get from the experiments conducted I will be able to conclude as to which would be a better way togo ahead. Additionally I would also like to suggest some of the other research experiments which can be conducted in future and was not possible due to time constraint.

CHAPTER 4

ANALYSIS

4.1 Introduction

In this chapter, we delve into the analytical processes that underpin our exploration of hybrid search techniques for text summarization using RAG-based methods. The effectiveness of these methods heavily relies on the quality and structure of the data employed. To this end, we first provide a detailed description of the datasets selected for our study: the CNN Daily dataset, RAG Mini BioASQ, and RAG Mini Wikipedia.

The subsequent sections outline our data preparation strategies, which include the elimination of irrelevant variables, transformation as maybe needed. Each of these steps is critical in ensuring that the data is not only clean but also conducive to robust analysis. Following the preparation phase, we conduct exploratory data analysis (EDA) to uncover insights through methods employed here as checking on the number of records and types of fields. Analyse the types of content and their lengths. We can also look at word clouds or bar charts of the most common words or phrases in the context or question and answers. Additionally comparing statistics and distributions of the dataset of the content/question/answer length against similar dataset in order to identify any features or any limitations.

Data visualization techniques will further enhance our understanding of the underlying patterns within the datasets. Through this comprehensive analysis, we aim to lay a solid foundation for evaluating the accuracy of RAG-based summarization methods, ultimately contributing to the advancement of hybrid search techniques in text summarization.

4.2 Dataset Description

The CNN Daily Mail is a very widely known benchmark for training and evaluation text summarization models. It mainly comprises of various news articles over 300K written by many journalists from CNN and Daily Mail. This dataset has a diverse range of topics and writing styles.

The RAG Mini BioASQ dataset is specifically designed for tasks involving biomedical question answering and text summarization. This dataset is a subset of the larger BioASQ challenge, which focuses on understanding and summarizing biomedical literature.

The RAG Mini Wikipedia dataset is designed for enhancing question-answering and summarization tasks using retrieval-augmented generation (RAG) methods. This dataset comprises a subset of Wikipedia articles, making it a rich resource for training models to generate coherent and contextually relevant responses. It covers articles with a wide array of subjects—from history to science—this dataset offers a broad context for evaluating summarization and retrieval capabilities across different domains. The retrieval-based approach allows models to leverage relevant context from the articles, enhancing their ability to generate accurate summaries and answers.

4.3 Data Preparation

Given that we sourced our datasets directly from the Hugging Face repository, we found that extensive data cleaning and transformation were mostly unnecessary. These datasets are well-structured and maintained, allowing us to focus more on exploring their content. However, we did undertake several steps to perform exploratory data analysis (EDA), which included comparing the datasets to uncover any relevant statistical insights. This comparative analysis not only helped us understand the characteristics of each dataset but also highlighted how they might interact in the context of our research on hybrid search methods for text summarization.

4.3.1 Basic Statistics

The CNN/Daily Mail dataset is organized into three distinct subsets: training, validation, and test sets, each designed to facilitate comprehensive evaluation of summarization models.

1. **Training Set**: Comprising 287,113 articles, this substantial dataset serves as the foundation for training our models, providing a rich source of diverse content.
2. **Validation Set**: With 13,368 entries, the validation set allows us to fine-tune model parameters and evaluate performance during the training process.
3. **Test Set**: Finally, the test set includes 11,490 articles, enabling a robust assessment of the model's summarization capabilities on unseen data.

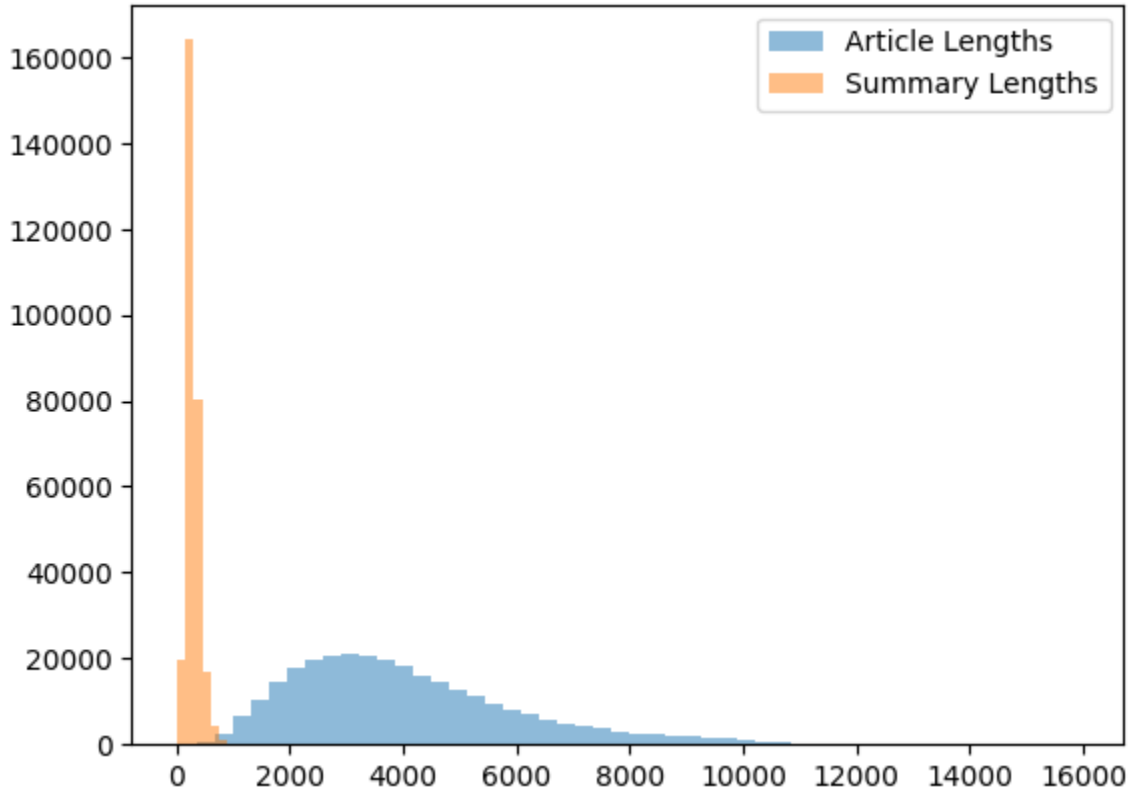


Figure 4.3.1 Overall length of the dataset

This structured format ensures that we can effectively train, validate, and test our approaches to hybrid search for text summarization. The dataset is structured with three key features that enhance its usability for summarization tasks: Article:- contains the full text of the news articles, which provides the primary content for summarization, Highlights :- we find the corresponding summaries of the articles, serving as the reference outputs for evaluating the model's performance, ID :- Each entry includes a unique identifier, which facilitates tracking and referencing specific articles throughout the analysis. From the training part of the dataset when we go ahead to check the length of the articles with a bin size of 50 the length varies from 0 upto 20K. Based on the

word cloud generated from the CNN Daily dataset, several key themes and topics emerge that can provide insights into the overarching narratives present in the articles. The most prominent words, such as “scroll video,” “United States,” “police officer,” and “New York,” suggest a focus on current events, particularly those that have significant social and political implications. Above figure 4.3.1 shows the overall distribution of Article lengths vs Summary lengths.

The phrase “police officer” alongside “New York” indicates a likely discussion around law enforcement issues, perhaps touching on police conduct or recent incidents that have garnered public attention. The inclusion of terms like “first time” and “last week” suggests that these discussions are tied to recent developments, emphasizing the dataset's relevance to ongoing societal debates.

While going ahead to implement topic modelling it reveals a compelling mix of themes that are both pressing and diverse, highlighting the multifaceted nature of the discussions within the dataset. Words like “country,” “war,” “state,” “attack,” and “military” suggest that a significant portion of the content revolves around geopolitical conflicts and issues of governance. This points to narratives that might delve into national security, military actions, and the implications of warfare on both a local and global scale. The focus on these terms indicates a societal preoccupation with stability and the consequences of conflict. Additionally the presence of “hospital,” “health,” “patient,” and “water” suggests that public health is another crucial theme. This could relate to discussions about healthcare systems, the impact of conflict on medical facilities, or the ongoing challenges posed by health crises. In order to look at the most repeated words below word cloud figure 4.3.2 shows it.

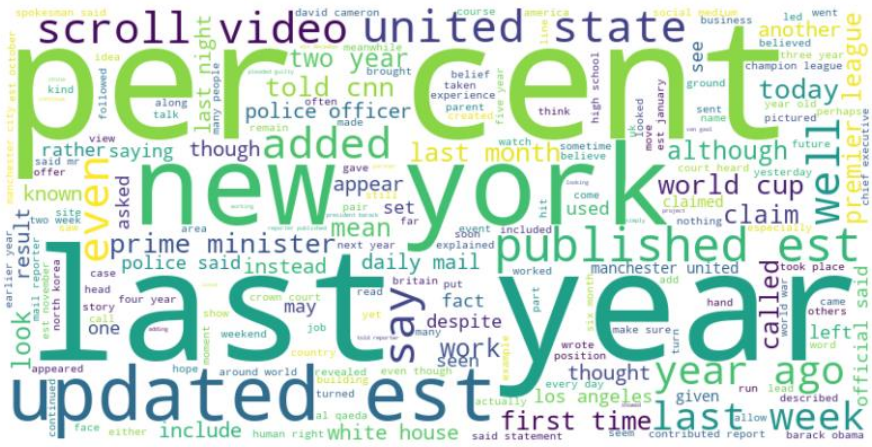


Figure 4.3.2 Word Cloud

Coming to the next one the **RAG Mini Wikipedia** dataset is composed of two distinct yet equally fascinating components that enrich our analysis. The text corpus and the question-answer pairs, each structured to support various tasks in text summarization and retrieval.

Text Corpus: This component consists of a carefully curated collection of passages, featuring a total of 3,200 entries. Each passage provides insightful and informative content, perfectly structured with two key features:

a) Passage: This feature contains the intriguing text drawn from Wikipedia, offering a wide variety of knowledge for our summarization and retrieval tasks.
b) ID: Each passage is accompanied by a unique identifier, which serves as a reliable reference point, making it easy to track and organize the data throughout our analysis.

Question-Answer Format: In addition to the rich text corpus, the dataset includes a dedicated question-answer component, comprising 918 entries. This format is designed to facilitate engaging interactions between questions and their corresponding answers:

a) Question: Each entry contains a question related to the passages, designed to evaluate the model's ability to retrieve relevant information.
b) Answer: This feature provides the precise answer, ensuring a clear and effective response to each question.
c) ID: A unique identifier for each question-answer pair, facilitating easy reference.

On creating a word cloud on the Passages in the dataset it reveals a great array of keywords that hint at several interconnected themes. Presence of names like “Roosevelt,” “Wilson,” and “Ford” suggests that a significant portion of the dataset is rooted in historical discussions. These figures represent pivotal moments in U.S. history and politics, pointing to topics related to leadership, governance, and the evolution of political parties. The inclusion of “first” indicates an exploration of groundbreaking events or individuals who shaped the course of history. Additionally terms like “Tesla,” “Newton,” and “elephant” (likely referring to the symbolism of the Republican Party) suggest that the dataset also addresses significant scientific advancements and contributions. The juxtaposition of scientific figures with historical ones might indicate a narrative that explores how these individuals influenced both science and society.

Exploring the dataset with the help of a topic modelling it reveals a captivating mix of themes, with the highlighted words suggesting various interconnected narratives. Terms such as “leopard,” “wolf,” “bear,” and “polar penguin” hint at discussions surrounding wildlife, possibly touching on

conservation efforts or the symbolism of these animals in political discourse. These animals are often associated with different traits, which might be used metaphorically to describe political parties or leaders. This interplay between wildlife and politics could provide a unique lens through which to examine societal attitudes and values. Again presence of words like “Ford,” “president,” “Lincoln,” “Cleveland,” and “Republican” suggests a strong focus on political history and figures in the United States. This indicates that the dataset likely delves into discussions about presidential administrations, key legislative changes, and the evolution of political parties over time. The mention of both historical (Lincoln, Cleveland) and more contemporary figures (Ford) illustrates the continuity and shifts within American politics. Overall statistics of the Rag Mini data is depicted below with figure 4.3.3 and the Distribution of the overall document lengths is depicted in figure 4.3.4

```
Statistics for Wikipedia:
Number of samples: 3200
Average length of documents: 12980.17
Minimum length of documents: 3
Maximum length of documents: 33420
```

Figure 4.3.3 Statistics(RAG Mini Wiki)

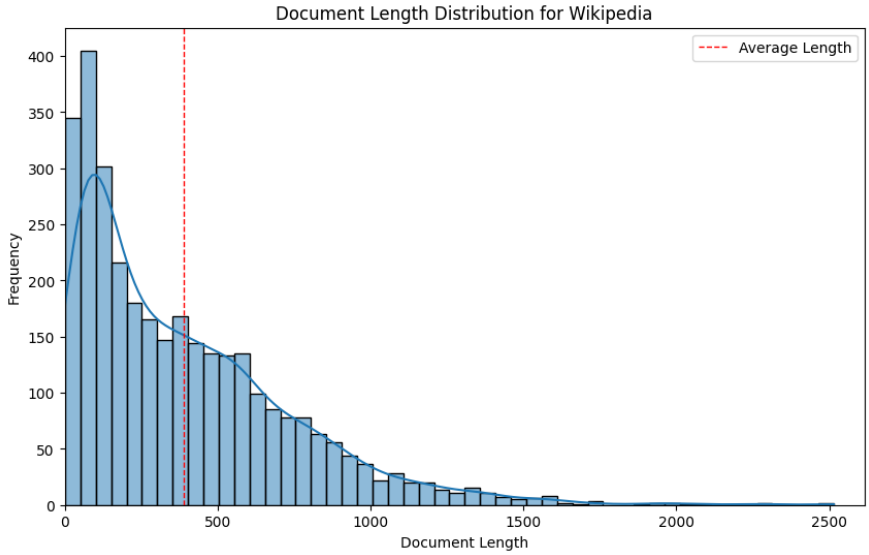


Figure 4.3.4 Document Length(RAG Mini Wiki)

On the next dataset which is **RAG mini Bioasq** is one which is supported by the National Library of Medicine of the National Institutes of Health. This dataset was created intending to support research in automatic question answering and information retrieval. It focuses on the biomedical field, making it suitable for applications like clinical decision support, literature review, and biomedical research assistance.

As per the dataset configuration it is done as per below :-

Text Corpus: This contains a collection of passages that can be used to provide context for answering questions. It includes a total of 40,221 passages with unique identifiers.

Question-Answer Passages: This configuration includes 4,719 question-answer pairs, where each entry comprises:

a) Question: The query prepared by the owners.
b) Answer: The corresponding answer to the question.
c) Relevant Passage IDs: A list of IDs indicating which passages are relevant for answering the question.
d) ID: A unique identifier for the question-answer pair.

The words revealed by the word cloud on this dataset completely shows that the data has themes and concepts that are central to discussions in biomedical research. Some of the words “protein,” “gene,” and “expression” indicate a strong emphasis on molecular biology and genetics. This suggests that the dataset likely includes studies exploring the roles of specific proteins and genes in various biological processes. These terms highlight the importance of understanding genetic factors and their influence on health and disease. Again terms like “treatment,” “patient,” and “effect” suggest that the dataset is also concerned with clinical outcomes and the implications of research for patient care. This implies that the articles may focus on how specific treatments can lead to measurable effects in patients, emphasizing the practical application of scientific findings in medical contexts.

Applying topic modelling reveals a range of interconnected themes that are central to contemporary biomedical research. Presence of the term “DNA” indicates a strong focus on genetic research. This suggests that the dataset likely includes discussions around the genetic basis of diseases, the role of DNA in health and disorders, and how genetic information can inform treatment strategies. Some of the points here somehow show scientific discoveries translate into clinical settings, particularly in diagnosing and treating various conditions. The focus on patient

outcomes underscores the dataset's relevance to healthcare. It also highlights a strong emphasis on understanding and addressing specific health challenges. This suggests that the dataset covers a range of medical conditions, exploring their pathophysiology, risk factors, and potential treatment options. Below figure 4.3.5 depicts the Statistics for the Rag Mini Bioasq dataset and figure 4.3.6 shows the overall distribution of Document lengths within the dataset.

```
Statistics for BioASQ:
Number of samples: 40221
Average length of documents: 1032.71
Minimum length of documents: 3
Maximum length of documents: 33420
```

Figure 4.3.5 Statistics(RAG Mini Bioasq)

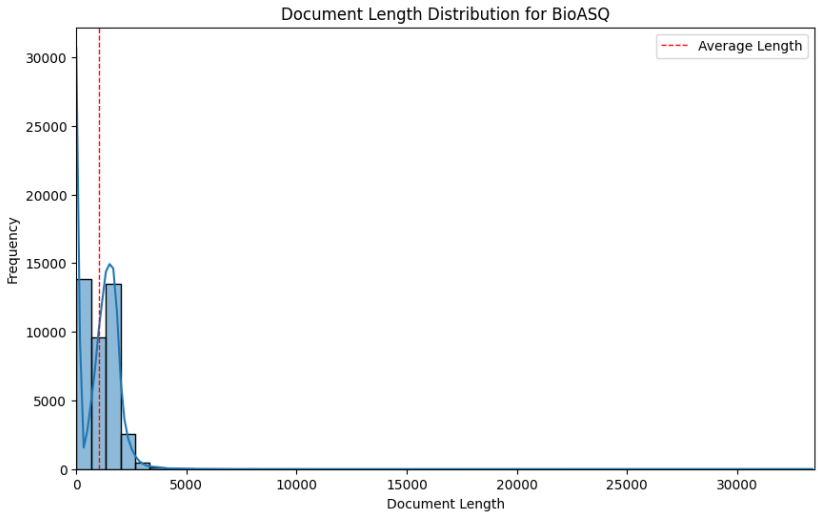


Figure 4.3.6 Document length(RAG Mini Bioasq)

4.4 Model Hyperparameters

In this case I have used Gpt-4o-mini and Gemini-1.5 Flash as 2 large language models for implementation and evaluation purpose for creating the hybrid search on textual data. Although both the models bring their unique strengths and weakness on the table, they still offer diverse perspectives on the generating the best content in terms of accuracy and concise as per the defined situation. Below are the detailed the hyperparameters used in along with the models to optimize their performance for Retrieval-Augmented Generation (RAG)-based methods.

1. **GPT-4o-mini**

GPT-4o-mini is a lightweight version of OpenAI's GPT-4 series designed for resource-constrained applications. Despite its smaller size, it retains the robust linguistic capabilities and reasoning skills characteristic of the GPT-4 architecture.

Some of the Key hyperparameters used for this study include : -

1. **Model Size**: Optimized for faster inference with fewer parameters compared to the full GPT-4 model.
2. **Temperature**: Set to 0.9 to balance creativity and coherence in text generation/summarization tasks.
3. **Top-p (Nucleus Sampling)**: Configured at 0.9 to focus on the most probable word choices while still allowing some diversity.
4. **Max Tokens**: Limited to 100 to generate concise and contextually relevant summaries.
5. **Context Window Size**: 128K tokens and 65.5K as max token output which seemed adequate for capturing extensive input text and maintaining consistency in text generation/summarization.

This model was very effective in terms of understanding the nuanced prompts and generating text with highest linguistic quality making it a very reliable choice.

2. Gemini-1.5 Flash

This language model is a fast and versatile multimodal for scaling across diverse tasks. It can handle various input and output formats, including text, images, audio, and video. This allows it to perform tasks like image and video captioning, data extraction from documents and tables, and more.

Some of the Key hyperparameters configured include :-

1. **Temperature**: Set to 1.0 to allow a higher degree of variability and richness in the generated summaries, suitable for capturing diverse contextual information.

2. **Top-p**: Maintained at 0.95 to ensure coherent yet creative outputs by focusing on a broader range of word probabilities.
3. **Top-k Sampling**: Configured with 64 to include the most relevant tokens while minimizing randomness.
4. **Max Output Tokens**: Set at 8912 to accommodate complex summarization tasks without truncation of critical information.

4.5 Model Implementation

4.5.1 Model Selection and Integration

Here the rationale behind choosing the Gpt-4o-mini model was because of its light weight architecture still offering computation efficiency which does not sacrifice on the language understanding capabilities. However the Gemini-1.5 flash model was because it was already tested on large scale generative tasks, mainly because of its ability to handle complex contextual information.

4.5.2 Prompt Design

Discussing on the structure of the prompts used for the models here, I have used Instructional Prompts in order to guide the LLM model to retrieve the required context effectively. Specific prompt templates ensures the consistency across text generation/summarization tasks. Since here the main task was to find out the required answer to a specific question asked hence the prompt was designed accordingly.

4.5.3 Hybrid Search Implementation

First I had created the dense and sparse vectors on that basis hybrid search was implemented. One of the best way to implement hybrid search was to leverage the strengths of dense and sparse retrieval techniques to fetch the most relevant context for downstream tasks. Sparse methods are computationally efficient and perform well on queries with exact or near-exact term matches. Again dense methods can match semantically similar phrases, even if the wording differs significantly.

Above diagram 3.2.4.1 refers the overall implementation process where we are creating the embedding and storing in vector database. Then we are retrieving the relevant documents with similarity search operation and reranking the top K documents are fed to the LLM, where after entering the right prompt we are querying or generating the answer with the LLM model.

4.6 Summary

This chapter uncovers the foundational elements that underpin this research, establishing a robust framework for exploring hybrid search strategies in text summarization or generation using RAG-based methods. Here the discussion began with a comprehensive overview of the datasets, detailing their format, content, and relevance to the research objectives. The exploratory data analysis (EDA) offered critical insights into the statistical properties of the datasets, highlighting patterns, distributions, and anomalies if any.

In terms of exploration of the models, their hyperparameters, and implementation processes underscored the importance of tailoring configurations to align with the unique characteristics of the dataset. The models selected for experimentation, including Gemini and GPT-4omini, were meticulously integrated into the hybrid search framework, leveraging their individual strengths to address the complexities. This chapter not only provided the analytical foundation for the research but also emphasized the iterative and adaptive nature of designing effective hybrid search strategies. These findings reaffirm the importance of a well-rounded approach, combining thorough data exploration, model fine-tuning, and innovative retrieval techniques to achieve improved outcomes in text summarization or text generation.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter aims to present and critically analyse the outcomes of applying Retrieval-Augmented Generation (RAG) methods for text summarization or text generation use cases. The primary objective of this study was to evaluate the accuracy and effectiveness of hybrid search techniques in generating text that retain key information while ensuring brevity. Mainly I focussed on if the hybrid search methods as compared to the traditional methods or approaches to understand how good we get the accuracy and how much we are also able to maintain the consistency. I am also able to conclude on how the RAG based models perform across diverse datasets. In order to evaluate the accuracy the evaluation metrics I used here was ROGUE score, BLUE Score, Uptrain and RAGAS. These findings are crucial for advancing the field of text summarization, especially in scenarios where accuracy and contextual relevance are paramount.

5.2 Findings and Discussion

When I tried to experiment without the RAG approach on either of the datasets rag-mini-bioasq or rag-mini-wikipedia with gpt-4o-mini, I realised that the score was not consistent or was very low as per the below figure 5.2.1 which explains various metrics. Below is the table for reference with gpt-4o-mini experiment.

ROGUE-1	ROGUE-2	ROGUE-L	BLEU Score
8.6%	4.54%	8.69%	0.86%

Figure 5.2.1 GPT Scores before using a RAG approach

Finally when I used gemini-1.5-flash-latest model on either of the dataset below was the overall performance scores with reference to figure 5.2.2.

Below is the table for reference with Gemini llm model experiment.

ROGUE-1	ROGUE-2	ROGUE-L	BLEU Score
19.67%	3.38%	19.67%	11.31%

Figure 5.2.2 Gemini Scores before using a RAG approach

Above are the scores without a RAG approach, where I realized that we might need a RAG approach if it could provide better accuracy, enhanced contextual understanding, and improved retrieval mechanisms to achieve more precise and informative summaries. The relatively low ROUGE-2 and BLEU scores indicate that the traditional methods may lack the depth in capturing relationships between words and long-range dependencies, which are crucial for generating high-quality text.

However when I changed my approach I started creating the embeddings for the complete context with text-embedding-3-small model from Openai. I have chosen this embedding model due to its balanced performance and efficiency. Its performance on the Massive Text Embedding Benchmark (MTEB) was average, offering a reliable baseline without compromising on retrieval quality as compared to the other models available. I inserted the embeddings to the Qdrant vector database in the form of Collection with the help of Points creation method since the similarity search methods would happen with respect to any text on the basis of points internally. I have also created dense and sparse vectors out of the complete context so that I can try to implement hybrid search for information retrieval and check the overall performance. Since the approach combines the strengths of two different types of vector representations hence I would expect better retrieving of relevant information.

	Scores			
	Bleu Score	**ROGUE-1**	**ROGUE-2**	**ROGUE-L**
Gemini	91.20%	96.96%	90.63%	96.96%
GPT 4o-mini	65.15%	90.14%	75.36%	90.14%

Figure 5.2.3 Scores after using a RAG approach

	Uptrain		RAGAS				
	Faithfulness	**Answer_correctness**	**score_context_relevance**	**score_factual_accuracy**	**score_response_completeness_wrt_context**	**score_response_completeness**	**score_response_conciseness**
Gemini	1	89.06%	1	1	1	1	1
GPT 4o-mini	1	100%	1	1	1	1	1

Figure 5.2.4 Scores after using a RAG approach

Above BLEU and ROGUE scores referred in figure 5.2.3 we have got by experimenting with different models like Gemini and GPT-4o-mini. Additionally we also tried to use the RAGAS library and uptrain metrics to understand how better the answer was relevant to the question which is explained by the figure 5.2.4.

After applying the RAG approach where we have created hybrid search with the help of dense and sparse vectors the overall accuracy has increased to a big extent. However there is also some of the hyper parameters involved here score_threshold to be a minimum of 50% and other parameters were within the function SearchParams(exact=True, hnsw_ef=128). Because of the use of these hyper parameters we are getting very high scores.

Considering the above Bleu scores we can clearly see that Gemini model achieved higher score of 91.20% which could be possibly due to its strong ability to generate text that closely aligns with the question or query asked. This suggests that the Gemini model excels in producing highly accurate and contextually relevant outputs, making it well-suited for tasks that require precise information retrieval. GPT-4o-mini model score could indicate a comparatively lower alignment with the reference summaries.

However if we consider the answer_correctness parameter from the uptrain then we can clearly see the GPT 4o-mini scored highest strong reliability in generating factually correct and precise information, making it highly effective for tasks where accuracy and correctness are critical as compared to Gemini. But here I would still prefer to go ahead with Gemini since even though it lags little behind in some of the scores still its faster in terms of inference.

5.3 Interpretation of Results

When we ran the Gemini or GPT model overall the questions from the dataset we realised that the answers generated were from the overall context the model was trained on. Hence the result was if the question was related to a specific fact or incident happened at a particular time the answer collected all the incidents happened at that particular time. So quite part of the answer generated was not required or out of context. The above hyper parameters additionally helped me to make sure that the search happens very specific to the question as asked or searches for the answer to be very specific to the contextual meaning of the sentence. The overall results are very impressive where we had used various accuracy metric to measure the model performance. Without the hyperparameters we were still able to see a drop in the overall score by 20-30%.

5.4 Summary

The results and discussions of the experiments conducted using Gemini and GPT-4 0minin models shed light on the evolving dynamics of hybrid search strategies in the domain of text summarization or text generation. By leveraging Retrieval-Augmented Generation (RAG) methods, this study demonstrated that hybrid search approaches, when designed with a careful balance of sparse and dense vector representations, can significantly enhance the accuracy and relevance of generated text.

A key takeaway from the experiments is the pivotal role of weight optimization in hybrid search. The interplay between sparse and dense retrieval, influenced by dataset characteristics and task-specific requirements, underscores the need for adaptive and fine-tuned approaches. Additionally, the comparative analysis highlighted how even subtle variations in search techniques and model configurations could lead to significant differences in performance, emphasizing the importance of continual experimentation and evaluation in this field.

Ultimately, the findings of this study affirm that hybrid search methodologies, when paired with advanced models like Gemini and GPT-4 0minin, hold immense potential for improving text generation accuracy. These results contribute valuable knowledge to the ongoing exploration of RAG-based methods and pave the way for future research aimed at further refining hybrid search strategies to address real-world challenges in text processing and summarization.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter concludes the study by putting out the key findings and insights gained from the exploration of hybrid search strategies for text summarization using Retrieval-Augmented Generation (RAG) methods. The research aimed to investigate the potential of combining sparse, dense and other retrieval techniques to enhance the accuracy and relevance of text summarization outcomes.

In this chapter, the primary conclusions are drawn from the analysis and experimentation presented, highlighting the performance of the models and techniques utilized. The chapter also provides recommendations for future research, offering guidance on potential areas of improvement and further exploration in hybrid search methodologies. By reflecting on the research objectives and outcomes, this chapter aims to underscore the significance of hybrid search in advancing the field of text summarization and to chart a path for future innovations in this domain.

6.2 Conclusion

After carrying out all the experiments the evaluation results reveal a good difference in performance between both the models used as indicated by various metric scores. It seems like Gemini model is decently good at producing high quality accurate information which is relevant to the output. The notable performance gap could be due to the models superiority in generating accurate texts, which seems likely due to the enhanced retrieval capabilities and better contextual adaption. Still GPT 4o-mini may still be valuable in other scenarios where priority could be in generating more diverse or creative texts rather then generating text based on some context matching. We can also conclude stating that the result somehow underscore the importance of

models based on the task requirements – Gemini could be favoured for context specific text generation or summarization however GPT 4o-mini could be great for tasks that expect demand flexibility and generalization.

In the same way if I consider some additional tasks like the storytelling, poetry writing or fictional content creation or if we are brainstorming for unique ideas for any action GPT 4o-mini would be very great choice. Additionally if in case a particular task will span across various fields or domains which would require flexibility across diverse topics GPT 4o-mini would be able to take care of these situations. Also situations with very limited computational resources the smaller version of this model would be able to offer a cost effective solution.

In addition to the above specific implementations of GPT model, I also understand that Gemini 1.5 flash model has its own advantages. Its strength mainly lies in the retrieving and adapting to contextual information making it ideal for extracting key insights. Mainly in case of legal or medical summaries and generating high quality technical reports or documentation it would be ideal to go for Gemini model. Additionally in case of data extraction, annotation tasks or generating FAQ or articles from structured sources it would be a good choice to go ahead with Gemini model.

6.3 Contributions to knowledge

This research highlights several key contributions to the various applications of text summarization or text generation through the exploration and evaluation of hybrid search techniques using RAG-based methods. The experiments were designed to systematically assess the effectiveness of different retrieval strategies and their impact on summarization accuracy. Since keyword based or lexical search specifically relies on the matching exact word by word or phrases that appear in a query with those in the documents. It has its limitations as it may not be able to handle misspellings, synonyms, or polysemy. Additionally if this search strategy has pick up 5-7 sentences from the knowledge base and the answer lies in one specific line then there are changes it might not be able to pick up the right sentence to device the answer. Also alone semantic similarity search is a techniques to analyse the meaning of words and their relationships. Although it can handle can handle misspellings, synonyms, and polysemy, however it also has its own limitations as it requires large amount of data to train which is computationally expensive and time consuming as well.

Hence this cannot be an effective strategy for short documents. This is why I chose to develop an optimized combination of these approaches to form a more effective hybrid search strategy. This integration leverages the precision of keyword-based retrieval and the contextual depth of semantic search, resulting in a more robust and adaptable system for retrieving relevant information across diverse or complex context. The major contributions are outlined below :-

1. Baseline Evaluation without Vector Database :-

   As a basic foundational step, experiments were conducted by directly querying the GPT-4 0mini and Gemini model without integrating a vector database. This approach provided a baseline to assess how well these models could generate summaries using only their inherent knowledge, without access to external context or any retrieval mechanisms. This experiment highlighted the limitations of relying solely on model memory for generating accurate and contextually rich summaries, especially when dealing with domain-specific or complex queries.

2. Implementation of Vector Database with Hybrid Search :-

   To enhance retrieval efficiency, a vector database was implemented using Qdrant, where embeddings of context were generated and stored. Hybrid search was applied by combining both sparse (keyword-based) and dense (semantic-based) vectors to retrieve relevant information for generation. Additionally, experiments were conducted with and without applying Qdrant's hyperparameters in order to evaluate their effect on improving retrieval precision and overall performance. Initially the approach without the hyper parameters increased the overall relevance and accuracy but additionally when I applied the hyper parameters as discussed above that lead to the jump in relevance and accuracy higher by around 20%.

6.4 Future Recommendations

One significant area for future exploration lies in enhancing the prompt design used in this research. While the current prompt have been designed to effectively guide the models in the context of the datasets and tasks at hand, their scalability and adaptability to larger and more complex knowledge bases remain a critical avenue for improvement.

As knowledge bases grow in size and diversity, the ability of prompts to efficiently guide the retrieval and generation processes becomes an important aspect. Future research could focus on developing more dynamic and context-aware prompt engineering techniques, incorporating mechanisms to adapt the prompts based on the size, structure, and content of the underlying knowledge base. This may involve leveraging advanced techniques such as few-shot learning, chain-of-thought prompting, or contextual embeddings to ensure that the search process remains robust and yields accurate, relevant results even in more demanding scenarios.

Another promising area for future research is the utilization and further enhancement of the built-in filtering functionality provided by the Qdrant vector database. This functionality allows the implementation of condition-based filters, enabling the search process to retrieve contexts based on specific conditions, such as the presence or absence of particular keywords. By leveraging this capability, it becomes possible to refine the retrieval process further, narrowing down the results to the most relevant contexts from extensive knowledge bases. For instance, filtering by specific keywords can help prioritize information aligned with the query, while excluding irrelevant terms ensures that noise is minimized in the retrieved data. This feature can significantly enhance the precision and relevance of hybrid search strategies, especially when dealing with large and complex datasets.

Future work could focus on systematically integrating and optimizing these filter functions within hybrid search workflows. Additionally, experimenting with advanced filtering criteria and combining them with prompt engineering techniques may provide even more powerful tools for extracting highly contextual and relevant information from massive knowledge repositories.

One more potential avenue for future improvement is the exploration and optimization of additional hyperparameters provided by the Qdrant vector database. Alongside the existing hyperparameters used in this research, Qdrant offers a range of configurable parameters that can influence search performance, such as vector quantization settings, and indexing strategies. Fine-tuning these parameters, in conjunction with the hybrid search implementation, can lead to significant gains in efficiency and retrieval accuracy.

Future studies could focus on systematically evaluating the impact of these hyperparameters in various scenarios, identifying configurations that best align with the characteristics of the dataset and task requirements. This could involve conducting sensitivity analyses to understand how adjustments to these settings affect performance and exploring automated approaches such as hyperparameter optimization frameworks to discover optimal values.

REFERENCES

Anon (2014) *2014 Iranian Conference on Intelligent Systems (ICIS) : 4-6 February 2014 : Bam, Iran*. Institute of Electrical and Electronics Engineers.

Copeck, T., Szpakowicz, S., Barker, K., Chali, Y. and Matwin, S., (1998) *The Design of a Configurable Text Summarization System The Design of a Conngurable Text Summarization System*. [online] Available at: https://www.researchgate.net/publication/2448918.

Delile, J., Mukherjee, S., Van Pamel, A. and Zhukov, L., (2024) Graph-Based Retriever Captures the Long Tail of Biomedical Knowledge. [online] Available at: http://arxiv.org/abs/2402.12352.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S. and Larson, J., (2024) From Local to Global: A Graph RAG Approach to Query-Focused Summarization. [online] Available at: http://arxiv.org/abs/2404.16130.

Garg, K.D., Khullar, V. and Agarwal, A.K., (2021) Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization. In: *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*. Institute of Electrical and Electronics Engineers Inc., pp.750–754.

Hanunggul, P.M., (n.d.) *The Impact of Local Attention in LSTM for Abstractive Text Summarization*.

Hou, A.B., Weller, O., Qin, G., Yang, E., Lawrie, D., Holzenberger, N., Blair-Stanek, A. and Van Durme, B., (2024) CLERC: A Dataset for Legal Case Retrieval and Retrieval-Augmented Analysis Generation. [online] Available at: http://arxiv.org/abs/2406.17186.

Ieee, (2012) *2012 International Conference on Advances in Engineering, Science and Management.* IEEE.

Ji, M., Fu, R., Xing, T. and Yin, F., (2021) Research on Text Summarization Generation Based on LSTM and Attention Mechanism. In: *Proceedings - 2021 International Conference on Information Science, Parallel and Distributed Systems, ISPDS 2021*. Institute of Electrical and Electronics Engineers Inc., pp.214–217.

Ji, Y., Li, Z., Meng, R., Sivarajkumar, S., Wang, Y., Yu, Z., Ji, H., Han, Y., Zeng, H. and He, D., (2024) RAG-RLRC-LaySum at BioLaySumm: Integrating Retrieval-Augmented Generation and Readability Control for Layman Summarization of Biomedical Texts. [online] Available at: http://arxiv.org/abs/2405.13179.

Jiang, J., Zhang, H., Dai, C., Zhao, Q., Feng, H., Ji, Z. and Ganchev, I., (2021) Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization. *IEEE Access*, 9, pp.123660–123671.

Jiang, W., Zou, Y., Zhao, T., Zhang, Q. and Ma, Y., (2020) A hierarchical bidirectional LSTM sequence model for extractive text summarization in electric power systems. In: *Proceedings - 2020 13th International Symposium on Computational Intelligence and Design, ISCID 2020*. Institute of Electrical and Electronics Engineers Inc., pp.290–294.

Karmakar, R., Nirantar, K., Kurunkar, P., Hiremath, P. and Chaudhari, D., (2021) Indian Regional Language Abstractive Text Summarization using Attention-based LSTM Neural Network. In: *2021 International Conference on Intelligent Technologies, CONIT 2021*. Institute of Electrical and Electronics Engineers Inc.

Khan, B., Shah, Z.A., Usman, M., Khan, I. and Niazi, B., (2023) Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey. *IEEE Access*, 11, pp.109819–109840.

Kundan Chaudhari, Raj Mahale, Fardeen Khan, Shradha Gaikwad and Vita Jadhav, (2024) Comprehensive Survey of Abstractive Text Summarization Techniques. *International Research Journal on Advanced Engineering and Management (IRJAEM)*, 607, pp.2217–2231.

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T. and Chen, E., (2024) CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models. [online] Available at: http://arxiv.org/abs/2401.17043.

Mehta, R., Mehta, N., Purohit, V., Saha, I. and Mishra, P., (2024) Text Summarization for Research Papers using Transformers. In: *2024 IEEE 9th International Conference for Convergence in Technology, I2CT 2024*. Institute of Electrical and Electronics Engineers Inc.

Naaz, R., Kavitha, R. and Yadav, S., (2024) Exploring the Impact of LSTM Parameters on Network Performance for Automatic Text Summarization. In: *2024 International Conference on Optimization Computing and Wireless Communication, ICOCWC 2024*. Institute of Electrical and Electronics Engineers Inc.

Patel, R.M. and Goswami, A.J., (2021) Abstractive Text Summarization with LSTM using Beam Search Inference Phase Decoder and Attention Mechanism. In: *ICCISc 2021 - 2021 International Conference on Communication, Control and Information Sciences, Proceedings*. Institute of Electrical and Electronics Engineers Inc.

Rahman, M.M. and Siddiqui, F.H., (2019) An optimized abstractive text summarization model using peephole convolutional LSTM. *Symmetry*, 1110.

Saraswathi, R.V., Chunchu, R.V., Kunchala, S., Varun, M., Begari, T. and Bodduru, S., (2022) A LSTM based Deep Learning Model for Text Summarization. In: *6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp.1063–1068.

Sawarkar, K., Mangal, A. and Solanki, S.R., (2024) Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. [online] Available at: http://arxiv.org/abs/2404.07220.

Shukla, N.K., Katikeri, R., Raja, M., Sivam, G., Yadav, S., Vaid, A. and Prabhakararao, S., (2023) Generative AI Approach to Distributed Summarization of Financial Narratives. In: *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023*. Institute of Electrical and Electronics Engineers Inc., pp.2872–2876.

Siddhartha, I., Zhan, H. and Sheng, V.S., (2021) Abstractive Text Summarization via Stacked LSTM*. In: *Proceedings - 2021 International Conference on Computational Science and Computational Intelligence, CSCI 2021*. Institute of Electrical and Electronics Engineers Inc., pp.437–442.

Wang, Z., Teo, S.X., Ouyang, J., Xu, Y. and Shi, W., (2024) M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions. [online] Available at: http://arxiv.org/abs/2405.16420.

Widowati, K.G., Budiman, N., Foejiono, K. and Purwandari, K., (2023) Abstractive Text Summarization Using BERT for Feature Extraction and Seq2Seq Model for Summary Generation. In: *Proceedings: ICMERALDA 2023 - International Conference on Modeling and E-Information Research, Artificial Learning and Digital Applications*. Institute of Electrical and Electronics Engineers Inc., pp.226–230.

Yang, R. and Chew, E.Y., (n.d.) Ascle: A Python Natural Language Processing Toolkit for Medical Text Generation (Preprint). [online] Available at: https://www.researchgate.net/publication/380681160.
Zhang, Y., Sun, R., Chen, Y., Pfister, T., Zhang, R. and Arik, S.Ö., (2024) Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. [online] Available at: http://arxiv.org/abs/2406.02818.

EXPLORING HYBRID SEARCH FOR TEXT SUMMARIZATION: A STUDY OF RAG BASED METHODS WITH ACCURACY EVALUATION

SUSHANT SUR

Research Proposal

AUGUST 2024

Abstract

Hybrid search is well known ensemble technique which combines different search techniques in order to give better results. It might be a combination of keyword based search technique, semantic search technique etc. For example if we search with keyword places to travel in Thailand then keyword based search may only give out result based on the keywords as places and Thailand, it may miss out some of the important tourist attractions. Also here alone semantic search will understand the intent to find all the places in Thailand for tourists.

Here hybrid search combines multiple approaches. Keyword search prioritizes exact matches, ensuring precision, while semantic search expands results by understanding related concepts and synonyms. This kind of comprehensive approach is ideal for RAG applications. This study aims to explore the efficacy of the hybrid search methods for text summarization or other use cases, focusing on the Retrieval-Augmented Generation (RAG) based techniques. Text summarization being a crucial task in the filed of natural language processing where some of the recent advancements have successful introduced hybrid search methods like RAG to enhance accuracy.

Even though there has been significant progress, existing text summarization techniques often seem to struggle with maintaining accuracy, leading in to further exploration of hybrid approaches. This research would imply a comparative analysis of RAG based methods in order to demonstrate superior accuracy in text summarization tasks when compared to the conventional techniques. The results suggest that hybrid search techniques can significantly improve the performance of the text summarization systems offering practical benefits for various applications. The study underscores the potential of RAG-based methods in advancing text summarization, paving the way for further research in hybrid approaches.

TABLE OF CONTENTS

Abstract	67
1. Background	70
2. Problem Statement OR Related Research OR Related Work	72
3. Research Questions (If any)	75
4. Aim and Objectives	75
5. Significance of the Study	76
6. Scope of the Study	77
7. Research Methodology	78
8. Requirements Resources	82
9. Research Plan	83
References	84

1. Background

Retrieval-augmented generation (RAG) is concept which is helpful to address various knowledge intensive tasks. It has the power to cover some of the limitations of the LLM's thereby improves the capability of the LLM models by leveraging external knowledge sources. Some of the common limitations that RAG is able to put an end is missing information and chances of creation misleading or inaccurate content. Even though evaluation of the RAG systems has been challenging in the past still there are some of the benchmarks available based on some of the specific use cases around.

The motivation behind this study stems from growing demand for quick, efficient and accurate text generation or summarization methods in the era of information getting overloaded. With huge amount of data being generated everyday from various field or domains from newspapers, magazines, blogs or articles the ability to condense and distil the information into concise and meaningful information is very crucial. So of the earlier techniques, were effective to some extent, but often they struggle to capture the full context and nuance of diverse datasets. Here some of the limitations are pretty evident when facing complex documents from diverse backgrounds where simple single strategy search has been leading to inconsistent or biased summaries.
The emergence of hybrid search techniques within the RAG framework have been offering solutions that is quite promising to the current challenges. However by combining search strategies and reranking methods hybrid approaches have the potential to significantly improve the retrieval process, leading to more accurate and contextual summaries. Additionally the advancement of power language models currently available can provide an unprecedented opportunity to push the boundaries of what better can be achieved.

Some of primary challenges lie in the traditional methods where the text summarization used to happen with wither extractive or abstractive in isolation. Here extractive text summarization used to produce summaries which are disjoint or lack coherence since they just pull the sentences from the original text. Other the other hand abstractive methods aims to generate new sentences based on an input which struggles with maintaining factual accuracy and preserving the original meaning of the context. Another challenge seems to be in the ability to handle diverse datasets from various backgrounds and large scale datasets.

In this paper we will be mainly focusing on the text summarization as use case. Text summarization mainly helps to extract the essence of any large corpus of document which would otherwise take huge amount of time to go through pages and pages, finally leading into enabling quick insights and decision making. There are mainly two types of text summarization mainly Extractive and Abstractive. Extractive method involves creating a concise overview of the document by identifying, extracting and compressing essential information for efficient comprehension. Text summarization is helpful to us in a variety of ways from document summarization, blog content, news articles, market research reports to social media content. Some researchers have commented as hybrid approaches which combines the elements of both extractive and abstractive methods, Despite various recommended methodologies, the generated summaries still exhibit noticeable differences compared to those created by humans (Khan et al., 2023). Some of the challenges highlighted here are controlling the output, not being able to identify linguistic features, Summarization evaluation metrics, information being overloaded, Falase information getting generated, and Multi-document summarization. Proposing on the evaluation methods it was understood that the core of the measure is covered by Latent Semantic Analysis which could capture the main topics of the documents. Results show a high correlation between Human rankings and LSA based evaluation measure. The measure was designed to compare a summary with its full text (Steinberger and Ježek, 2009).

The primary objective of the research is mainly to explore and evaluate various hybrid search strategies with the RAG framework for text summarization or other use cases. The goal is ultimately to identify approaches that :-

a) Improve the relevance and accuracy of the generated text or summaries.

b) Preserve the context and coherence of the content

c) Enhance the overall efficiency and scalability of the summarization process.

This study aims to contribute to the field of Natural language processing by providing better insights into the effectiveness of the search strategies with RAG, ultimately leading to the uncover practical application of text processing technologies. The main purpose of the study is to try and test out various hybrid approaches within the framework of Retrieval-Augmented Generation

(RAG) along with reranking if they are able to give us better result with respect to text summarization or other use cases. Considering the volume of digitization has been continuously growing, the need for efficient and accurate summarization methods has always been increasingly important. Traditional techniques has been falling short due to some or other challenges mainly in terms of complexity or diversity in modern datasets. By implementing simple search methods has not been able to enhance the retrieval and summarization process. This specific research will focus on experimenting with different hybrid search strategies, optimizing their performance through weight assignments, and rigorously evaluating the performance using state of the art algorithms like LLAMA, GPT4 etc. The final goal would be to identify and develop advanced methods that significantly improve the quality and accuracy of the text summarization or other use cases contributing to valuable insights to the field of Natural Language Processing and Information retrieval.

2. Problem Statement OR Related Research OR Related Work

Summarization is considered to be a complex task with a series of sub tasks, where each of the sub tasks could affect the potential in order to generate good quality summaries. In the past papers researchers had proposed LSA based approaches which can be applied to multiple document summarization. Some of the challenges discussed here is when this is applied to multiple documents there could be a scenario where two documents might have text about the same event/topic hence there is a challenge to resolve the redundancy. Additionally the other challenge is LSA would expect long sentences since they would contain important terms than a shorter sentence (Ježek and Katedra, n.d.)

.

Some of the common problems faced in case of extractive text summarization was that finding out the position of the sentence and the concurrency of the words in the sentence. Another problem which came back was extracting the information out of the context (Widyassari et al., 2022). Traditional challenges were more related to content selection, coherence and informativeness.

These problem would impact researchers who would be dependent on accurate and concise summaries on the past research papers in order stay updated with the latest advancements. It would

additionally impact journalists and content creators since they would need to quickly extract key information from large corpus of documents on a daily basis. If in case the problem is not solved then users will get overloaded with information which would make it difficult to extract relevant information. Manually doing the same task would not be an appropriate way and would reduce overall productivity. Since there is exponential growth of digital information and increase in demand of access to relevant information it is important to have method where we are easily being able to extract accurate summaries from wide corpus of data. The problem of text summarization is more prevalent in digital platforms, educational institutions and research communities mainly. Fixing the problem to have a better and accurate text summarization method would lead to enhanced decision making, save time and help with continuous learning.

In the initial papers it was also discussed that there might be a close relation between the text mining and text summarization. According to difference in requirements summary with respect to input text, established summarization systems should be created and classified based on the type of input text. Finally, the most fundamental proposed evaluation methods are considered (Ieee, n.d.). Some of the past papers on Punjabi Extractive text summarizer was developed using unsupervised machine learning techniques. Here the methodology consisted of tokenization of punjabi text, removal of stop words, generation of similarity matrix, ranking based on similarity matrix finally summary was proposed. It included various forms of ROGUE score to measure the evaluations (Garg et al., 2021). The past papers have also suggested use of Fuzzy Inference System where summary of the document is created based upon the level of the importance of the sentences in the document (Babar and Patil, 2015). Researchers have also proposed a method based on multi-agent particle swarm optimization approach is proposed to improve the extractive text summarization (2014 Iranian Conference on Intelligent Systems (ICIS) : 4-6 February 2014 : Bam, Iran, 2014).

Some of the researchers have also discussed text categorization and summarization approach in order to analyze the input text. A text analyzer was developed to derive the structure of the input text sing rule reduction technique in 3 stages namely token creation, feature identification and Categorization and Summarization. This analyzer was tested with sample input texts which gave noteworthy results. Extensive experimentation validates the selection of parameters and the efficacy of the approach for text classification (Ieee, 2012). In the recent past

papers we see that there has been some improvement in the text summarization experiments where we started implementing models like Bert to both the types of summarization. This model was able to capture the meaning of a document and create numerical representations of the sentence(Liu and Lapata, 2019). Here researchers overcame the limitation of the initial phase of Bert model which could only have positional embeddings for max length of upto 512, by adding additional position embeddings. Finally there was advancement where researchers have been adding multiple hybrid search strategies in order to get the better context out of the source document (Sawarkar et al., 2024). Some of the additional challenges highlighted in this paper are that there are no standard datasets available on which Retriever and RAG benchmarks are currently available. Retriever might have been considered as a separate problem to be treated in specific domains while RAG is more considered in LLM domain.

Some papers have addressed the issue where most of the researchers have only focused on the evaluation of the LLM components in the RAG pipeline in their experiments but have completely neglected the influence of the retrieval components and external knowledge database construction (Lyu et al., 2024). It is also discussed that RAG fails on global questions directed at an entire text corpus when there is a query focused summarization task. In order to combine the strengths of the contracting methods Graph RAG Approach is proposed to question answering over private text corpora which scales with generality of user questions and quantity of source text to be indexed (Edge et al., 2024). But in case there are many documents with large amount of information extractive text summarization would arise as NP completed problem, in order to solve these metaheuristic algorithms were used Ježek and Katedra (n.d.). When we talk about automatic text summarization for specific domains not having domain specific dataset could be a problem where it was proposed for fully automating the fine tuning of the text summarization model in a specific domain without involvement of human annotators(Avramelou et al., 2023). Some prior investigations revealed that LLMs when left to their own devices, often struggle to effectively summarize financial documents. Here it was discussed that when LLMs were guided with the DiSum framework exhibited substantial improvement in the quality of financial report summarization (Shukla et al., 2023).

With the recent advancements in RAG models have shown quite promising improvement in text summarization by taking the help of external knowledge source and retrieval mechanisms.

However the real potential of hybrid strategies with RAG framework still remains unexplored. Hybrid search strategies which combine various search techniques and assigning weights to it, could potentially enhance the performance of the RAG models.

3. Research Questions (If any)

The following research questions are suggested for each of the research objective as highlighted as follows.

1. How does using hybrid search techniques improve the quality of text summarization compared to traditional methods?
2. Which combination of keyword-based and other search techniques provides the best results for text summarization?
3. Does reranking improve the accuracy of the summaries generated through hybrid search methods?
4. How do the proposed hybrid search and reranking methods compare with existing text summarization techniques in terms of performance and accuracy?
5. Can the proposed hybrid search methods be effectively used for summarizing or other use cases with respect different types of documents, such as news articles, research papers, and other large text corpora?

4. Aim and Objectives

The main aim of this research is to propose a better hybrid search strategy with RAG, additionally I would also like to check if reranking would bring in a value add to the overall output. This would be achieved through looking into other applicable search techniques and experimenting on the available dataset.

The research objectives are formulated based on the aim of this study which are as follows:

- To implement multiple hybrid search techniques with Retrieval-Augmented Generation (RAG).
    - ✓ Here the objective involves with integrating and designing various hybrid strategies within RAG framework. By combining different search strategies the study will aim to enhance the retrieval and generation capabilities of the various models tested.
- To experiment with different weight assignments for various search techniques to optimize output.
    - ✓ Focus will be on the impact of the different weight being assigned to various search techniques within the hybrid model. Once the weights are adjusted in a proper manner this study seeks to identify the optimal combination which maximizes the performance and accuracy both at the same time.
- To evaluate and compare the performance of different hybrid search methods
    - ✓ In order to have a proper evaluation and comparison of various hybrid search strategies the study will employ some of the evaluation metrics. The overall comparative analysis will also help in identifying the most effective hybrid strategy for text summarization or other use cases.
- To develop scripts to automate the testing and evaluation process.
    - ✓ I would be going ahead with developing the scripts in order to automate the testing and evaluation process which would be essential for considering the efficiency and consistency of the experiments. Automated scripts will be streamlining the process, reduce manual intervention as well.
- To create logging tools to track and manage the performance metrics of the hybrid search techniques.
    - ✓ Development of logging tools will be systematically record and manage the performance metrics. Such kind of tools is crucial for monitoring the behavior of hybrid search techniques over time and for analyzing the effectiveness in various situations.

5. Significance of the Study

This study should be able to contribute to the advancement of the Natural Language processing field by exploring into various advanced techniques in text summarization. Integrating RAG with reranking techniques this study should be able to further enhance the accuracy as compared to the past research papers. This advancement or experiments will be able to lead to the development of more complex and skilled summarization models which could be adopted in various practical applications.

I would also aim to improve the information retrieval process in case when there are multiple documents in pace. Having a efficient summarization process will enable users to quickly summarize essential information from large corpus of documents in a very easy manner without addressing any form of information overload. This could be beneficial in wide variety of fields namely legal, medical and research areas. Although the text summarization process was already in research still some advanced techniques could aid to better decision making. The findings from this study could also lead to wide variety of development of tools for many professionals across fields.

This study will also be addressing any limitations if any through experimenting with different hybrid search techniques and reranking methods. By having a measure of the performance of various methods this study aims to overcome the current challenges and put forward more reliable models. This will finally lead to better user experiences and efficient information extraction process.

6. Scope of the Study

Considering in scope I will be having multiple documents as source for data input against which I will be doing an indexing operation. Then I will be entering a input query along with its embedding, the same will go ahead with similarity based hybrid search based on the embedding. Reranking mechanism will be able to retrieve the top 2-3 document sources. Based on this the input query will go the LLM in order to extract the answer. Finally I will apply the accuracy metric to understand the performance of the hybrid search.

If I consider out of scope then I will be starting with fine tuning the model with the multiple documents then I will send the multiple documents as data source for indexing operation to be done. Repeating the above operation I will be entering a input query with its embedding, which would next go for similarity based hybrid search. Next reranking operation would be able to retrieve top documents in order to fetch the answer. Input query next goes to the LLM in order to generate the final answer. Next we compare the output with the human based answer generated in order to understand the performance.

7. Research Methodology

Architecture Diagram

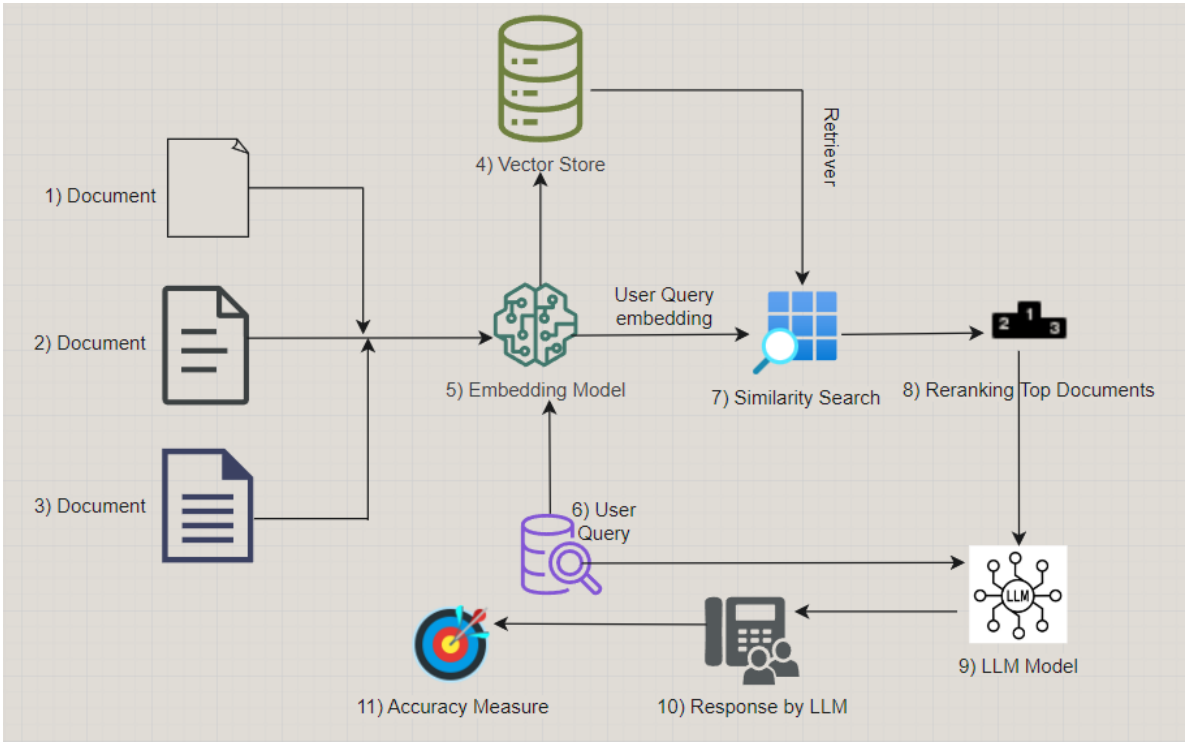


Figure 1. Hybrid Retrieval-Augmented Generation (RAG) Framework for Text Summarization

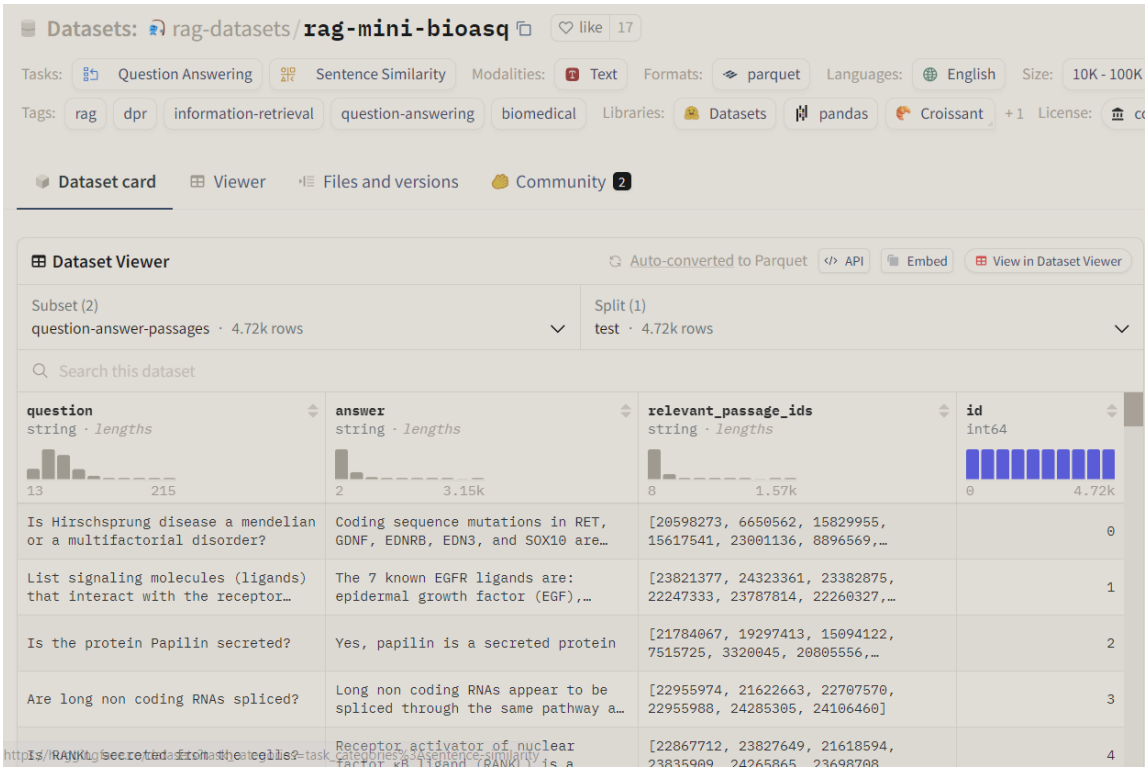


Figure 1.2. Snapshot of the Rag-Mini-Bioasq dataset

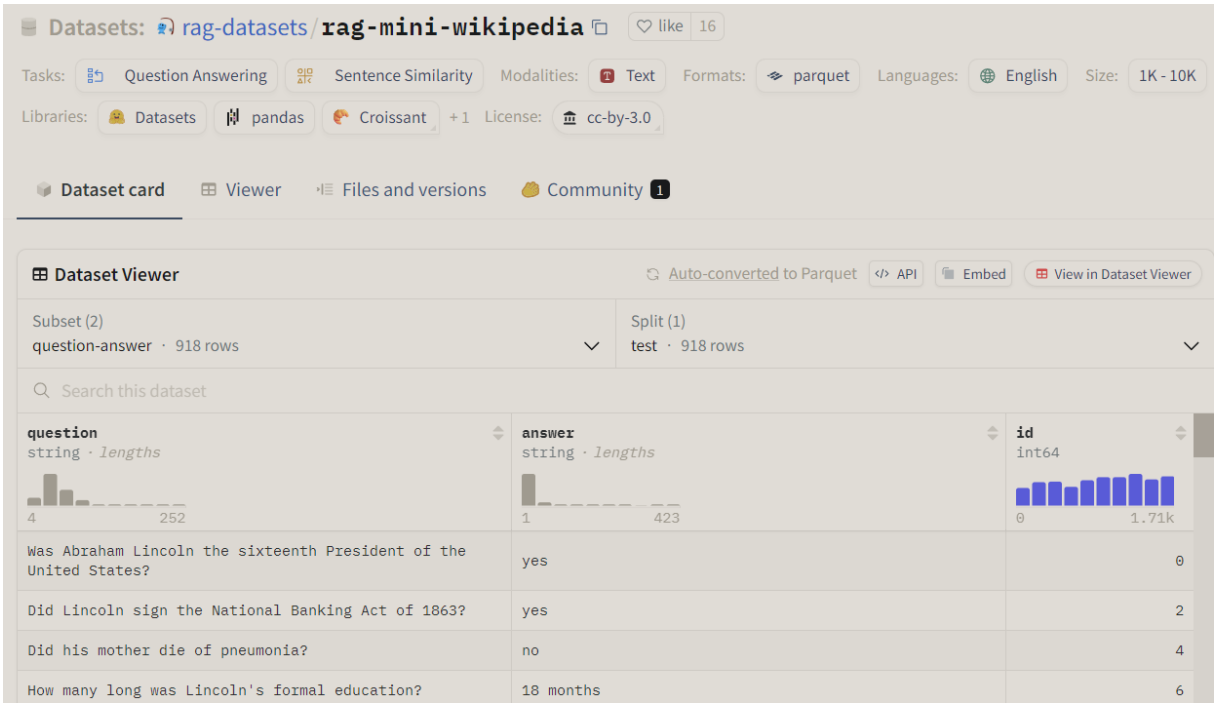


Figure 1.3 Snapshot of the Rag-Mini-Wikipedia Dataset

Dataset

7.1 Data Collection:

- Collect multiple sources of documents to be used for text summarization or other use cases. **(Figure1(1), Figure1(2), Figure 1(3) As per Architecture diagram [Figure 1])**

7.1.1 Embedding Layer

1. **Document Embedding**:
    - Send multiple sources of documents to the embedding layer. **Figure 1(5)**
    - Use an embedding model to convert the documents into vector representations.
    - Store the embedded documents in a document store. **Figure 1(4)**
2. **Query Embedding**:
    - Send the user query to the embedding layer/model. **Figure 1(6)**
    - Convert the user query into a vector representation using the same embedding model.

7.1.2 Indexing and Similarity Search

1. **Indexing**:
    - Perform an indexing operation on the document store to organize the embedded documents for efficient retrieval.
2. **Similarity Search**:
    - Conduct a similarity search **Figure 1(7)** with various hybrid techniques on the indexed document store.
    - Use the embedded user query to search for similar documents in the document store.
3. **Reranking**:
    - Apply a reranking operation in the similarity search layer. **Figure 1(8)**
    - Retrieve the top N documents based on their similarity to the user query.

7.2 Models

7.2.1 Document and Query Processing:

- Send the top N retrieved documents to the LLM model. Here we will be using Llama2 or GPT4 or any other best model available to generate a response based on the user query and the top N documents. **Figure 1(9)**

Generative Pretrained Transformer 4 is a state of the art language model developed by Open AI. This is designed to generate human like text by predicting the next word in a sequence based on the context provided. Its capabilities include text generation to summarization, to translation producing coherent contextually relevant text. Comparing to this Llama model was developed by Meta AI based on the transformer architecture which is designed to generate high quality text by understanding and predicting the next word sequences. It has been trained on diverse textual data enabling it to capture nuances and provide more accurate summaries across different domains.

Using Llama 2 or 3 version would be important since this model is designed to generate human like text by predicting the next word, this model is also know for high performance in NLP tasks, this is also designed in order to handle scalability and this is also suitable for hybrid search techniques. Also GPT 4 could be one of the best model here since this is built on the transformer architecture and has been already trained on a vast corpus of text data. This has better capability in handling of complex queries and also known for providing superior accuracy and coherence in text. This model has also proven its effectiveness in various applications like question answering and content generation.

Here we will be choosing the final model based on the factors as speed, accuracy, scalability, adaptability and ethical implications. Finally we will get the response by the LLM model. **Figure 1(10)**

7.3 Evaluation Metrics

1. **Accuracy Measurement**:
    - Compare the generated response with the ground truth to measure accuracy.
    - Use established evaluation metrics like ROGUE score or blue score or fuzzywuzzy to assess the performance of the text summarization or other use cases.**(11)**
2. **Performance Tracking**:
    - Develop scripts to automate the testing and evaluation process.

- Create logging tools to track and manage the performance metrics of the hybrid search techniques.

8. Requirements Resources

1. **Hardware Resources**:
   - ✓ **High-end GPU**: Essential for training and running RAG models efficiently. Consider GPUs like NVIDIA RTX 3060 or higher.
   - ✓ **High-performance Computing Cluster**: If available, it can significantly speed up training and experimentation processes.
   - ✓ **High-memory RAM**: At least 32 GB, preferably 64 GB or more, to handle large datasets and complex computations.
   - ✓ **Storage**: Sufficient SSD storage (1TB or more) for datasets and model checkpoints.
2. **Software Resources**:

   Python: The primary programming language for implementing and testing RAG models.

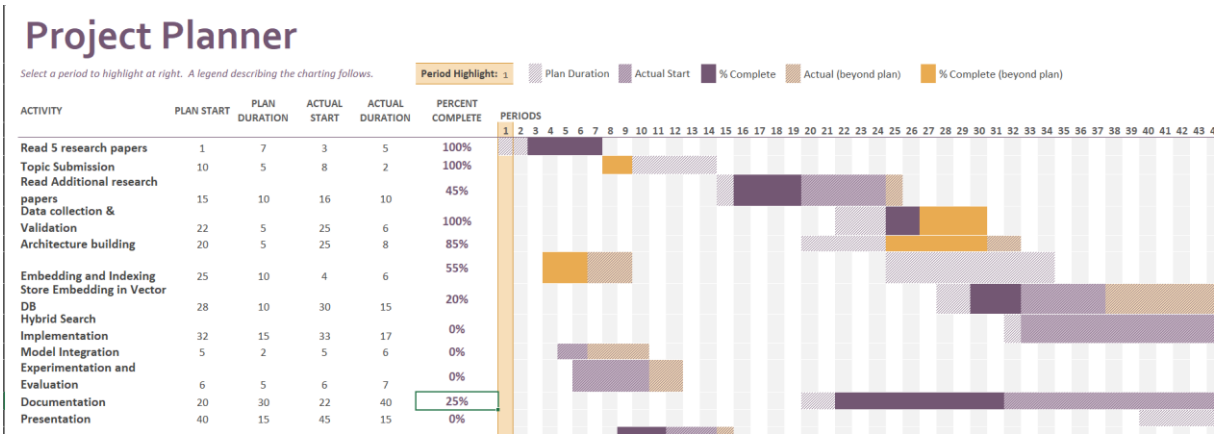
   - **Deep Learning Libraries**:
     - ✓ **PyTorch or TensorFlow**: For building and training neural networks.
     - ✓ **Transformers**: From Hugging Face, for leveraging pre-trained models and implementing RAG.
     - ✓ **LangChain**: For managing and chaining language models in RAG.
       - **NLP Libraries**:
     - ✓ **NLTK**: For natural language processing tasks.
     - ✓ **spaCy**: For efficient NLP operations.
       - **Data Processing Libraries**:
     - ✓ **Pandas**: For data manipulation and analysis.
     - ✓ **NumPy**: For numerical operations.
       - **Evaluation Tools**:
     - ✓ **ROUGE**: For evaluating the quality of summaries.
     - ✓ **BLEU**: For additional evaluation metrics.
3. **Datasets**:

- ✓ **Benchmark Datasets**: Such as CNN/Daily Mail, rag-mini-wikipedia, rag-mini-bioasq and other publicly available datasets for text summarization.
- ✓ **Custom Datasets**: Depending on your specific research focus, you might need to curate datasets relevant to your domain of interest.

4. **Development Environment**:
   - ✓ **Jupyter Notebooks**: For interactive development and experimentation.
   - ✓ **Integrated Development Environment (IDE)**: Such as PyCharm or VS Code, for more extensive coding tasks.
5. **Documentation and Reporting Tools**:
   - ✓ **Mendeley**: For Bibliography and citations we are considering the university approved font/style as suggested.
6. **Cloud Services (Optional)**:
   - ✓ **AWS/GCP/Azure**: For scalable computing resources, especially useful if local hardware resources are limited.

9. Research Plan

Gantt Chart



1. **Literature Review**
   - ✓ Month 1: Review existing literature on hybrid search, RAG, and text summarization techniques. Identify gaps and establish a theoretical framework.
2. **Data Collection and Preprocessing**
   - ✓ Month 2: Collect datasets for text summarization (e.g., CNN/Daily Mail, rag-mini-wikipedia, rag-mini-bioasq) and preprocess the data.
3. **Model Implementation and Initial Testing**
   - ✓ Month 3: Implement baseline models for text summarization. Develop and integrate hybrid search techniques with RAG. Conduct initial testing.
4. **Experimentation and Optimization**
   - ✓ Month 4: Experiment with different hybrid search techniques and weight assignments. Implement reranking methods and optimize the models.
5. **Evaluation and Performance Metrics**
   - ✓ Month 5: Evaluate model performance using metrics like ROUGE and BLEU. Compare results with baseline models and document findings.
6. **Tool Development and Final Analysis**
   - ✓ Month 6: Develop scripts for automated testing, evaluation, and logging. Configure a monitoring dashboard. Analyze results, draw conclusions, and document final findings.

References

Refer: Harvard Referencing Guide

Anon (2014) *2014 Iranian Conference on Intelligent Systems (ICIS) : 4-6 February 2014 : Bam, Iran*. Institute of Electrical and Electronics Engineers.

Avramelou, L., Passalis, N., Tsoumakas, G. and Tefas, A., (2023) Domain-Specific Large Language Model Finetuning using a Model Assistant for Financial Text Summarization. In: *2023 IEEE Symposium Series on Computational Intelligence, SSCI 2023*. Institute of Electrical and Electronics Engineers Inc., pp.381–386.

Babar, S.A. and Patil, P.D., (2015) Improving performance of text summarization. In: *Procedia Computer Science*. Elsevier B.V., pp.354–363.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S. and Larson, J., (2024) From Local to Global: A Graph RAG Approach to Query-Focused Summarization. [online] Available at: http://arxiv.org/abs/2404.16130.

Garg, K.D., Khullar, V. and Agarwal, A.K., (2021) Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization. In: *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*. Institute of Electrical and Electronics Engineers Inc., pp.750–754.

Ieee, (2012) *2012 International Conference on Advances in Engineering, Science and Management.* IEEE.

Ieee, (n.d.) *A. Text mining applications Spam identification: Supervision: Aliases identification Concepts relationship: Search and Retrieval Classification and clustering data Text summarization Case Folding: Stemming: Stop Words: N-grams: Tokenization: B. Text mining and text summarization*.

Ježek, K. and Katedra, J.S., (n.d.) *Automatic Text Summarization (The state of the art 2007 and new challenges)*.

Khan, B., Shah, Z.A., Usman, M., Khan, I. and Niazi, B., (2023) Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey. *IEEE Access*, 11, pp.109819–109840.

Liu, Y. and Lapata, M., (2019) Text Summarization with Pretrained Encoders. [online] Available at: http://arxiv.org/abs/1908.08345.

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T. and Chen, E., (2024) CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models. [online] Available at: http://arxiv.org/abs/2401.17043.

Sawarkar, K., Mangal, A. and Solanki, S.R., (2024) Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. [online] Available at: http://arxiv.org/abs/2404.07220.

Shukla, N.K., Katikeri, R., Raja, M., Sivam, G., Yadav, S., Vaid, A. and Prabhakararao, S., (2023) Generative AI Approach to Distributed Summarization of Financial Narratives. In: *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023*. Institute of Electrical and Electronics Engineers Inc., pp.2872–2876.

Steinberger, J. and Ježek, K., (2009) *EVALUATION MEASURES FOR TEXT SUMMARIZATION*. *Computing and Informatics*, .

Widyassari, A.P., Rustad, S., Shidik, G.F., Noersasongko, E., Syukur, A., Affandy, A. and Setiadi, D.R.I.M., (2022) *Review of automatic text summarization techniques & methods*. *Journal of King Saud University - Computer and Information Sciences*, .