Synthesising Summaries: A novel Retrieval-Augmented Generation-based pipeline for multi-document summarisation

Martin Thomas Callaghan

A thesis submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Doctor of Philosophy

August 2025

Candidate Declaration

I hereby declare that:

1. I have not been enrolled for another award of the University, or other academic or professional organisation, whilst undertaking my research degree.

2. None of the material contained in the thesis has been used in any other submission for an academic award.

3. I am aware of and understand the University's policy on plagiarism and certify that this thesis is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.

4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.

5. The word count of the thesis is 58,401

Name	Martin Thomas Callaghan
Date	11 August 2025
Award	PhD
Research Institute	Industry and Innovation Research Institute
Director of Studies	Dr. Laurie Hirsch

Table of Contents

Figures ***10***

Tables ***11***

Code Listings ***12***

Expressions ***12***

Glossary of Key Concepts ***13***

Published works as outputs of this research ***17***

Acknowledgements ***18***

Abstract ***19***

Chapter 1: Introduction ***20***

1.1 Background and Motivation **20**

1.1.1 Importance of Effective Scientific Communication 20

1.1.2. Challenges in Processing and Understanding Large Volumes of Scientific Literature 21

1.1.3. The Role of Multi-Document Summarisation (MDS) in Mitigating These Challenges 22

1.2. Research Problem **24**

1.2.1. Problem Statement: MDS of Scientific Papers 24

1.2.2. Identifying the Key Challenges and Limitations of Existing Methods 24

1.2.3. Leveraging Recent Advances in NLP and Machine Learning to Address These Challenges 25

1.2.3. Significance of Addressing this Research Problem 27

1.3. Research Objective **30**

1.3.1. Developing an Efficient MDS Framework Focused on Abstractive Techniques 30

1.3.2. Investigating the Use of Advanced Natural Language Processing Techniques 32

1.3.3. Evaluating the Performance of the Proposed Framework on Diverse Scientific Datasets 33

1.3.4. Identifying Potential Applications and Implications of the Research Findings 34

1.3.5 Research Questions and Objectives 35

1.4.2. Limitations of the Research Context, Data Sources, and Methodologies 38

1.4.3. Assumptions and Potential Biases in the Research Approach 40

1.5. Overview of Methodology **42**

1.5.1. Brief Introduction to the Main Research Methodologies 42

1.6 Dissertation structure **43**

1.7 Chapter Conclusion **45**

Chapter 2: Literature Review *47*

2.1 Introduction **47**

2.2 A background to text summarisation **49**

2.2.1 Abstractive vs Extractive summarisation 51

2.2.2 Hybrid Approaches: Combining Abstraction and Extraction 54

2.2.3 Why concentrate on abstractive summarisation? 55

2.3 Multi-Document Summarisation (MDS) **57**

2.3.1 Introduction to MDS and Its Complexities 57

2.4 Challenges in Multi-Document Summarisation **59**

2.4.1 The Challenges Described 59

2.4.2 Pre-trained language models such as BERT, RoBERTa, GPT-3/4, and T5 61

2.4.3: Hybrid techniques in Multi-Document Summarisation 63

2.5 Applying pre-trained language models to the summarisation problem **67**

2.5.1 Outline of Studies Utilising LLMs for Summarisation Tasks 68

2.6 Domain-Specific Summarisation: Focusing on Scientific Papers **69**

2.6.1 Review of previous work on summarising scientific literature 69

2.6.2 Gaps and potential areas for improvement in the current methodologies 71

2.7 Application of Advanced NLP techniques to MDS **73**

2.7.1 Review of Studies Applying These Techniques in MDS 75

2.8 Choosing and preparing data for MDS **85**

2.8.1 Importance of dataset selection for MDS 85

2.8.2 Techniques for data pre-processing and formatting 87

2.8.3 Quality and characteristics of suitable datasets for MDS 90

2.9 Choice of evaluation metrics: assessing summarisation quality **93**

2.9.1 Automated Metrics 93

2.9.2 Strengths and Limitations of Automated Metrics 95

2.9.3 Choosing the Right Metric: A Research Perspective 96

2.9.4 Human Evaluation: The Gold Standard 96

2.9.5 Comparative studies – automated metrics compared to human judgement 97

2.10 Chapter summary **99**

Chapter 3: Modern LLM tools and techniques and their applications to Multi Document Summarisation *101*

3.1 Introduction **101**

3.2 Advances in LLM Architectures: GPT-4, Google Gemini, LLaMA, and Mistral **103**

3.2.1 Overview of LLM Architectures 103
3.2.2 Modern LLM Architectures: GPT-4, Google Gemini, LLaMA, and Mistral 105
3.2.3 Development of Large Language Models (LLMs) from 2022 Onwards 107
3.2.4 Key Milestones and Breakthroughs 109
3.2.5 Comparative Analysis of Notable Models 114

3.3 Understanding LLM Tokens 114

3.3.1 Definition and Role of Tokens in LLMs 114
3.3.2 Tokenisation Techniques 115
3.3.3 Impact of Tokenisation on Model Performance 117

3.4 Pre-Training and Fine-Tuning of LLMs 117

3.4.1 Overview of Pre-Training: Objectives and Techniques 118
3.4.2 Fine-Tuning Methods for Specific Tasks 119
3.4.3 Case Studies of Successful LLM Implementations 120

3.5 Augmenting LLMs with External Knowledge 121

3.5.1 Introduction to Retrieval-Augmented Generation (RAG) 121
3.5.2 Techniques for Integrating External Knowledge 122
3.5.3 Benefits and Challenges 124
3.5.4 Case Study: Multi-Document Summarisation with RAG 124

3.6 Incorporating Knowledge Graphs in LLMs 126

3.6.1 Overview of Knowledge Graphs 126
3.6.2 Techniques for Integrating Knowledge Graphs with LLMs 127
3.6.3 Applications in Enhancing Summarisation Tasks 128

3.7. Modern LLM Techniques in Multi-Document Summarisation 130

3.7.1 Comparative Analysis of Traditional vs. Modern Approaches 130
3.7.2 Comparison of models, features and examples 131
3.7.3 Performance Metrics and Evaluation 133

3.8. Challenges and Future Directions 136

3.8.1 Current Limitations of LLMs in Summarisation 136
3.8.2 Potential Improvements and Innovations 137
3.8.3 Ethical Considerations and Responsible AI 141

3.9. Chapter Conclusion 144

3.9.1 Key Points 144
3.9.2 The Future of MDS with LLMs 145

Chapter 4: Methodology 148

4.1 Introduction 148

4.1.1 Research Objectives and Questions 148

4.1.2 Theoretical Framework 149
4.1.3 Overview of Methodological Approach and Evolution of Models 150
4.2 Mixed Methods Approach 153
4.2.1 Rationale for Mixed Methods 153
4.2.2 Quantitative Components 154
4.2.3 Qualitative assessment components 155
4.2.4 Integration of Quantitative and Qualitative Approaches 157
4.3 Data Collection and Preprocessing 159
4.3.1 Dataset Selection and Justification 159
4.3.2 Data Preprocessing Techniques 161
4.3.3 Ethical Considerations in Data Usage 162
4.4 Software Development and DevOps Techniques 164
4.4.1 Programming Languages and Frameworks 164
4.4.2 Version Control and Collaboration 164
4.4.3 Containerisation and Environment Management 165
4.4.4 Continuous Integration and Deployment 165
4.4.5 Performance Optimisation and Scalability 166
4.5 Error Analysis and Limitations 166
4.6 Reproducibility and Scalability 167
4.7 Comparative Analysis with Existing Methods 168
4.8 Ethical Considerations 168
4.9 Chapter Conclusion 169
Chapter 5: Description of experimental structure 171
5.1 Introduction 171
5.2 Experimental Plan for RAG-based Hybrid Summarisation 173
5.3 Retrieval-Augmented Generation (RAG) Pipeline 174
5.4 Experimental Design 176
5.5 Evaluation Frameworks 178
5.5.1: Automated Metrics 178
5.5.2: Evaluation 178
5.5.3: LLM-as-a-Judge Evaluation 179
Chapter 6: Experiments and results – model, chunking and LLM 181
6.1 Embedding model evaluation and fine-tuning 181

6.1.1 Baseline Embedding Model Evaluation 181
6.1.2 Dataset 181
6.1.3 Preprocessing 181
6.1.4 Embedding Models 181
6.1.5 Evaluation Metrics 182
6.1.6 Methodology for Embedding Creation and Testing 183
6.1.7 Fine-tuning Methodology 185
6.1.8 Results: Base Models 187
6.1.9 Results: Fine-tuned Models 188

6.2 Chunking Strategy Evaluation 190

6.3 LLM evaluation 192

6.3.1 Off-the-shelf LLM Evaluation 192

6.4 Parameter-Efficient Fine-Tuning (PEFT) for Gemma Models 196

6.4.1 LoRA and QLoRA Overview 197
6.4.2 Computational Requirements 197
6.4.3 Hyperparameters and Fine-Tuning Process 197
6.4.4 Challenges and Observations 199
6.4.5 Conclusion 199
6.4.6 Impact of Hyperparameter Choices on Fine-tuning and Model Performance 199

6.5 RAG Pipeline Implementation and Testing 203

6.5.1 Retriever Component 204
6.5.2 Generator Component 211
6.5.3 RAG Integration 216
6.5.4 End-to-end Evaluation 218

Chapter 7: Evaluation Methods - Human Study and LLM-as-a-Judge 222

7.1 Human Evaluation Study 222

7.2 LLM-as-a-Judge Evaluation 223

7.2.3 Comparative Analysis and Insights 223

Chapter 8: Overall RAG Pipeline Evaluation Results and Data Evaluation 225

8.1 Evaluation Results 225

8.2 Correlation Analysis 225

8.3 Discussion of Evaluation Results 226

Chapter 9: Conclusions and Recommendations 228

9.1 Summary of Research 228

9.2 Key Findings and Contributions 229
9.2.1 Retriever Component Refinement 230
9.2.2 Integration and End-to-end Optimisation 231
9.2.3 Re-evaluation and Results 231
9.2.4 Addressing Research Questions 232
9.3 Implications of the Research 233
9.4 Limitations of the Study 236
9.4.1 Dataset Limitations 236
9.4.2 Methodological Constraints 236
9.4.3 Technological Limitations 237
9.5 Future Research Directions 237
9.5.1 Enhancing RAG Techniques for Scientific Literature 237
9.5.2 Exploring Other Large Language Models 238
9.5.3 Improving Evaluation Metrics for Scientific Summarisation 238
9.5.4 Cross-domain Applicability and Generalisation 239
9.5.5 Integration with Scientific Workflow Systems 239
9.6 Concluding Remarks 240
Appendices 242
Appendix 1: The LSTM and the problem with 'Attention' 243
Background to the LSTM 243
The Attention Mechanism 243
Limitations of LSTM in Summarisation 245
The Attention Problem 245
Impact on Summarisation 246
Case Studies and Examples 247
Concluding comments 247
Appendix 2: Human Evaluation Study 249
Methodology: 249
Study Design 250
Analysis 255
Results 255
Appendix 3: LLM-as-a-Judge 257
Rationale: 258
Methodology: 258
Evaluation Process 260
Data Storage 261

Analysis 261
Limitations and Considerations 261
Bibliography 262

Figures

Figure 1: Comparison of extractive vs. abstractive approaches 54
Figure 2: Attention Mechanism 78
Figure 3: Stacked Attention Layers 79
Figure 4: Graph based sentence relationships 81
Figure 5: Timeline of Transformer-based model releases 103
Figure 6: BERT architecture 104
Figure 7: T5 architecture 105
Figure 8: GPT architecture 105
Figure 9: Chunking techniques for RAG 111
Figure 10: Example of chunking techniques applied to text 112
Figure 11: The chunking process 114
Figure 12: Word-based, character-based, and subword-based embeddings 117
Figure 13: RAG retriever-generator 122
Figure 14: The RAG-LLM retrieval process 126
Figure 15: Knowledge graph 126
Figure 16: Knowledge graph integration 128
Figure 17: A proposed KG integration model 130
Figure 18: A comparison of approaches to summarisation 131
Figure 19: Impact of limitations on the summarisation process 137
Figure 20: Flowchart of the Retrieval Augmented Generation (RAG) process 139
Figure 21: Document summarisation pipeline 140
Figure 22: Ethical framework 142
Figure 23: MDS framework 151
Figure 24: Progression of research 153
Figure 25: Inter-relationship of mixed methods approaches 158
Figure 26: Dataset selection strategy 160
Figure 27: Exemplar preprocessing pipeline 162
Figure 28: Container architecture 165
Figure 29: High-Level RAG pipeline 174
Figure 30: Experimental plan flowchart 173
Figure 31: Average Cosine Similarity 183
Figure 32: Structure of the LSTM cell 244
Figure 33: Response form as it appears to the respondent 254
Figure 34: Summary of the LLM-as-a-judge process 257
Figure 35: LLM-as-a-judge implementation 259

Tables

Table 1: Methodological approaches 42
Table 2: Comparison of models 114
Table 3: Summary features and applications of Transformer-based models 131
Table 4: Summary of LLM developments and potential impact on MDS 145
Table 5: Performance of Base Embedding Models 188
Table 6: Performance of Fine-tuned Embedding Models 188
Table 7: Chunking strategies compared 190
Table 8: Off-the-shelf Performance of Gemma Models on SciSummNet 195
Table 9: Key hyperparameters for LoRA/QLoRA fine-tuning 197
Table 10: Hyper-parameter learning rates 200
Table 11: LoRA rank evaluation 200
Table 12: Batch size evaluation 201
Table 13: LoRA dropout rates 202
Table 14: Retrieval Method Performance 210
Table 15: Prompt types and examples 213
Table 16: Performance of Different Prompting Strategies 215
Table 17: RAG Pipeline Optimisation Results 218
Table 18: End-to-end Evaluation Results 218
Table 19: Results from RAG evaluation 225

Code Listings

Code Listing 1: Document Ingestion 125
Code Listing 2: Query and retrieval 125
Code Listing 3: Summary Generation Pseudocode 129
Code Listing 4: Pseudocode for main evaluation 184
Code Listing 5: Fine-tuning process pseudocode 187
Code Listing 6: Pseudocode function to calculate ROUGE score 193
Code Listing 7: Pseudocode function to calculate BLEU score 194
Code Listing 8: Pseudocode function to calculate BERTScore 194
Code Listing 9: pseudocode for data preparation and model setup for Gemma 7B using QLoRA 199
Code Listing 10: Model integration pseudocode 212
Code Listing 11: Pipeline structure pseudocode 217

Expressions

Expression 1: TF-IDF formula 206
Expression 2: BM25 formula 206
Expression 3: Cosine similarity formula 207
Expression 4: Precision@k formula 207
Expression 5: Recall@k formula 207
Expression 6: Mean Reciprocal Rank formula 209

Glossary of Key Concepts

Abstractive Summarisation: A summarisation technique that generates new sentences to capture the essence of the original text, often producing more concise and coherent summaries than extractive methods.

Attention Mechanism: A component in neural networks that allows the model to focus on different parts of the input when generating each part of the output, essential for improving the quality of generated summaries.

BERT (Bidirectional Encoder Representations from Transformers): A transformer-based machine learning model for NLP pre-training, designed to understand the context of a word in a sentence by looking at the words that come before and after it.

BERTScore: An automatic evaluation metric for text generation that computes a similarity score for each token in the generated text with each token in the reference text using contextual embeddings.

BLEU (Bilingual Evaluation Understudy): An algorithm for evaluating the quality of text which has been machine-translated from one natural language to another, often used in summarisation evaluation.

Chain-of-Thought (CoT) Prompting: A prompting technique that guides language models through a step-by-step reasoning process, often improving performance on complex tasks.

Chunking: The process of breaking down large documents into smaller, manageable pieces for processing by language models with limited context windows.

Corpus: a collection of documents. These may or may not be closely related but there is a complex and multilevel relationship between the topics identified within and across documents.

Cross-encoder Re-ranking: A technique used in information retrieval to improve the relevance of retrieved documents by using a BERT-based model to re-score the initial results.

Dense Retrieval: A method of information retrieval that uses dense vector representations of both queries and documents to find relevant information.

Document: several elements of text, each containing information about a single topic. A document therefore will normally contain information about several topics.

Embeddings: Dense vector representations of words, sentences, or documents that capture semantic meanings in a high-dimensional space.

Extractive Summarisation: A summarisation technique that selects and orders existing sentences from the source text to form a summary.

Few-shot Learning: A machine learning approach where a model is trained to perform a task with only a few examples, leveraging its pre-existing knowledge.

Fine-tuning: The process of further training a pre-trained model on a specific task or domain to improve its performance.

Gemma Models: A family of open-source language models developed by Google, available in different sizes (e.g., 2B and 7B parameters).

Large Language Models (LLMs): Advanced AI models trained on vast amounts of text data, capable of understanding and generating human-like text.

LLaMA (Large Language Model Meta AI): A family of foundation language models developed by Meta AI, designed for efficiency and open-source accessibility.

LLM-as-a-Judge: A methodology that uses a large language model to evaluate the quality of outputs from other AI systems, such as summarisation models.

LSTM (Long Short-Term Memory): A type of recurrent neural network architecture used in deep learning, designed to address the vanishing gradient problem.

Multi-Document Summarisation (MDS): The task of producing a single, coherent summary from multiple source documents.

Parameter-Efficient Fine-Tuning (PEFT): Techniques like LoRA and QLoRA that allow for efficient adaptation of large language models to specific tasks with minimal computational resources.

QLoRA (Quantized Low-Rank Adaptation): An efficient fine-tuning technique that combines quantization and low-rank adaptation to reduce the computational and memory requirements of adapting large language models.

RAG (Retrieval-Augmented Generation): A hybrid AI framework that combines information retrieval with text generation to produce more accurate and contextually relevant outputs.

Reciprocal Rank Fusion: A method for combining multiple ranked lists in information retrieval, often used to merge results from different retrieval strategies.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): A set of metrics used for evaluating automatic summarisation and machine translation in natural language processing.

SPECTER: A pre-trained language model specifically designed for scientific document representation, often used in academic information retrieval tasks.

Summary: A text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that.

Text: small volume of text (a paragraph, say) that contains information about a single topic.

TF-IDF (Term Frequency-Inverse Document Frequency): A numerical statistic used to reflect the importance of a word in a document relative to a collection of documents, commonly used in information retrieval and text mining.

Tokenisation: The process of breaking down text into smaller units (tokens), which can be words, subwords, or characters, for processing by natural language models.

Transfer Learning: A machine learning technique where a model developed for one task is reused as the starting point for a model on a second task, often applied in fine-tuning large language models.

Transformer: A deep learning model architecture that relies entirely on an attention mechanism to draw global dependencies between input and output.

Vector Database: A database optimized for storing and querying high-dimensional vector data, often used in conjunction with embedding models for efficient information retrieval.

Zero-shot Learning: The ability of a machine learning model to solve a task it hasn't been explicitly trained on, leveraging its general knowledge or understanding.

Published works as outputs of this research

Callaghan, M. (2023) 'Cloud Computing for Metagenomics: Building a Personalized Computational Platform for Pipeline Analyses', in S. Mitra (ed.) *Metagenomic Data Analysis*. New York, NY: Springer US, pp. 261–279. Available at: https://doi.org/10.1007/978-1-0716-3072-3_13.

Callaghan, M. (2024) 'Chatting with My Data: LLMs for Biodata'. *Festival of Genomics 2024*, London, UK, 24 January 2024.

Callaghan, M. (2024) 'Multimodal AI for Enhanced Information Extraction from Complex HPC Documentation'. *Second Workshop on Multimodal AI*, 25 June 2024. Available at: https://multimodalai.github.io.

Acknowledgements

I would like to express my gratitude to my supervisors, Dr. Laurie Hirsch and Prof. Alessandro di Nuevo, for their guidance and intellectual input throughout this research. Their expertise has been invaluable to the development of this study.

I am also thankful to my various employers, managers and colleagues over the past years for providing a stimulating environment to discuss and refine ideas. Their insights and the conversations we had have helped me develop ideas that have contributed to this work.

Finally, I owe particular thanks to my wife, Jill, for her steadfast support and understanding throughout this journey. Her encouragement has been instrumental in the completion of this thesis.

Abstract

The rapid growth of scientific literature in recent years has created a requirement for efficient methods to synthesise information from multiple related documents. This thesis addresses this challenge by developing and evaluating novel approaches to multi-document summarisation (MDS) of scientific papers, with a focus on hybrid and deep learning techniques leveraging both extractive and abstractive methods.

The research explores the application of state-of-the-art large language models (LLMs), specifically Google's Gemma 2B and 7B models, to the task of scientific literature summarisation. A key innovative approach is the integration of Retrieval-Augmented Generation (RAG) techniques to enhance the summarisation process. The study employs a mixed-methods approach, combining quantitative evaluation metrics with qualitative human assessment and the recently developed novel LLM-as-judge methodology.

A comprehensive literature review provides the theoretical foundation, covering the evolution of summarisation techniques, the emergence of transformer-based models, and recent advances in LLMs and related tools and techniques. The experimental design involves fine-tuning embedding models, optimising chunking strategies, and developing a RAG pipeline that integrates retrieval mechanisms with generative LLMs.

Results demonstrate significant improvements in summary quality, coherence, and factual accuracy compared to baseline methods. The fine-tuned Gemma models, coupled with RAG techniques, show promise in handling the complexities of scientific text. The study also shows interesting trade-offs between model size and performance, with implications for resource-constrained applications.

This research contributes to the field by advancing the state-of-the-art in scientific literature summarisation, providing insights into the effective application of LLMs to this area, and suggesting improved evaluation methodologies. The findings have potential implications for enhancing scientific communication, accelerating literature reviews, and improving access to scientific knowledge.

Chapter 1: Introduction

1.1 Background and Motivation

1.1.1 Importance of Effective Scientific Communication

Effective scientific communication is very important for the advancement of knowledge and innovation (White, 2001; De Semir, 2009; Bautista *et al.*, 2022). It ensures that research findings are disseminated accurately and comprehensively, allowing researchers from various disciplines to build upon each other's work (Gross, Harmon and Reidy, 2002).

Furthermore, it fosters collaboration, interdisciplinary understanding, and contributes to the broader scientific community's ability to address complex problems (Committee on Science, Engineering and Public Policy, 2009). Accessible and clear communication can bridge the gap between experts and non-experts, enhancing the public's understanding of scientific research and its impact on society (Brownell, Price and Steinman, 2013). This communication is not only **for** and **between** other researchers but also to the wider world and can be one of the factors that enhances the impact of a research project.

Despite the impact of social media and pre-print services, scientific papers in journals are still the primary medium for communicating research findings, with countless articles being published every year across a wide range of disciplines. As the volume and frequency of publication of scientific literature grows, it is becoming increasingly challenging for researchers to stay abreast of new developments and synthesise the existing knowledge in their fields; as much research is increasingly collaborative and interdisciplinary in order to produce significant breakthroughs. Saw (2020) emphasises this and Roche and Rickard, (2017) have described how interdisciplinary collaboration in scientific research helps address gaps in knowledge and tackle complex problems. This need to work across disciplines increases the volume of research that needs to be read and synthesised; making it even harder to stay updated with developments in a particular field.

This information overload can hinder the effective communication and application of scientific discoveries, ultimately impeding the progress of research and innovation. In recent research by Lehman and Miller (2020), they discuss some approaches to managing this information overload, they indicate that having on hand a human expert who is able to summarise and distil information is one way to manage this. Of course, not every research team has access to

such human experts so this opens the way to an automated tool which can create summaries in a particular knowledge domain.

1.1.2. Challenges in Processing and Understanding Large Volumes of Scientific Literature

This huge growth of scientific literature presents several challenges for researchers, including the difficulty of finding relevant articles, the time taken to read and understand complex texts and the need to synthesise information from multiple sources. This is echoed by Reis, (2021) who discusses this in relation to the rapid growth and change in a range of scientific fields and the impact this is having on the development of public opinion. This work took place during the "Covid-19" period where very large volumes of medical research were conducted and published during a relatively short period of time.

As the number of published papers continues to increase, so researchers are confronted with an ever-growing volume of information to read, sift and understand, often resulting in them missing key insights or connections. This is especially prevalent in the initial stages of a research project and with those researchers who are new to a particular domain. Shahaf, Guestrin and Horvitz (2012) describe this difficulty and elaborate further, describing the usefulness of visual maps of the interrelationship between concepts in documents (something they term 'Metro Maps of Science') and how such visualisation can help those new to a domain make sense of what they read.

In crossing domains however, researchers are faced with the additional challenge that scientific papers are often written in inaccessible and specialised language, requiring significant background knowledge to understand fully. This is difficult enough in a first language, but, as Pérez-Llantada, Plo and Ferguson (2011) describe, this becomes even more of a challenge for researchers attempting to read research in a different language to their own. This can further contribute to the time and effort required for researchers to process and integrate new information into their work. Additionally, interdisciplinary research increasingly demands the understanding of literature from diverse domains, amplifying the challenge of navigating and synthesising information from multiple sources.

The enormous volume and complexity of scientific literature, coupled with the time constraints faced by researchers, highlight the need for efficient and effective methods to distil and summarise the wealth of knowledge available. In some domains, the rate of production of new

papers and knowledge poses a very real risk that an individual researcher could never keep updated with their field.

Effective Multi Document Summarisation (MDS) of collections of scientific papers can address these challenges by providing concise and coherent synopses of relevant articles, enabling researchers to quickly assimilate critical information and facilitate better-informed decision-making in their work. Several research groups have attempted to automate this process. One notable example is COVIDScholar devised by Trewartha *et al.* (2020); which attempts to automate the discovery and summarisation process by using Natural Language Processing (NLP) techniques to aggregate, analyse and search for (in this case) COVID-19 research literature.

1.1.3. The Role of Multi-Document Summarisation (MDS) in Mitigating These Challenges

MDS has emerged as a promising approach to address the challenges associated with processing and understanding large volumes of scientific literature (Nenkova and McKeown, 2012). By automatically generating concise and understandable summaries from multiple documents on similar topics, MDS can help researchers quickly grasp the main ideas and key findings presented in a set of related articles (Hafeez *et al.*, 2018).

Recent advancements in natural language processing (NLP) and machine learning techniques have significantly improved the capabilities of MDS systems (Radev, Hovy and McKeown, 2002). These systems can now identify and extract important and relevant information from various sources and present it in a coherent and easy-to-understand manner (Zhang, Xu and Wang, 2019). In addition, they can also analyse and aggregate data from different perspectives, thus facilitating a more comprehensive understanding of complex research topics (Ma *et al.*, 2021). This ability to analyse and aggregate from new perspectives is one feature currently being explored with the new generation of Large Language Models (LLMs). Indicating the continued relevance of Transformer based approaches as described by Amatriain *et al.*, (2023), in a paper from Lui and colleagues (Liu *et al.*, 2021), they discuss using RoBERTa (a variant of the Transformer model) and its ability to rapidly learn new linguistic knowledge across domains.

By incorporating MDS techniques into the research process, researchers can efficiently navigate the growing volume of scientific literature, better understand complex interdisciplinary topics, and more easily stay up-to-date with the latest developments in their fields. In some pre-

transformer work by Wan et.al (2013), they describe a topic-based approach to analysing topics within research papers across domains that is able to do just this; identifying and mapping sub-topics across documents.

Furthermore, these generated summaries can also serve as valuable resources for non-experts, helping them understand scientific research and its implications more easily (Wang *et al.*, 2016). In this regard, MDS of scientific papers has the potential to significantly enhance the effectiveness of scientific communication and contribute to the overall progress of research and innovation.

1.2. Research Problem

1.2.1. Problem Statement: MDS of Scientific Papers

Multi-Document Summarisation (MDS) creates concise overviews from multiple related documents. In a review of techniques undertaken by Ma and colleagues (Ma *et al.*, 2021), they describe MDS as potentially leading to more cogent summaries but offset by complexity and contradiction often seen in multiple perspectives on a similar set of topics. The purpose of this study then is to create a computational pipeline, using a range of techniques, to produce accurate summaries quickly at the lowest possible computational cost (so using smaller and more energy efficient techniques and models).

In the context of scientific papers then, MDS aims to provide a comprehensive overview of the key findings, methods and implications discussed in a set of articles (Li, Li and Li, 2012), whilst avoiding some or most of the complexities and contradictions. This enables researchers to efficiently access and assimilate information from multiple sources, so enhancing their understanding and decision-making in the research process (John, Premjith and Wilscy, 2017). In this study, the MDS process does not aim to simplify or explain the concepts further, only to summarise and distil.

1.2.2. Identifying the Key Challenges and Limitations of Existing Methods

Despite significant progress in MDS research, several challenges and limitations remain, particularly in the context of scientific papers. Some of these challenges include:

1. **Domain-specific terminology and concepts**: Scientific papers often contain specialised language and concepts that may be difficult for summarisation systems to understand and process accurately. This can lead to the generation of summaries that are unclear or misleading, particularly for non-expert readers. Goldstein *et al.*, (2000) discuss this, together with describing an approach to address these challenges with domain-independent techniques (ie. slightly different approaches for different knowledge domains) within a modular framework for MDS, highlighting the difficulty in processing domain-specific terminology
2. **Identifying relevant and novel information**: Summarising scientific papers requires identifying the most relevant and novel contributions across multiple documents. Existing methods may struggle to accurately distinguish between significant and trivial findings, leading to summaries that lack focus or omit important details. In some early work by Mani and Bloedorn (1997), they emphasise the importance of identifying

relevant and novel information in MDS to improve accuracy and topic relevance in summaries. This identification of 'novel' information (or topics within documents) then could form a key element of creating effective summaries.

3. **Preserving coherence and logical structure**: MDS systems must not only extract relevant information but also present it in a coherent and well-structured manner (Nenkova and McKeown, 2012). Ensuring that the generated summaries are easy to read and understand, particularly when dealing with complex scientific topics, remains a challenge for many existing methods.
4. **Evaluation and benchmarking**: Evaluating the quality and effectiveness of MDS systems is a complex task, given there is often no single "correct" summary for a given set of documents. Developing or selecting appropriate evaluation metrics and benchmark datasets for assessing the performance of scientific paper summarisation methods remains an ongoing challenge. In a series of studies, Koh *et al.*, (2023) discuss the complexity and challenge of evaluating MDS systems, including analysing benchmark datasets, models, and metrics. Furthermore, they describe a number of interesting hybrid approaches around the selection of narratives within documents as a model to focus on relevant material.

1.2.3. Leveraging Recent Advances in NLP and Machine Learning to Address These Challenges

Recent advancements in natural language processing (NLP) and machine learning, particularly in the field of deep learning, have opened new avenues for addressing the challenges associated with MDS of scientific papers (Devlin *et al.*, 2018). These advancements have led to the development of more sophisticated and effective summarisation techniques, which can better understand and process complex domain-specific language and generate more coherent and informative summaries (Gao *et al.*, 2024)

Some of the key advances in NLP and machine learning that can potentially improve MDS include:

Pre-trained language models: The introduction of pre-trained language models, such as BART, BERT, the OpenAI GPT family, and RoBERTa, has significantly improved the performance of various NLP tasks, including summarisation (Devlin *et al.*, 2018; Liu *et al.*, 2021). These models can be fine-tuned for the specific task of summarising scientific papers, enabling them to better capture domain-specific terminology and generate more accurate and coherent summaries. Kumar, Choudhary and Cho (2021) show how pre-trained models like BERT and RoBERTa can be

used for data augmentation, which helps in adapting the models for specific domains like scientific paper summarisation, enabling the models to handle specialised language more effectively and Venkataramana, Srividya and Cristin (2022) have described the efficiency of the BART model in summarising and extracting important information from large documents, making it particularly suitable for analysing and condensing information from scientific papers. This ability to condense complex information into short summaries is therefore very important for managing the dense and technical content often found in scientific literature.

Neural abstractive summarisation: Following on from the above, neural abstractive summarisation techniques, such as sequence-to-sequence models and transformer-based architectures, have demonstrated the ability to generate more fluent and informative summaries compared to older extractive methods. These techniques can potentially address the limitations of existing summarisation methods in terms of coherence and logical structure. Some recent studies (Song *et al.*, 2019; Subramanian *et al.*, 2020) have confirmed that abstractive methods can produce a broad range of summaries, from purely factual to ones that are a little more generative. These often outperform those traditional (extractive) approaches.

Graph-based approaches: Graph-based approaches, such as Knowledge Graphs and Graph Neural Networks (GNNs), have proven to be effective in modelling complex relationships between entities and concepts in scientific papers. Studies have shown that GNNs can significantly outperform related methods in capturing these intricate relationships, thereby enhancing the extraction of relevant and novel information (Kipf and Welling, 2017; Zhou *et al.*, 2020). By capturing these relationships, GNNs can potentially improve the identification and extraction of relevant and novel information in MDS, but have also shown promise in tasks as varied as modelling physics systems, learning molecular fingerprints and classifying diseases.

Interpretable and explainable AI: The development of interpretable and explainable AI (XAI) techniques is increasingly important in the field of scientific summarisation by aiming to make AI decisions more transparent and understandable, especially in high-stakes fields such as medicine and law. Some of the current challenges in XAI include defining model explainability and designing evaluation measures. However, advancements in this area can help researchers understand the rationale behind generated summaries, facilitate the identification of potential biases, and lead to more accurate and trustworthy summaries. Murdoch *et al.* (2019)provide a comprehensive overview of interpretable machine learning, discussing its importance across various domains, including in scientific research. They state that as AI systems become more

complex and are applied to critical decision-making processes, the need for interpretability and explainability becomes even more important. In the context of scientific summarisation then, the authors argue that explainable AI can help researchers understand how summaries are generated, identify potential biases in the summarisation process, and so increase trust in AI-generated summaries. This is particularly important in fields like medicine and law, where the implications of decisions based on these summaries could be far-reaching. The paper also highlights the ongoing challenges in defining and measuring explainability.

1.2.3. Significance of Addressing this Research Problem

Addressing the research problem of MDS of scientific papers has wide-ranging implications and benefits for the scientific community and beyond. By developing more effective and efficient summarisation techniques, researchers can better navigate the ever-growing volume of scientific literature, facilitating a number of significant outcomes:

1. **Accelerated scientific progress**: As researchers can more efficiently access and assimilate information from multiple sources, the time required to understand and build upon existing knowledge is reduced, so potentially accelerating scientific discovery and innovation (Hey and Trefethen, 2020)
2. **Enhanced interdisciplinary research**: By providing comprehensive overviews of key findings, methods, and implications across different domains, MDS systems can facilitate interdisciplinary research, fostering collaboration and promoting the exchange of ideas between researchers from diverse fields (Cabanac, Frommholz and Mayr, 2019)
3. **Improved decision-making**: Summaries generated by MDS systems can provide valuable insights for decision-makers in academia, industry, and policy-making. These insights can support evidence-based decision-making, leading to more informed choices and better resource allocation (Grimshaw *et al.*, 2012).
4. **Increased research accessibility**: High-quality summaries of scientific papers can make complex research findings more accessible to non-expert readers, including the general public, journalists, and policymakers. This can help bridge the gap between academia and society, fostering a more informed and engaged public (Nisbet and Scheufele, 2009).
5. **Encouraging open science**: Effective MDS systems can contribute to the open science movement by enabling researchers to quickly and easily access and understand research findings from various sources, regardless of their background or expertise

(Nosek *et al.*, 2015).

In addition to the direct benefits for the scientific community, advancements in MDS techniques can have broader applicability in various sectors, as these technologies are not limited to the summarisation of scientific papers. Some of these broader implications include:

1. **Enhanced business intelligence**: MDS can be applied to analyse and summarise large volumes of business documents, such as market reports, financial statements, and news articles. This can provide decision-makers with a comprehensive overview of relevant information, enabling them to make more informed business decisions and better understand market trends (Gupta, Hanges and Dorfman, 2002).
2. **Improved legal research and analysis**: The legal sector can benefit from the application of advanced MDS techniques to analyse and summarise legal texts, such as case law, legislation, and regulatory documents. This can help legal practitioners more efficiently access and understand complex legal information, leading to better legal advice, and more effective advocacy (Susskind, 2010).
3. **Enhanced educational resources**: MDS can be employed to generate concise and coherent summaries of educational materials, such as textbooks, articles, and lecture notes. These summaries can serve as valuable study aids for students, helping them to better understand and retain complex concepts and information. For example, Oliveira *et al.* (2022) describe a learning tool with associated text mining functions can support students develop academic writing skills.
4. **Improved crisis response and management**: In times of crisis, such as natural disasters or public health emergencies, the ability to quickly and effectively process and summarise large volumes of data from various sources can be important for decision-makers. MDS systems can facilitate the rapid assimilation of critical information, enabling a more effective response and better allocation of resources. For instance, the CORD-19 dataset, which compiled over 500,000 scholarly articles about COVID-19 and related coronaviruses, has been used with various MDS techniques to help researchers quickly synthesise information from this vast corpus, accelerating the understanding of the virus and potential treatments (L. L. Wang *et al.*, 2020)
5. **Streamlined media monitoring and analysis**: Advanced MDS techniques can be employed to analyse and summarise large volumes of news articles, social media posts, and other media content, providing comprehensive insights into public opinion, sentiment, and trending topics and possibly to even rapidly identify “fake news”. This

can benefit organisations in diverse sectors, including public relations, marketing, and journalism, by allowing them to better understand and respond to the rapidly changing media landscape (Liu, Hu and Cheng, 2005).

1.3. Research Objective

1.3.1. Developing an Efficient MDS Framework Focused on Abstractive Techniques

To address the research problem of MDS of scientific papers, the development of an efficient and effective framework is essential. In this section, the focus is on abstractive summarisation techniques which have been shown to generate more coherent and informative summaries compared to extractive methods. Abstractive techniques are able to synthesise text and generate novel sentences, providing more accurate and context-aware summaries than other techniques . The key components and considerations involved in designing and implementing an abstractive summarisation framework are described, with an initial emphasis on the relatively simple BERT and BART models, and the exploration of hybrid techniques.

1. **Data collection and pre-processing**: The first step in building a MDS framework is to collect a large and diverse dataset of scientific papers from various sources, such as academic journals, preprint repositories, and conference proceedings. Pre-processing techniques, including tokenisation, stemming, and stopword removal, must be applied to clean and normalise the text data, ensuring that the input data is consistent and suitable for further analysis (Isinbayeva and Przepiorka, 2024).
2. **Domain-specific language model fine-tuning**: To improve the summarisation system's ability to handle domain-specific terminology and concepts, pre-trained language models, such as BERT or BART, can be fine-tuned on the collected dataset of scientific papers (Beltagy, Lo and Cohan, 2019a; Gururangan *et al.*, 2020). This fine-tuning process allows the model to adapt to the unique linguistic characteristics and patterns present in scientific texts, leading to better performance in the summarisation task.
3. **Summarisation model selection and training**: In this step, the focus is on abstractive methods for generating summaries. Transformer-based architectures, such as BERT and BART, have shown great promise in generating high-quality abstractive summaries (Cachola *et al.*, 2020; Lewis *et al.*, 2021). To further enhance the framework's effectiveness, hybrid techniques that identify sections of interest in scientific papers for summarisation can be explored. This approach combines the strengths of extractive methods in identifying key information with the abstractive capabilities of generating coherent and informative summaries.
4. **Evaluation metrics and benchmarking**: To assess the performance and effectiveness of the developed summarisation framework, appropriate evaluation metrics must be employed. These may include ROUGE scores for measuring the overlap between generated summaries and human-authored reference summaries, as well as qualitative

assessments of coherence, informativeness, and readability (Bhandari *et al.*, 2020; Fabbri *et al.*, 2021). The framework's performance should therefore be benchmarked against existing summarisation systems and techniques to ensure that it represents a significant improvement over the work that came previously.

5. **Interpretability and explainability**: Incorporating interpretable and explainable AI techniques into the summarisation framework can help researchers understand the rationale behind the generated summaries and identify potential biases and limitations (Danilevsky *et al.*, 2020). These techniques can also facilitate the fine-tuning of the model, enabling the development of more accurate and trustworthy summaries.

1.3.2. Investigating the Use of Advanced Natural Language Processing Techniques

The proposed research will potentially explore a number of advanced natural language processing (NLP) techniques that may be used to enhance the performance of an MDS framework. These techniques can aid in the extraction of key information, the generation of coherent summaries, and the adaptation of models to the domain-specific context of scientific papers. Techniques include:

Pre-trained language models: Pre-trained language models, such as BERT, RoBERTa, GPT-3/4, and T5, together with newer Large Language Models such as Google's Gemma series of models, have shown enormous potential in various NLP tasks, including summarisation (Beltagy, Peters and Cohan, 2020; Zaheer *et al.*, 2021). By incorporating these models into the framework, the system can leverage their ability to understand and generate human-like text, resulting in improved summary quality.

Transfer Learning: Transfer learning techniques, allow models to "store" knowledge acquired from one domain and then apply it to another, can be used to enhance the summarisation framework's performance. For instance, fine-tuning a pre-trained language model on a dataset of scientific papers can help the model adapt to the unique linguistic patterns and terminologies of the scientific domain (Gururangan *et al.*, 2020). Transfer learning can work at multiple levels of fine tuning; for example, a model trained on general Computer Science papers can then be further trained on a specialised sub-domain with an appropriate set of papers.

Attention mechanisms: Attention mechanisms, originally discussed by (Bahdanau, Cho and Bengio, 2016), have proven to be highly effective at various NLP tasks, including summarisation. These mechanisms allow models to selectively focus on different parts of the input text, providing a more nuanced understanding of the content and improving the coherence and informativeness of generated summaries (J. Zhang *et al.*, 2020). An original focus of this PhD research project was to explore the applicability of LSTM (long short-term memory) architectures to summarisation tasks but this approach was discarded as LSTMs are no longer regarded as "State Of The Art"; Transformer based models (the new "State Of The Art" following work from Vaswani *et al.* (2017)) are therefore the focus of this research, augmented with more traditional ML (Machine Learning) and NLP (Natural Language Programming) approaches.

Graph-based methods: Graph-based methods, such as graph neural networks (GNNs), can be utilised to capture the complex relationships between different parts of the input text. By representing scientific papers (or rather the relationships between concepts in and between

papers) as graphs, GNNs can learn meaningful representations that help identify key information and generate higher quality summaries (D. Wang *et al.*, 2020).

Domain adaptation via fine tuning: To further improve the performance of the summarisation framework, domain adaptation techniques can be employed (Hua and Wang, 2017). These methods help the model to learn and generalise from a smaller set of labelled data by leveraging information from related, but distinct, domains. This can be particularly useful when working with those scientific papers that span multiple disciplines or sub-domains.

1.3.3. Evaluating the Performance of the Proposed Framework on Diverse Scientific Datasets

The effectiveness and robustness of the developed Multi-Document Summarisation (MDS) framework should be evaluated across a range of scientific datasets spanning various domains and disciplines. Several key considerations and methodologies are important for this evaluation process and are described below.

Dataset selection is very important. A diverse set of scientific datasets should be chosen to assess the framework's performance across multiple domains. These datasets may include large-scale collections of scientific articles, such as the ACL Anthology (Bird *et al.*, 2008), PubMed (Lu, 2011), or arXiv (McKiernan, 2000). The selection should cover a wide range of topics and disciplines to ensure a comprehensive evaluation.

Data preprocessing and partitioning are also essential steps in preparing the datasets for evaluation. This includes tokenisation, stemming, and stopword removal (Manning, Raghavan and Schütze, 2008). Furthermore, the data should be partitioned into training, validation, and test sets to ensure that the framework's performance is evaluated on unseen data, providing a more reliable estimate of its generalisability (X. Liu *et al.*, 2018).

The application of suitable evaluation metrics is essential for quantitatively measuring the performance of the summarisation framework. Metrics such as **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (C.-Y. Lin, 2004), **BLEU** (Bilingual Evaluation Understudy) (Papineni *et al.*, 2002), and **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) (Lavie and Agarwal, 2007) can be used to compare the generated summaries with human-authored reference summaries. Additionally, qualitative evaluations involving human

judgments of coherence, informativeness, and readability can provide valuable insights into the framework's performance.

Comparison with existing methods is also essential for contextualising the framework's performance. Benchmarking against existing techniques, including classic models such as BERTSUM (Liu, 2019), PEGASUS (J. Zhang *et al.*, 2020), and GPT-3 (T. B. Brown *et al.*, 2020), can help identify the framework's strengths and weaknesses, as well as potential areas for improvement.

Finally, generalised evaluation across various scientific domains and disciplines is necessary to assess the framework's robustness and adaptability. By measuring performance across different datasets, researchers can identify potential domain-specific biases and limitations, informing future improvements and adaptations of the framework (Cohan and Goharian, 2015).

1.3.4. Identifying Potential Applications and Implications of the Research Findings

Advancing the field of MDS as applied to scientific papers can have implications and applications across various domains thus increasing the potential applications of the research findings and their broader impact on academia, industry, and society. Some potential applications are as follows:

1. **Accelerating scientific discovery**: By providing coherent and informative summaries of scientific papers, the proposed summarisation framework can help researchers quickly grasp the main findings and contributions of relevant literature, thereby accelerating scientific discovery and promoting cross-disciplinary research (H. Wang *et al.*, 2023). This can lead to more efficient knowledge dissemination and a better understanding of novel research developments.
2. **Enhancing literature review and meta-analysis**: MDS techniques can aid researchers in conducting literature reviews and meta-analyses by extracting key insights from a large volume of articles (Cohen *et al.*, 2006). This can help identify trends, gaps, and areas for future research, fostering innovation and collaboration in various scientific domains.
3. **Supporting research communication and public engagement**: The generated summaries can be utilised to communicate research findings to non-expert audiences, facilitating public engagement with scientific research and promoting science literacy (Baram-Tsabari and Lewenstein, 2013). This can help bridge the gap between researchers and the public, fostering an informed dialogue on scientific issues.

4. **Information retrieval and recommendation systems**: The summarisation framework can be integrated into information retrieval and recommendation systems, improving the relevance and quality of search results and suggestions (Manning, Raghavan and Schütze, 2008). This can enhance user experience and facilitate access to pertinent scientific knowledge.
5. **Industrial applications**: Beyond academic research, a developed MDS framework can be applied to various industries, such as healthcare, finance and the legal sectors, where rapid access to summarised information from multiple documents is critical for decision-making and analysis (Nenkova, 2011).

1.3.5 Research Questions and Objectives

The primary aim of this research is to develop an efficient and effective multi-document summarisation framework for scientific papers, leveraging advanced natural language processing techniques and focusing on abstractive approaches. The research questions for this study have been formulated based on the current state of the field, the identified gaps in the literature, and the potential practical applications of the proposed framework(s) (Afantenos, Karkaletsis and Stamatopoulos, 2005; Devlin *et al.*, 2018; Lewis *et al.*, 2020).

The following research questions will guide the investigation:

RQ1: What are the key features and characteristics of an efficient hybrid multi-document summarisation framework for scientific papers, and how can Retrieval-Augmented Generation (RAG) techniques be effectively incorporated to identify and use sections of interest?

RQ2: How can state-of-the-art language models be adapted and fine-tuned for the task of multi-document summarisation of scientific papers, and what advantages do newer LLMs (such as Gemma 2B/7B) offer over earlier models (like BERT and BART)?

RQ3: How does the performance of the proposed hybrid framework compare to existing approaches, both extractive and abstractive when evaluated using standard metrics (e.g., ROUGE, BLEU) and on diverse scientific datasets?

To address these research questions, the study will address the following evaluation objectives:

1. Develop and implement multi-document summarisation frameworks based on advanced natural language processing techniques, with a focus on abstractive methods and the adaptation of pre-trained models such as BERT and BART and other language models as may be appropriate (Devlin *et al.*, 2018; Lewis *et al.*, 2020).
2. Investigate the potential benefits of hybrid techniques, possibly combining extractive and abstractive approaches, to identify and summarise relevant sections of scientific papers effectively.
3. Evaluate the performance of the proposed framework using established evaluation metrics, such as ROUGE (C.-Y. Lin, 2004) and BLEU (Papineni *et al.*, 2002), and compare it to existing state-of-the-art methods on a variety of scientific datasets such as SciTLDR and SciSummNet. These are widely used datasets of scientific papers, abstracts and summaries.

1.4. Scope and Limitations

1.4.1. Justifying the Focus on Scientific Papers

The focus on Multi-Document Summarisation (MDS) of scientific papers is based on several key factors that underscore the challenges and opportunities presented by this domain. This section describes the reasons behind concentrating on scientific papers and explores the potential benefits of developing summarisation techniques tailored to this particular context.

As already discussed, the exponential growth of scientific literature in recent years has caused an information overload, making it increasingly challenging for researchers to remain abreast of the latest findings and developments in their respective fields (Bornmann and Mutz, 2015). The development of effective MDS techniques for scientific papers can then significantly aid researchers in swiftly identifying and assimilating relevant knowledge, thereby accelerating scientific discovery and facilitating interdisciplinary research.

Peer-reviewed research articles in academic journals (referred to as 'scientific papers') are characterised by complex terminologies, technical jargon, and dense writing styles which pose substantial challenges for traditional summarisation methods (Afantenos, Karkaletsis and Stamatopoulos, 2005). By focusing on this domain, the proposed research can address these challenges and develop summarisation techniques that are robust and adaptable to the unique patterns and structures in scientific literature.

The structured nature of academic research papers papers, typically comprising sections such as abstract, introduction, methodology, results, and conclusion, presents an opportunity for the MDS framework to leverage this inherent structure (Gupta and Lehal, 2010). This can facilitate the generation of more coherent and informative summaries that accurately capture the key contributions and findings of the source papers.

Furthermore, summarising scientific papers can enhance research accessibility, making findings more comprehensible to a wider audience, including non-experts, policymakers, and practitioners (Baram-Tsabari and Lewenstein, 2013). By developing effective summarisation techniques tailored to scientific papers, this research can contribute to promoting scientific literacy and fostering public engagement with scientific research.

The techniques and insights gained from the development of MDS methods for scientific papers have broader applicability. They can be adapted to other domains that require the

synthesis of complex, structured, and domain-specific information (El-Kassas *et al.*, 2021). This potential for wider application can lead to the development of more effective and versatile summarisation tools with utility across various sectors.

In conclusion, the focus on MDS of scientific papers is justified by the pressing need to manage the growing volume of scientific literature, the unique challenges presented by scientific writing, the potential to leverage document structure, the opportunity to enhance research accessibility, and the broader applicability of the developed techniques. This research direction promises to yield significant advancements in the field of automatic summarisation while addressing critical needs in scientific communication and knowledge dissemination.

1.4.2. Limitations of the Research Context, Data Sources, and Methodologies

While the focus on Multi-Document Summarisation (MDS) of scientific papers offers numerous advantages, it is still important to acknowledge and critically examine the limitations inherent in the research context, data sources, and methodologies. This section describes these constraints and their potential impact on the generalisability and applicability of any research findings.

The domain-specific challenges presented by scientific literature present a double-edged sword. On one hand, they offer unique opportunities for developing sophisticated summarisation techniques. On the other hand, they may limit the direct applicability of these methods to other domains. Cohan and Goharian (2016) highlight the distinct linguistic and structural characteristics of scientific papers, including their specialised terminology, complex syntax, and a standardised rhetorical (way of presenting argument) structure. These features, while useful for summarisation within the scientific domain, may not translate directly to other texts such as news articles or social media content, which have been the focus of many previous studies (Nallapati *et al.*, 2016; J. Zhang *et al.*, 2020). Still, the insights gained from addressing these challenges could inform the development of more versatile summarisation methods adaptable to various contexts.

The selection of data sources for training and evaluation introduces potential biases and limitations that need careful consideration. Bornmann and Mutz (2015) suggest that the exponential growth and diversification of scientific literature highlights the challenge of obtaining a truly representative sample. Focusing on specific scientific domains or journals might result in a narrow representation of the diversity of scientific literature. To mitigate this limitation, researchers should try to use diverse datasets that cover a broad range of topics and disciplines. For instance, (Cohan *et al.*, 2018) introduced the PubMed and arXiv datasets, which

span multiple scientific domains, providing a more comprehensive basis for evaluation. However, even these larger datasets may not fully capture the global diversity of scientific literature, particularly from non-English language sources or emerging research fields. That notwithstanding, there are now more comprehensive datasets such as SciSummNet (Yasunaga *et al.*, 2019) and TLDR; (Cachola *et al.*, 2020) that provide focussed corpora of papers that may be more suitable for these summarisation tasks.

The choice of evaluation metrics and benchmarks significantly influences the interpretation of the summarisation framework's performance. Traditional automatic metrics such as ROUGE (C.-Y. Lin, 2004) and BLEU (Papineni *et al.*, 2002) have been widely used in summarisation tasks. However, these metrics have limitations in capturing the nuanced quality of summaries, particularly in the scientific domain. Cohan and Goharian (2018) proposed the use of citation-based evaluation metrics for scientific document summarisation, arguing that they better reflect the importance of content in scientific papers. Nevertheless, reliance solely on automatic metrics may not fully capture the cognitive and qualitative aspects of human-authored summaries. Therefore, incorporating human evaluation becomes essential, despite its time-consuming and potentially subjective nature. Fabbri *et al.* (2021) introduced a comprehensive human evaluation protocol for summarisation, which could be adapted for scientific document summarisation to obtain a more holistic assessment of summary quality.

Methodological limitations arise from the choices of NLP techniques, pre-processing steps, and model architectures. These decisions can significantly impact the generalisability, efficiency, and interpretability of the developed summarisation framework. For instance, Liu and Lapata (2019) demonstrated the effectiveness of pre-trained language models for summarisation tasks. However, these models often require substantial computational resources and may struggle with the longer document lengths typical in scientific literature. Additionally, D. Wang *et al.* (2020) proposed a graph-based approach for scientific document summarisation which explicitly models the document discourse structure. While this approach shows promise in capturing the logical flow of scientific arguments, it may be less effective for documents with less explicit structure. Future research could investigate hybrid approaches that combine the strengths of different techniques to optimise the framework's performance and robustness across various scientific domains and document types.

Ethical considerations and potential misuse of advanced summarisation techniques also raise important concerns. The risk of plagiarism or misrepresentation of research findings, as highlighted by Baram-Tsabari and Lewenstein (2013) becomes more pronounced with increasingly sophisticated summarisation tools. Moreover, the inherent biases in training data, such as gender or geographical biases in scientific publishing (Holman, Stuart-Fox and Hauser,

2018), could be perpetuated or even amplified by summarisation models. Researchers must actively work to identify, mitigate, and transparently communicate these biases. Additionally, there is a need to develop guidelines and best practices for the responsible use of scientific summarisation tools to ensure they enhance rather than undermine the integrity of scientific communication and to avoid possible misleading or inaccurate summaries. In recent years, this is one of the areas of AI ethics that has developed into the concept of **guardrails**. These guardrails serve as important safeguards, ensuring that AI-generated summaries maintain factual accuracy, coherence, and ethical integrity. Recent research, such as the work by (Dong *et al.*, 2024), has highlighted the critical importance of addressing issues like hallucination prevention, bias mitigation, and source attribution in multi-document summarisation tasks.

1.4.3. Assumptions and Potential Biases in the Research Approach

In any research activity, it is important to acknowledge the assumptions and potential biases that may influence the research approach. This section discusses the assumptions and potential biases that may be present in the development of a multi-document summarisation framework for scientific papers and their implications on the research findings.

1. **Assumptions about language and domain-specific features**: The research approach assumes that the language and domain-specific features of scientific papers can actually be captured and modelled effectively using advanced natural language processing techniques (Beltagy, Lo and Cohan, 2019). This assumption may overlook the complexity and variability of these features across different scientific disciplines, potentially impacting the generalisability and adaptability of the developed summarisation framework. As this research programme will focus on summarisation of scientific (and especially computer science) papers, this is a significant risk but one that is acceptable considering the scope and scale of typical PhD research.
2. **Assumptions about model architectures**: The research approach assumes that certain model architectures, such as BERT and BART (although likely to be extended to more modern Large Language Model architectures), are well-suited for the task of multi-document summarisation of scientific papers (Zaheer *et al.*, 2021). This assumption may lead to potential biases in the selection and evaluation of model architectures, potentially overlooking alternative approaches that may offer better performance or efficiency. The following chapter (literature review) will explore the applicability of these and other architectures and approaches in more detail.

3. **Assumptions about data quality and representativeness**: The research approach assumes that the datasets used for training and evaluation are of high quality and adequately represent the diversity of scientific literature (L. L. Wang *et al.*, 2020). This assumption may overlook potential biases or limitations in the data sources, such as the over-representation of certain topics or disciplines, which could impact the generalisability and applicability of the research findings. In this study, it is likely that it will rely on standard datasets used in other studies to give some degree of comparability.
4. **Assumptions about evaluation metrics**: The research approach assumes that the chosen evaluation metrics, such as ROUGE and BLEU, provide an accurate and meaningful assessment of the summarisation framework's performance (Bhandari *et al.*, 2020). This assumption may overlook potential biases or limitations in these metrics, which may not fully capture the subtlety of human-authored summaries, or the specific challenges posed by the multi-document summarisation of scientific papers.
5. **Assumptions about user needs and preferences**: The research approach assumes that the generated summaries meet the information needs and preferences of the intended users, such as researchers, policymakers, and practitioners (Lauscher *et al.*, 2018). This assumption may overlook potential variations in user needs and preferences across different contexts, potentially impacting the relevance and utility of the developed summarisation framework. The study will therefore need to consider the use of a small-scale human study to evaluate this.
6. **Assumptions about the stability of research topics**: The research approach assumes that the topics and areas of interest in scientific literature remain relatively stable during the study (Boyack and Klavans, 2019). This assumption may overlook the dynamic nature of scientific research, where new areas of interest or emerging topics can arise rapidly. To mitigate this limitation, the developed summarisation framework should be designed to adapt to evolving research landscapes and maintain its performance and relevance over time.
7. **Assumptions about ethical considerations**: The research approach assumes that the developed summarisation framework will be used responsibly and ethically by researchers, policymakers, and practitioners (Nanayakkara, Hullman and Diakopoulos, 2021). This assumption may overlook potential risks, such as the misuse of the framework for plagiarism or misrepresentation of research findings. Researchers should develop guidelines and safeguards to ensure the responsible use and

dissemination of the summarisation framework and engage in ongoing discussions about the ethical implications of their work.

1.5. Overview of Methodology

1.5.1. Brief Introduction to the Main Research Methodologies

To effectively address the research questions and objectives outlined in this dissertation, a multi-method research approach was employed, combining both qualitative and quantitative techniques. A more detailed description of all aspects of methodology is in Chapter 4.

Table 1: Methodological approaches

Approach	**Justification**
Literature review and meta-analysis: An extensive and systematic review of the existing literature on multi-document summarisation techniques, advanced natural language processing models, and evaluation metrics will be conducted (Allahyari *et al.*, 2017). This will include a meta-analysis of existing methods and their performance in order to identify gaps and opportunities for improvement. The review will inform the development of the proposed framework and provide a solid theoretical foundation for the study.	Conducting a systematic literature review and meta-analysis is essential for establishing a strong theoretical foundation for the study and identifying gaps in the existing research. This methodology enables a more complete understanding of the current state of the field, including the most recent advancements in multi-document summarisation techniques and natural language processing models. Moreover, it provides useful insights into the performance of existing methods and the potential avenues for improvement, guiding the development of the proposed framework.
Computational modelling and simulation (experimental approach): The development of the proposed multi-document summarisation framework will involve the adaptation and fine-tuning of state-of-the-art natural language processing models, such as BERT and BART, as well as the exploration of hybrid techniques that combine extractive and abstractive approaches. Computational modelling and simulation will be employed to iteratively refine and optimise the framework, ensuring its efficiency and effectiveness.	The use of computational modelling and simulation is vital for the development and optimization of the multi-document summarisation framework, given the complex nature of the task and the reliance on advanced natural language processing techniques. This methodology allows for the iterative refinement of the framework and the exploration of different approaches, such as the adaptation of pre-trained models and the incorporation of hybrid techniques. Computational modelling and simulation provide a means to experiment with various configurations and settings (such as hyperparameters), ensuring the efficiency and effectiveness of the proposed framework.
Evaluation and benchmarking (mixed methods approach- experimental and human evaluation): The performance of the proposed	A rigorous evaluation and benchmarking process is critical for determining the performance of the proposed framework and

framework(s) will be assessed using established evaluation metrics, such as ROUGE and BLEU, and compared to existing state-of-the-art methods. This will involve the curation of diverse scientific datasets, representing various disciplines and document types, to ensure the generalisability of the results. The evaluation process will also include qualitative assessments of the generated summaries, seeking input from human subjects to validate the framework's effectiveness in capturing the essential content, maintaining coherence and enhancing understanding.	validating its effectiveness in comparison to existing state-of-the-art methods. By employing established evaluation metrics and diverse scientific datasets, this methodology ensures that the results are reliable, generalisable, and comparable to other approaches in the field. Furthermore, the inclusion of qualitative assessments from human readers enhances the validity of the evaluation process, confirming the framework's ability to generate coherent and informative summaries.
Case studies and applications: To further demonstrate the practical utility of the proposed framework, case studies will be suggested to illustrate its potential applications in various contexts, such as interdisciplinary research, science communication, and decision-making processes. These case studies will provide real-world examples of how the developed summarisation framework can be employed to facilitate information synthesis and knowledge dissemination.	The use of case studies and applications is essential for demonstrating the practical utility of the proposed multi-document summarisation framework in real-world contexts. By illustrating how the framework might be employed in various situations, such as interdisciplinary research, science communication, and decision-making processes, this methodology helps to bridge the gap between theoretical advancements and practical applications. Additionally, the case studies provide valuable feedback on the usability and potential impact of the framework, informing further refinements and developments.

1.6 Dissertation structure

This dissertation is organised into several chapters, each focusing on a specific aspect of the research on Multi-Document Summarisation (MDS) of scientific papers. The following overview provides a brief description of each chapter's content and purpose.

Chapter 1 is an introduction to the research, presenting the background and motivation for the study. It discusses the importance of effective scientific communication and highlights the challenges researchers face in processing and understanding large volumes of scientific literature. The chapter introduces MDS as a potential solution to these challenges, defining the research problem and objectives. It also outlines the key challenges and limitations of existing MDS methods, particularly in the context of scientific papers. The chapter also discusses recent research in Natural Language Processing (NLP) and machine learning that are relevant to addressing these challenges in MDS.

Chapter 2 describes the purpose of the literature review and its relevance to the research topic. It then explores the background of document summarisation, covering its history, importance in various fields, and the distinction between single and multi-document summarisation.

A significant portion of the chapter is dedicated to comparing extractive and abstractive summarisation techniques, explaining their respective advantages and disadvantages. The focus then shifts to multi-document summarisation, discussing its complexities and challenges. The chapter also explores the use of pre-trained language models like BERT, RoBERTa, BART, GPT-3/4, and T5 in summarisation tasks. It addresses domain-specific summarisation, particularly for scientific papers, and the application of advanced NLP techniques in MDS.

Chapter 3 examines the recent research and development in Large Language Models (LLMs) and their application to MDS. It focuses on state-of-the-art models like GPT-4, Google Gemini, LLaMA, and Mistral, tracing their evolution from earlier models such as BERT and T5.
The chapter also explores key architectural developments, including the introduction of Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA and QLoRA, which have significantly reduced computational requirements for model adaptation. It also discusses advancements in embeddings, chunking techniques, and vector databases, describing their importance in managing large text corpora. The role of tokens and tokenisation in LLMs is explained, showing the impact of effective tokenisation on model performance. The chapter covers the phases of pre-training and fine-tuning in LLM development and introduces techniques like Retrieval-Augmented Generation (RAG) and the integration of knowledge graphs to enhance LLM capabilities.

Chapter 4 outlines the methodology employed in the study of multi-document summarisation of scientific papers. The research aims to develop and evaluate novel hybrid approaches that combine extractive and abstractive methods, with a focus on adapting large language models like Gemma 2B and 7B. The study is based on techniques in natural language processing, information retrieval, and machine learning, leveraging transformer-based architectures, Retrieval-Augmented Generation (RAG), and transfer learning. A mixed-methods approach is adopted, integrating quantitative and qualitative methodologies to provide a comprehensive understanding the impact of the summariser.

Chapter 5 presents a structured experimental approach and sets of results for evaluating and refining a RAG-based hybrid summarisation system for scientific papers. The plan is divided into seven key phases, each addressing essential aspects of the summarisation pipeline.

Chapter 6 covers the core parts of the summarisation framework, namely the embedding fine-tuning and evaluation, the chunking strategy, evaluation of the LLMs, PEFT fine-tuning & evaluation and concludes with the full pipeline implementation and testing.

Chapter 7 covers evaluation of the full pipeline, using both a human study and the 'LLM-as-a-judge' methodology.

Chapter 8 pulls together the evaluations and presents a correlation analysis and a discussion of the evaluation results.

Chapter 9 provides the conclusions derived from the research, including a summary of the research and how the research objectives have been met, together with recommendations, limitations and future work.

1.7 Chapter Conclusion

This introductory chapter has outlined the research objectives and questions to be answered in this study. It explained the importance of effective scientific communication in the advancement of knowledge and innovation as a key driver of the study. It has highlighted the growing challenges researchers face in processing and understanding the ever-increasing volume of scientific literature, particularly in interdisciplinary fields. The chapter has also introduced Multi-Document Summarisation (MDS) as a promising solution to these challenges, emphasising its potential to help researchers efficiently navigate and synthesise information from multiple sources.

The research problem of MDS for scientific papers has been clearly defined, along with the key challenges and limitations of existing methods. These include dealing with domain-specific terminology, identifying relevant and novel information, preserving coherence and logical structure, and the complexities of evaluation and benchmarking.

Recent advances in Natural Language Processing (NLP) and machine learning, particularly in deep learning and pre-trained language models, have been discussed as potential avenues for

addressing these challenges. The chapter has also indicated the significance of this research problem, not only for accelerating scientific progress and enhancing interdisciplinary research but also for its broader applications in various sectors such as business intelligence, legal research, education, crisis management, and media analysis.

The research objective has been established: to develop an efficient MDS framework focused on abstractive techniques. This framework aims to leverage state-of-the-art NLP models and techniques to generate more coherent, informative, and accurate summaries of scientific papers.

The subsequent chapters will explore key research in the field, the methodological approach, the implementation of the proposed framework and the evaluation of its effectiveness in addressing the challenges of scientific literature summarisation. This research aims to make contributions to the field of scientific communication and information management, potentially revolutionising how researchers interact with and synthesise knowledge from the enormous collection of scientific literature.

Chapter 2: Literature Review

2.1 Introduction

The purpose of this review is to describe some of the key features of text summarisation, the different methods currently in use for single document and multiple document summarisation and to describe approaches to evaluate the summaries created by software tools and programs.

The key features of a summary are that it should be (Radev, Hovy and McKeown, 2002):

- Short and concise
- Contains the important information from the donor documents
- Contains information from one or more documents

The area and discipline of modern text summarisation, despite being a relatively young field of study with its advent in the early 2000s, has rapidly evolved with advancements in technology and the growth of computational power. This advancement has been particularly propelled by the rise of high performance computing (HPC), general-purpose graphic processing units (GP-GPUs), and cloud computing, enabling the development and deployment of sophisticated models (Manning, 2015).

Deep learning, a subset of machine learning that relies on complex, multi-level artificial neural networks, has been instrumental in the progress of text summarisation. Lawrence and LeCun, renowned researchers in the field, envisage natural language processing (NLP) to be the next significant milestone for deep learning (Manning, 2015). The goal is to equip machines with an understanding that extends beyond individual words to entire sentences and paragraphs.

However, the intricacy of human language and the current limitations of available tools and methodologies present a significant challenge in this endeavour. Specifically, there is yet to be a unified, simple, and efficient method to accurately and concisely summarise even a single structured document. This challenge is compounded when the task extends to summarising a corpus of documents, particularly as new concepts are added over time. Nenkova and McKeown (2012) discuss the complexities involved in text summarisation, highlighting the challenges of developing a unified, efficient method for summarising even single documents, let alone multiple documents. They point out that while significant progress has been made in

certain areas, the task of accurately and concisely summarising complex documents remains a considerable challenge.

Despite the challenges in text summarisation, certain aspects of this field are well-researched, with promising tools and techniques emerging to address known issues, especially those associated with summarising long and complex documents. These well-researched areas include extractive summarisation techniques, which have been refined to effectively select and arrange existing sentences from source texts (Nallapati, Zhai and Zhou, 2016). Additionally, the application of attention mechanisms in abstractive summarisation has significantly improved summary quality by enabling models to focus on relevant input text portions (See, Liu and Manning, 2017). Domain-specific summarisation methods have been tailored for types of documents such as scientific literature or legal texts (Cohan and Goharian, 2018), while multi-document summarisation approaches have been designed to consolidate information from multiple sources (P. J. Liu *et al.*, 2018). Furthermore, the development and refinement of evaluation metrics like ROUGE have provided standardised ways to assess summary quality (C. Lin, 2004).

The literature reviewed in this chapter is important to shape and ground the research. This review will provide insights into the development and current state of text summarisation and the computational methods used within it, including high-performance computing, deep learning, natural language processing, and network theory. The evolution of these techniques, their applications in the field of text summarisation, and their strengths and limitations will be explored in detail.

Specifically, this review will explore the critical differences and advantages of extractive and abstractive summarisation, two principal methodologies in text summarisation. This differentiation is essential as the proposed research heavily leans towards abstractive summarisation.

Further, the literature review will probe the use of pre-trained language models in summarisation tasks. These models represent some of the most advanced tools currently available and will form the basis for the proposed summarisation framework. This, and the following chapter will also discuss recent developments in Large Language Models, such as the OpenAI GPT models, the Google Gemma models and several other open-source models and techniques and their applicability to text summarisation.

This review of existing work on domain-specific summarisation, particularly of scientific papers, will show some of the unique challenges and opportunities in this field. This is particularly relevant as the proposed research aims to develop a summarisation framework specifically for scientific literature.

Lastly, the review will evaluate the application of advanced NLP techniques in multi-document summarisation, which is the focus of this research. Understanding the potential and the limitations of these techniques will guide the development and refinement of the proposed summarisation framework.
In summary, the relevance of the reviewed literature lies in its ability to provide a comprehensive understanding of the current state of text summarisation. It allows the identification of gaps and opportunities in the existing methods ultimately informing the development of a new approach that is grounded in, yet advances beyond, the current state of the art.

2.2 A background to text summarisation

Document summarisation is a developing field that has seen a notable increase in interest and application over the past two decades and an acceleration in development in just the past five years. At its core, document summarisation is the process of reducing a larger text document into a concise summary that retains the main points and salient details of the original text (Nenkova and McKeown, 2012). This process can be executed manually by human reviewers, but the rise of computational power and advanced algorithms developed through machine learning and deep learning has shifted the focus towards automatic summarisation techniques.

The history of document summarisation as a computational rather than a purely linguistic problem began in the late 1950s with the work of Hans Peter Luhn (Luhn, 1958). Luhn, at the time working at IBM, developed an algorithm that identified the most frequent words in a document (excluding common words like 'and', 'the', etc.), and used this as a basis to identify key sentences for inclusion in the summary. This area of research remained little more than a curiosity and it was only in the early 2000s, with the advent of more powerful computing capabilities (such as GPUs and other vector processor chips) and the rise of the Internet, that automatic text summarisation started to gain significant momentum in research and practical applications (Mani, 2001).

Following Luhn's important and critical work, various techniques for automatic summarisation were explored. These included the use of semantic networks, graph-based models, and statistical methods. Spärck Jones (2007) provides an extensive overview of the evolution of automatic summarisation techniques since Luhn's original and very important work. She discusses various approaches that emerged over time, including the use of semantic networks, graph-based models, and statistical methods. The paper not only covers these techniques but also evaluates their effectiveness and discusses the challenges faced in the field of automatic summarisation. However, these early methods often struggled with maintaining the coherence and readability of summaries. The development of Machine Learning and Natural Language Processing (NLP) in the late 1990s and into the 2000s paved the way for more sophisticated approaches. Notably, the introduction of extractive and abstractive summarisation techniques have allowed rapid development in the field, offering more nuanced and accurate summarisation capabilities (Radev, Hovy and McKeown, 2002; Rush, Chopra and Weston, 2015).

The importance of document summarisation extends to various fields and sectors. In academia, summarisation can aid in literature reviews and the synthesis of research findings across multiple studies, thereby enhancing knowledge dissemination and uptake (Cohn and Lapata, 2008). For businesses, summarisation tools can help process large amounts of textual data, such as customer reviews or business reports, providing essential insights quickly and efficiently (Das and Martins, 2007).
In the news industry, automatic summarisation can assist in generating concise news briefs or digesting multiple reports on the same event, such as Columbia's 'Newsblaster' which is an early tool to cluster news into events (McKeown *et al.*, 2003). More recently, the healthcare sector has begun leveraging document summarisation to condense patient records and medical literature, facilitating efficient information retrieval and decision-making (Spasić *et al.*, 2014).

In the legal field, automatic summarisation can support the review of lengthy legal documents, contracts, and case laws, making the process more efficient and manageable (Mochales and Moens, 2011). In the domain of scientific research, automatic summarisation is particularly vital due to the exponential growth of publications. It helps researchers stay abreast of new developments, trends, and breakthroughs across multiple disciplines (Teufel and Moens, 2002). In the era of big data, the ability to summarise vast amounts of textual information will prove invaluable across a wide range of sectors.

Document summarisation then, despite being a relatively recent field of study, has quickly become an important tool across various sectors due to its ability to distil large amounts of text into digestible, informative summaries.

Whilst the bulk of the research in the field initially focused on single document summarisation, the challenge and importance of multi-document summarisation have grown in prominence (Goldstein, Mittal, J. Carbonell, *et al.*, 2000). Single document summarisation concerns itself with generating a summary for a lone input document. However, multi-document summarisation (MDS) aims to create a cohesive and non-redundant summary from a set, or 'corpus', of documents. This could be multiple reports on the same event, a collection of research papers on a particular topic, or an entire body of work from a specific author. MDS presents additional challenges such as identifying and resolving conflicting information, avoiding redundancy, and maintaining a coherent narrative when drawing from multiple sources (Nenkova and McKeown, 2012).

2.2.1 Abstractive vs Extractive summarisation

Text summarisation is a well explored field but in common with many other problems in natural language programming, difficult and computationally expensive. As researchers in the field are dealing with natural human language (English in this case), slight differences in the context or semantics of a summary can lead to different understanding by different human readers. This is made even more complex in knowledge domains where terminology and phraseology have specific meaning, such as in clinical and engineering fields.

The summarisation process, as identified by Radev, Hovy and McKeown (2002), comprises two primary phases. The initial stage involves the identification and extraction of salient material from source documents. This is followed by a second phase where the extracted material is merged, modified or edited to produce a new set of sentences that concisely and fluently summarise the original text.

Earlier work Hahn and Mani (2000) classified summarisation techniques as either **knowledge-rich** or **knowledge-poor**. Knowledge-rich approaches consider the meaning of the document and its constituent sentences, enabling more effective reduction and better summary creation through understanding. In contrast, knowledge-poor approaches treat documents as unordered, context-free collections of words. Despite their limitations, knowledge-poor

approaches offer advantages in their ability to be applied unchanged to different documents across new application areas, making them relatively straightforward to reapply.

Prior to this, the 1980s and 1990s saw extensive use of statistical methods to produce moderately successful summaries. These methods often employed a 'bag of words' approach, treating documents as large collections of words where meaning and context were deemed unimportant. However, the establishment of the Document Understanding Conference (DUC) series was a shift towards recognising the importance of contextualisation and the inter-relationship of sentences and paragraphs within a document. As Verma and Lee (2018) noted, the meaning and relative context of sentences are very important considerations in effective summarisation.

As indicated earlier, Nenkova and McKeown (2012) have described several processes common to many text summarisation tools. One key process is intermediate representation, where an intermediate step occurs in the summarisation process. An important method within this is topic representation, which identifies key topics and their relationships in the text or document. This can be accomplished through various means, including frequency identification, TF*IDF (term frequency-inverse document frequency), and topic-word approaches.

- In the **frequency identification** approach, the number of times a word or phrase appears in a document can be taken as an indicator of its importance.
- The **TF*IDF** approach (actually a product of two different statistical indicators) (Sparck Jones, 1972) is used to weight a particular keyword in the context of the document(s) in which it appears. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic used to evaluate the importance of a word in a document within a collection or corpus. It consists of two main components: Term Frequency (**TF**) and Inverse Document Frequency (**IDF**). TF measures how frequently a term appears in a document, calculated as the number of times a word appears in a document divided by the total number of words in that document and IDF measures how important a term is across the entire corpus, calculated as the logarithm of the total number of documents divided by the number of documents containing the term. TF-IDF is the product of these two metrics, increasing proportionally to the number of times a word appears in a document but offset by the frequency of the word in the corpus. This

adjustment helps make up for the fact that some words appear more frequently in general.

- In **topic word** approaches, a simple table summary of words and weightings can be produced where highly weighted words can be regarded as more indicative of the topic.

Once candidate sentences or phrases have been identified, the summariser then needs to select the most appropriate set of important sentences to produce a paragraph-length summary for each of the important topics identified in the document.
Automatic text summarisation, as a field of study, has traditionally been split into two distinct strategies: extractive and abstractive summarisation.

Extractive Summarisation, whilst a simpler approach, remains widely used and highly effective in certain contexts. This approach involves identifying and extracting key segments or snippets from the original document(s). These 'extracts', usually whole sentences deemed most representative of the overall content, are then concatenated to form the final summary. One might visualise this as creating a highlight reel of the most significant points raised in the text. Algorithms for extractive summarisation often rely on statistical and linguistic features, such as word frequency, sentence position, named entities, or similarity to the title.

The main advantage of extractive summarisation lies in its relatively straightforward implementation. Moreover, as it uses text directly from the source, it minimises the risk of generating grammatically incorrect or nonsensical sentences. But, notable limitations exist. Extractive summaries often lack coherence when sentences are taken out of context and can struggle with avoiding redundancy when summarising multiple documents containing overlapping information. Additionally, they inherently cannot condense information beyond the sentence level, making them less effective for longer, more complex texts.

Abstractive Summarisation, on the other hand, attempts to emulate human summarisers by generating new text that encapsulates the main ideas in the source document(s). This involves a deeper understanding of the text, as the summarisation model must parse the semantic meaning of the source, identify key points, and then reformulate this information in a concise and coherent manner. This process often involves complex NLP techniques, including parsing, semantic interpretation, inference, and text generation.

The potential of abstractive summarisation is considerable, as it can generate more natural, concise, and coherent summaries, especially for longer and more complex texts. It can synthesise information from multiple sentences or even from different documents in the case of multi-document summarisation, creating a summary not limited by the original phrasing or sentence structure. However, the challenges are equally substantial. Abstractive summarisation models require larger amounts of computational resources and run the risk of generating summaries that may misrepresent the original text or contain factual inaccuracies, as they are not bound to use the exact wording of the source material. The diagram below (fig. 1) compares the two summarisation techniques on a simple phrase.

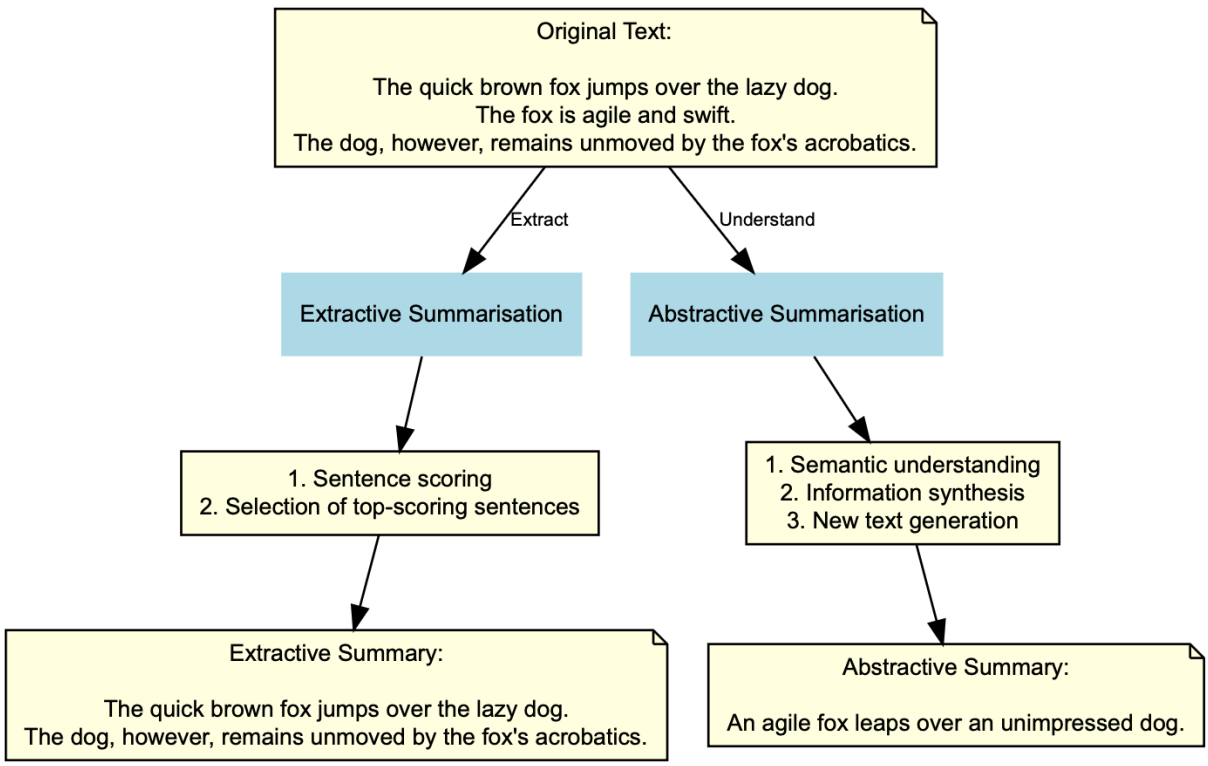

Figure 1: Comparison of extractive vs. abstractive approaches

2.2.2 Hybrid Approaches: Combining Abstraction and Extraction

Recent years have seen growing interest in hybrid approaches that combine elements of both extraction and abstraction. These methods aim to leverage the strengths of each approach while mitigating their respective weaknesses (Gehrmann, Deng and Rush, 2018). For instance, an extractive step might identify salient sentences or concepts, followed by an abstractive step to generate a concise and coherent summary. This approach can potentially reduce

redundancy and improve summary readability while maintaining a close tie to the source content.

The implementation of such hybrid models introduces additional complexity and computational overhead. Careful design is required to ensure that the integration of the two methods does not inadvertently introduce new issues, such as discrepancies between the extracted and abstracted content. In the following chapter, some of the hybrid techniques (such as Retrieval Augmented Generation) are discussed further.

As research in summarisation techniques continues, there is a clear trend towards more sophisticated abstractive methods. These approaches may overcome the limitations of purely extractive techniques, offering the potential for more flexible, concise, and human-like summaries. The following section will further explore abstractive summarisation, exploring its mechanisms, challenges, and future prospects.

2.2.3 Why concentrate on abstractive summarisation?

Recent research in neural language models have significantly enhanced the capabilities of abstractive summarisation. Models such as BERT and BART (Devlin *et al.*, 2019; Lewis *et al.*, 2020) have demonstrated the ability to understand and generate language in a subtle manner, aligning closely with the requirements of creating meaningful, coherent, and contextually accurate summaries of complex scientific texts. These models, pre-trained on vast amounts of text, can generate novel sentences and be fine-tuned for specific domains, making them particularly suited for summarising scientific literature.

Building on these foundations, the more recent development of Large Language Models (LLMs) has pushed the boundaries of what is possible in text generation and understanding. GPT-3 (T. Brown *et al.*, 2020) was a significant milestone, demonstrating impressive few-shot learning capabilities across various NLP tasks, including summarisation. More recent models like GPT-4 (OpenAI *et al.*, 2024) have further refined these capabilities, demonstrating enhanced ability to understand and generate nuanced text, which is particularly valuable for summarising complex scientific literature.

The incorporation of hybrid techniques (such as RAG), which combine elements of both extractive and abstractive methods, has shown promising results in recent studies. Zhang et al. (2020) suggested a hybrid approach that first extracts relevant sentences and then applies

abstractive techniques to generate a final summary. This method leverages the strengths of both approaches, potentially overcoming the limitations of purely extractive or abstractive methods.

In the context of scientific document summarisation, the work of Menick *et al.* (2022) on evaluating GopherCite demonstrated how LLMs can be used to generate summaries with faithful attribution to source documents, addressing one of the key challenges in abstractive summarisation: ensuring factual accuracy and traceability of information.

The emergence of open-source LLMs, such as BLOOM (BigScience Workshop *et al.*, 2023) and LLaMA (Touvron *et al.*, 2023), has democratised access to these powerful models, allowing researchers to explore their application in specialised domains like scientific literature summarisation. Models such as these offer the potential for fine-tuning on domain-specific corpora, which could significantly enhance their performance in summarising scientific papers. While these advancements in LLMs offer exciting possibilities for abstractive summarisation, they also present new challenges and areas for investigation. Issues such as computational efficiency, model interpretability, and the mitigation of biases in generated summaries remain active areas of research (Bender *et al.*, 2021).

In conclusion, the focus on abstractive techniques in this research is justified by their alignment with the complex requirements of summarising scientific literature, their potential to address key challenges in this domain, and the recent research in neural language models that have significantly enhanced abstractive summarisation capabilities. The exploration of state-of-the-art LLMs in the context of multi-document scientific summarisation is a promising area for research as it builds on the strengths of earlier models while taking advantage of the enhanced capabilities of more recent LLMs to address the complex requirements of summarising scientific literature.

2.3 Multi-Document Summarisation (MDS)

The progression from single document summarisation naturally leads to the domain of multi-document summarisation (MDS). MDS extends beyond summarising a single document to encompass the summarisation of an entire set of documents. Whilst the primary objective remains the generation of a concise and coherent summary, the complexity and challenges are significantly amplified due to the increased volume of text and the requirement to capture relevant content from multiple documents which may present related or diverse viewpoints.

2.3.1 Introduction to MDS and Its Complexities

Multi-Document Summarisation involves creating a comprehensive summary from multiple source documents. These documents may present similar or contrasting perspectives on a topic, thereby introducing additional layers of complexity to the summarisation task. The summarisation process must account for both commonalities and discrepancies across these documents, ensuring that the final summary accurately represents the information without bias (Endres-Niggemeyer *et al.*, 1998). In MDS, the challenges inherent in single-document summarisation are compounded by the additional complexities introduced by multiple document inputs. The system must address issues such as redundancy, contradiction, and coherence maintenance while synthesising information from multiple sources (Radev *et al.*, 2004). These challenges are particularly pronounced when dealing with large document sets or when the documents cover a wide array of sub-topics.

Redundancy management is a critical aspect of MDS. When summarising multiple documents on a similar topic, the likelihood of encountering repeated information is high. A robust MDS system should identify these redundancies and eliminate them to prevent unnecessary repetition in the generated summary. Carbonell and Goldstein (1998) introduced the Maximal Marginal Relevance (MMR) criterion, which aims to reduce redundancy while maintaining relevance in the summary. This approach has been widely adopted and adapted in subsequent research. Conversely, the system must also manage contradictions or differences in the presented facts or views across documents. This becomes particularly important in fields such as news reporting or scientific literature, where different documents might present varied viewpoints or findings. Balancing these contradictions while ensuring the summary remains objective is a complex task. Dou *et al.* (2021) proposed a system called GSum, which specifically addressed the issue of contradictory information in news articles, demonstrating early attempts to tackle this challenge.

Maintaining coherence in the final summary presents another significant challenge. The summary should not only capture key information from multiple documents but also present it in a manner that ensures a logical and easily comprehensible narrative. This involves understanding the relationships between different pieces of information and structuring the summary accordingly. Barzilay, McKeown and Elhadad (1999) proposed an information fusion approach to improve the coherence of multi-document summaries, which has influenced subsequent research in this area.

In the context of scientific literature, MDS faces the additional complexity of handling specialised language and technical terms commonly used in scientific texts. The system must possess a robust understanding of these terms and their context to generate accurate and meaningful summaries. Cohan and Goharian (2015) addressed this challenge by proposing a method specifically designed for scientific article summarisation, which takes into account the unique structure and language of scientific documents.

Despite these challenges, MDS serves as an essential tool in a number of fields where dealing with large amounts of textual data is commonplace. Its applications range from summarising news articles (Radev et al., 2004) to condensing scientific literature (Qazvinian, 2010) and even summarising user-generated content on social media platforms (Inouye and Kalita, 2011).

More recently, researchers have begun exploring hybrid approaches that combine the strengths of both extractive and abstractive methods. For example, Liu and Lapata (2019a) proposed a two-stage approach where an extractor first identifies relevant content, which is then paraphrased by an abstractor to produce the final summary.

The rapid development of large language models (LLMs) has opened new avenues for MDS research. These models have shown remarkable capabilities in understanding and generating text, which could potentially be leveraged for more sophisticated MDS systems. However, the application of these models to MDS also brings new challenges, such as ensuring factual consistency and managing computational resources.

As the field of MDS continues to evolve, researchers are increasingly focusing on domain-specific applications and the integration of external knowledge to improve summary quality.

2.4 Challenges in Multi-Document Summarisation

Recent rapid advancements in Natural Language Processing (NLP) have transformed Multi-Document Summarisation (MDS) into a dynamic and rapidly evolving research field. However, the task of generating concise, informative, and coherent summaries from multiple documents still presents a unique set of challenges. These challenges are increased by the complexities of language, the diversity of document types, and the inherent subjectivity in defining what constitutes a "good" summary.

2.4.1 The Challenges Described

2.4.1.1 Information Overload and Redundancy

One of the main challenges in MDS is managing the sheer volume of information present in multiple documents. This often leads to redundancy, where multiple documents contain overlapping information. Carbonell and Goldstein (1998) introduced the Maximal Marginal Relevance (MMR) criterion to address this issue, balancing relevance and novelty in summary generation. More recent approaches, such as the work by (Liu and Lapata, 2019a), have used neural architectures to capture cross-document relationships and reduce redundancy. Their hierarchical transformer-based model demonstrated significant improvements over previous methods in handling redundant information across multiple documents.

Additionally, documents may present conflicting viewpoints or contradictory information, requiring the summarisation system to reconcile these differences or present a balanced perspective. In SummEval, Fabbri *et al.* (2021) dealt with this challenge by introducing a multi-agent pointer-generator network that can effectively synthesise information from multiple, potentially contradictory sources.

2.4.1.2 Coherence and Structure

Maintaining coherence across multiple documents poses challenges too, particularly in preserving thematic consistency and temporal or causal relationships between concepts addressed in the documents. Wan (2008) proposed a graph-based method to improve multi-document summary coherence by considering both intra-document and inter-document relationships. More recently, Coavoux, Elsahar and Gallé (2019)introduced a neural approach that explicitly models discourse structure to generate more coherent summaries.

The preservation of temporal and causal relationships is very important for accurate summarisation, especially in news or historical document summarisation. Gholipour Ghalandari and Ifrim (2020) addressed this by developing a timeline-aware neural model that

can capture temporal dependencies across documents, showing improvements in summary quality for time-sensitive topics.

2.4.1.3 Context Preservation

Preserving context while condensing information is a delicate balance, with the risk of oversimplifying complex ideas or losing subtle but important details. This challenge is particularly high when summarising specialised documents such as scientific papers or legal documents. Cohan *et al.* (2018) proposed a discourse-aware attention model specifically designed for summarising scientific documents, which outperformed general-purpose summarisation models on scientific article datasets.

In the legal domain, Zhong *et al.* (2019) discusses a novel approach to legal document summarisation using iterative masking which could effectively summarise legal case documents while preserving critical legal terminology and concepts. Iterative masking is a technique where parts of the input text are systematically hidden or "masked" in successive rounds, allowing the model to focus on different aspects of the document in each iteration, thereby improving its ability to identify and extract the most salient information for summarisation. Their approach demonstrated the importance of domain-specific adaptations in MDS systems.

2.4.1.4 Evaluation Complexity

Assessing the quality of multi-document summaries presents unique challenges due to the subjectivity involved and the limitations of automatic metrics. While metrics like ROUGE (Lin, 2004) are widely used, they may not fully capture the nuances of summary quality, especially for abstractive summaries. Recognising these limitations, Fabbri *et al.* (2021) introduced SUMMEVAL (as mentioned earlier), a comprehensive framework for summary evaluation that combines multiple automatic metrics with human judgments.

The challenges in evaluation have led to increased interest in human evaluation methods. However, as Kryscinski *et al.* (2020) pointed out, human evaluation can be inconsistent due to individual biases and interpretations. They therefore proposed a unified framework for human evaluation of summarisation, aiming to standardise the process and improve reliability.

2.4.1.5 Scalability and Efficiency

As the volume of data grows, scalability becomes a significant concern in MDS. Processing large document sets requires substantial computational power, particularly for neural network-based approaches. Liu and Lapata (2019a) addressed this by introducing an efficient transformer-based model for MDS that could handle larger input sizes than previous approaches.

For applications requiring quick summarisation, such as news aggregation, balancing speed and quality becomes essential. In early work, Zhang *et al.* (2013) proposed a real-time multi-document summarisation system for twitter topic summarisation, demonstrating the feasibility of MDS in time-sensitive scenarios.

2.4.1.6 Linguistic and Structural Diversity

The variety in document types and linguistic styles presents additional challenges in MDS. Multi-lingual summarisation, in particular, adds layers of complexity. Lample and Conneau (2019) introduced a cross-lingual summarisation framework that can generate summaries in a target language different from the source documents, addressing the growing need for multilingual information synthesis.

Handling different document formats requires flexible and adaptable summarisation techniques. As an example, Zhong *et al.* (2020) developed a discourse-aware neural summarisation model that can effectively handle various document structures, from academic papers to news articles, by explicitly modelling document discourse.

2.4.1.7 Ethical Considerations and Bias

The ethical implications of MDS are increasingly becoming evident, particularly regarding bias mitigation and transparency. Summarisation systems may inadvertently amplify biases present in the source documents or introduce new biases. Celikyilmaz, Clark and Gao (2021) addressed this issue by proposing a framework for evaluating and mitigating gender bias in abstractive text summarisation.

Ensuring that the summarisation process is transparent and the results are explainable is very important, especially in sensitive areas. For example, DeYoung *et al.* (2020) introduced ERASER, a benchmark for evaluating rationales in NLP models, including summarisation, which can help in improving the explainability of MDS systems.

2.4.2 Pre-trained language models such as BERT, RoBERTa, GPT-3/4, and T5

In recent years (and leading into the development of the latest Large Language Models), the use of pre-trained language models, such as BERT, RoBERTa, BART, GPT-3/4, and T5, have improved numerous tasks in natural language processing (NLP), including multi-document summarisation. These models are trained on a large, often specialist, corpus of text data and learn to predict a word (or a sequence of words) based on its context. This pre-training step

allows the models to learn a rich representation of language, capturing various linguistic patterns and structures. All these models mentioned below have shown great promise in a variety of NLP tasks, including multi-document summarisation. However, their effective use in the specific context of scientific paper summarisation needs more careful exploration of their strengths and limitations, as well as the unique challenges posed by this task. This will be further explored in later parts of this chapter.

BERT (Bidirectional Encoder Representations from Transformers), developed by Devlin *et al.* (2019), is a transformer-based model that is pre-trained on a large corpus of text from the Internet. Unlike previous models, BERT is trained in a bidirectional manner, allowing it to understand the context of a word from both its left and right. This makes BERT particularly effective at tasks requiring an understanding of the context, such as multi-document summarisation.

RoBERTa (A Robustly Optimised BERT Pretraining Approach) is a variant of BERT that was introduced by Liu *et al.* (2019). RoBERTa modifies the pre-training process of BERT, including training the model on a larger amount of data, using a larger batch size, and removing the next sentence prediction objective, which leads to improved performance.

BART (Bidirectional and Auto-Regressive Transformers) is another variant of the transformer model that combines the strengths of BERT and GPT, and has demonstrated strong performance in text generation tasks such as summarisation. BART, introduced by Lewis *et al.* (2020), is pre-trained by corrupting text with an arbitrary noising function and learning to reconstruct the original text. It uses a standard transformer-based neural machine translation architecture which, unlike many models, is initialised with a cross-entropy loss that allows it to be parallelised across multiple GPUs and across numerous documents.

This has the effect that BART is especially suitable for tasks that require understanding the document context and generating fluent, coherent text. In the context of multi-document summarisation of scientific papers, BART's ability to encode an entire document context and generate meaningful, coherent summaries can be highly valuable. However, like other pre-trained language models, careful fine-tuning and adaptation are necessary to handle the unique challenges of scientific paper summarisation effectively.

GPT-3/4 (Generative Pretrained Transformer), introduced by OpenAI (OpenAI *et al.*, 2024), is another transformer-based language model. Unlike BERT and RoBERTa, GPT-3/4 is trained in an autoregressive manner, predicting each word in a sequence based on its preceding words. This allows GPT-3/4 to generate more coherent and fluent text, which can be highly beneficial for tasks like summarisation.

T5 (Text-to-Text Transfer Transformer), introduced by Raffel *et al* (2020), is a versatile transformer-based model that can be trained to perform any text-to-text task. T5 views every NLP task as a text generation task, enabling it to perform a wide range of tasks, including summarisation, translation, and question answering.

2.4.3: Hybrid techniques in Multi-Document Summarisation

Hybrid techniques in multi-document summarisation (MDS) represent a convergence of several strategies that aim to tackle the challenges in MDS from different angles. These techniques often combine traditional methods, such as text ranking, graph networks and topic modelling, with the advanced capabilities of large language models (LLMs).

The advantage of these so-called hybrid techniques lies in their ability to take advantage of the strengths of different approaches while mitigating their weaknesses. For example, text ranking can be used to identify key sentences or passages in a collection of documents, which can then be fed into an LLM for generating a more coherent and concise summary. Similarly, topic modelling or knowledge graph analysis can help in understanding the main themes across multiple documents, providing an additional layer of context for the LLM to generate more meaningful summaries.

The successful application of hybrid techniques in MDS requires a deep understanding of both traditional summarisation methods and LLMs, and the ability to effectively integrate them. The following sections will provide a more detailed review of some of these techniques and discuss their potential for improving the performance of MDS frameworks. Further discussion follows in Chapter 3, with an wider exploration of the tools and techniques used in developing 'modern' LLM applications.

2.4.3.1 Text Ranking

Text ranking is a fundamental task in many NLP applications, including Multi-Document Summarisation (MDS). It involves assigning importance scores to sentences or paragraphs

based on criteria such as relevance to a particular topic or overall informativeness. While traditional techniques like Term Frequency-Inverse Document Frequency (TF-IDF), PageRank, and LexRank continue to be relevant, recent advancements have led to more sophisticated approaches.

TF-IDF continues to be a baseline for many text ranking tasks. Its effectiveness in identifying key terms has been demonstrated in various summarisation contexts (Bano *et al.*, 2018). However, recent work has shown that combining TF-IDF with neural networks can yield improved results. For instance, Rossiello, Basile and Semeraro (2017) proposed a neural attention-based model that leverages TF-IDF features, demonstrating superior performance in extractive summarisation tasks.

Graph-based ranking algorithms, such as PageRank and LexRank, have evolved to incorporate more complex linguistic features. TextRank, an adaptation of PageRank for text processing tasks, has seen continued development. Mihalcea and Tarau (2004) initially introduced TextRank for keyword extraction and sentence extraction, and subsequent research has expanded its applications. More recently, Mallick *et al.* (2018) proposed Graph-based Unified Model (GUM), which integrates TextRank with semantic similarity measures to improve multi-document summarisation.

LexRank, which extends PageRank by considering cosine similarity of TF-IDF vectors, has also seen advancements. Erkan and Radev (2004) originally introduced LexRank for text summarisation, and it has since been adapted for various summarisation tasks. A notable extension is RankSum, proposed by (Joshi *et al.*, 2022), which incorporates continuous semantic spaces to capture more nuanced relationships between sentences.

In the context of MDS, these text ranking techniques are increasingly being used in conjunction with deep learning models. For example in the work of Zhong *et al.* (2020), the research team introduced a neural extractive summarisation model that combines graph-based ranking with BERT embeddings, demonstrating state-of-the-art performance on multi-document summarisation tasks.

In addition, the integration of text ranking with Large Language Models (LLMs) is opening new avenues for MDS. Li *et al.* (2023) proposed a framework that uses text ranking to guide the generation process of LLMs through a technique they term ‘stimulus prompting, where a small

model is used to fine-tune a prompt for a larger model. This results in more focused and relevant summaries. The approach leverages the strengths of both traditional ranking methods and the advanced language understanding capabilities of LLMs.

2.4.3.2: Topic Modelling

Topic modelling is a statistical approach for discovering abstract themes within a collection of documents. In the context of Multi-Document Summarisation (MDS), topic modelling serves as a powerful tool for identifying key themes across multiple texts, guiding the summarisation process towards more comprehensive and thematically coherent outputs.

Latent Dirichlet Allocation (LDA), introduced by Blei, Ng and Jordan (2003), is still a core topic modelling technique. LDA represents documents as mixtures of topics, where each topic is characterised by a distribution over words. This approach has been widely applied in MDS to identify overarching themes across document sets. For instance, Gong and Liu (2001) demonstrated the effectiveness of topic-based summarisation by using LDA to extract key sentences that best represent the main topics.

Recent advancements have seen the integration of neural networks with topic modelling. Srivastava and Sutton (2017) proposed Autoencoded Variational Inference For Topic Models (AVITM), which use variational autoencoders to learn topic distributions. This approach offers more flexibility and potentially better performance than traditional probabilistic topic models, especially when dealing with large and diverse document sets typical in MDS tasks.

The combination of Large Language Models (LLMs) with text ranking and topic modelling techniques is an interesting area for future development of hybrid MDS techniques. This integration uses the strengths of each approach to produce summaries that are both informative and well-written.

Text ranking techniques, such as those discussed in the previous section, can identify the most salient sentences or passages across multiple documents. Topic modelling enhances this process by revealing the main themes, guiding the selection and aggregation of content for the summary. For example, Ma *et al.* (2024) proposed a multi-document summarisation method that uses topic modelling to cluster similar content and guide the extraction of key information, which is then refined using a language model.

Once key content has been identified through text ranking and topic modelling, LLMs can be employed to generate a summary that is not only informative but also coherent and fluent. The advanced language understanding and generation capabilities of LLMs ensure that the summary maintains the original meaning and context of the source content while being accessible to readers. As mentioned earlier, Li *et al.* (2023) demonstrated this approach by using a large and smaller LLM together to generate abstractive summaries guided by topic-aware extractive content, resulting in summaries that effectively captured main themes while maintaining readability.

However, challenges remain in effectively combining these techniques. Ensuring that the LLM-generated summary accurately reflects the topics and key information identified in earlier stages is very important. Xu *et al.* (2020) addressed this by proposing a reinforcement learning framework that optimises the LLM's output based on topic coherence and information coverage metrics.

2.5 Applying pre-trained language models to the summarisation problem

Pre-trained language models (PLMs) have demonstrated significant utility in various natural language processing (NLP) tasks, including text summarisation. These models, trained on vast corpora of textual data, have developed rich representations of language that capture both syntactical structures and semantic meanings. This capability makes them particularly suitable for generating coherent and meaningful summaries (Liu and Lapata, 2019b).

A key advantage of using PLMs in text summarisation is their ability to generate fluent, human-like text. Summaries produced by these models often maintain a high degree of readability, enhancing their accessibility and utility for end users. Furthermore, the extensive pre-training of these models on diverse text data equips them with broad world knowledge, potentially leading to summaries that are not only accurate but also insightful (Lewis *et al.*, 2020). What PLMs do may appear almost magical (and indeed they are often treated as a black box) but it is important to understand that they produce fluent text through their advanced understanding of language patterns and structures acquired during their extensive pretraining phase. This pretraining involves exposure to vast amounts of text data, allowing the models to learn the intricacies of grammar, syntax, and semantic relationships. When generating summaries, PLMs use this learned knowledge to construct coherent sentences and paragraphs that flow naturally. They can maintain context over long sequences of text, ensuring that ideas are connected logically and that the overall narrative remains consistent. In addition, PLMs can adapt to different writing styles and tones, further enhancing the human-like quality of their output. To the end-user, the impact is that this enables them to produce summaries that read as if they were written by a skilled human writer.

PLMs offer a degree of flexibility often lacking in traditional extractive summarisation techniques. Their text generation capabilities enable the creation of summaries that go beyond merely reusing existing text from source documents. This allows for paraphrasing, rephrasing, or even commenting on the original text, providing a level of abstraction important for distilling complex documents into concise summaries (J. Zhang *et al.*, 2020).

However, despite these advantages, the application of PLMs in summarisation still has some challenges. Their effectiveness can vary significantly depending on the task nature and specific summarisation requirements. Recent studies exploring the use of PLMs for summarisation and other NLP tasks, along with their effectiveness and future potential in multi-document

summarisation (MDS), will be discussed in subsequent sections of this chapter and in Chapter 3.

2.5.1 Outline of Studies Utilising LLMs for Summarisation Tasks

While pre-trained language models have shown some effectiveness in single-document summarisation, their application to multi-document summarisation (MDS) has some unique challenges and complexities.

As described earlier, a primary challenge in MDS is the need to process and integrate information from multiple documents. This requires the summarisation model not only to generate accurate and coherent summaries but also to identify and handle overlapping, redundant, or conflicting information across documents (Lample and Conneau, 2019).

Existing pre-trained models such as BERT, RoBERTa, and GPT-3 have demonstrated capability in understanding individual document semantics and generating meaningful summaries. However, their usefulness in handling information from multiple documents remains an active area of research (Lim and Song, 2023).

One promising approach to addressing the challenges of MDS is the integration of advanced learning techniques with pre-trained language models. For instance, Liu and Liu (2021) proposed a contrastive learning framework for abstractive summarisation that can be applied to multi-document scenarios. Their method, SimCLS, demonstrates how contrasting different candidate summaries can lead to improved summarisation quality, outperforming other fine-tuning approaches on various datasets. The application of transformer-based models like BART and T5 to MDS has also shown promising results. These models' inherent text generation abilities make them well-suited for abstractive MDS. A study by Dou *et al.* (2021) presented a BART-based MDS model that outperformed previous models on several benchmark datasets. Nevertheless, the optimal utilisation of pre-trained language models for MDS remains an open question. Further research is needed to develop strategies for integrating these models with techniques such as clustering, topic modelling, and text ranking to effectively manage the complexity of MDS.

2.6 Domain-Specific Summarisation: Focusing on Scientific Papers

As discussed earlier, summarising scientific papers presents both unique challenges and advantages compared to summarising other types of documents. The advantages mainly arise from the structured nature of scientific papers, which tend to follow a standardised format consisting of sections such as the abstract, introduction, methodology, results, discussion, and conclusion (Mensh and Kording, 2017).

This format can facilitate the extraction of relevant information from specific sections, thus aiding in the summarisation process. Oh, Nam and Zhu (2022) proposed a structured abstract summarisation method using the IMRaD format (Introduction, Methods, Results, Discussion) to balance the emphasis on each section, improving the overall summary quality. Conroy and Davis (2018) introduced section mixture models for summarising scientific documents, which estimate term weights and optimise sentence extraction for comprehensive coverage. Rai *et al.* (2021) demonstrated a focused summarisation framework that ensures essential scientific content is included, particularly useful in large repositories like CORD-19 (a large repository of COVID-19 relevant literature).

The challenges of summarising scientific papers are numerous and complex. One main difficulty is the specialised language and complex concepts commonly found in scientific literature, which include terminologies, symbols, equations, and extensive references to other studies. Summarisation models not specifically trained on scientific texts often struggle with these complexities (Altmami and Menai, 2018). Also, summarising scientific papers involves not only extracting factual content but also understanding and conveying the paper's contributions to the existing body of knowledge. This requires a deep semantic understanding to grasp complex arguments, compare results with prior studies, and identify the novel contribution of the research (P. Wang *et al.*, 2023). The interdisciplinary nature of many scientific papers further complicates the process, as models need to handle diverse subjects and their interconnections (Oh, Nam and Zhu, 2022). Lastly, the structural features of scientific documents, such as the relationships between different sections, are often underused in summarisation models, leading to poorer summaries (Zhao, Yang and Cai, 2022).

2.6.1 Review of previous work on summarising scientific literature

Over the years, there has been considerable effort in the research community to develop effective techniques for summarising scientific literature. The path to optimise these techniques has seen the transition from basic sentence extraction methods to more sophisticated neural network architectures and pre-trained language models.

One of the earlier notable works in this area was by Teufel and Moens (2002). They proposed a sentence extraction method for summarising scientific papers. The focus was on sentences that convey the rhetorical status (referring to the role or function that a particular piece of information plays within the structure and argument of a document. It is a very important aspect of understanding not just what information is presented, but why it is presented and how it contributes to the overall message or argument of the text. of information), a vital aspect in understanding the crux of the paper.

As research progressed, the focus shifted towards machine learning and natural language processing techniques for summarising scientific papers. The development of LSTM (Long Short-Term Memory) models provided a new direction for these tasks. LSTM is a type of recurrent neural network that has the ability to learn long-term dependencies in text, making them suitable for text summarisation tasks (Bedi, Bala and Sharma, 2023). However, a key limitation of LSTM-based models is their inability to effectively handle long sequences due to the vanishing gradient problem (the vanishing gradient problem is a difficulty encountered when training artificial neural networks, particularly those using backpropagation and gradient-based optimisation methods, such as LSTMs). Additionally, they process sequences step-by-step, which can be computationally expensive for large documents.

To overcome some of the shortcomings of LSTM models, attention mechanisms were introduced. These allowed the model to focus on specific parts of the input sequence that are relevant for each step of the output sequence, which significantly improved the performance of neural network models in summarisation tasks (Bahdanau, Cho and Bengio, 2016). However, a common criticism of such models is that they often appear as black boxes, making it hard to interpret how they make their decisions. The attention mechanism and how this related to the LSTM is explained further in Appendix 1.

Advancements in transformer-based models, specifically the development of BERT (Bidirectional Encoder Representations from Transformers), made rapid contributions to the field of text summarisation. As noted earlier, BERT, proposed by Devlin *et al.* (2019), uses a bidirectional transformer, allowing it to understand the context of a word based on all of its surroundings (left and right of the word). This marked a significant improvement over previous models, which viewed the context in one direction (either left to right or right to left).

Expanding upon the benefits of BERT, (Beltagy, Lo and Cohan, 2019) developed SciBERT, a variant of BERT specifically trained on scientific literature. SciBERT demonstrated improved effectiveness in several tasks, including summarisation, demonstrating the importance of domain-specific pre-training and how it can improve performance in particular tasks.

Following this, BART (Bidirectional and Auto-Regressive Transformers), a variant of BERT that is specifically designed for text generation tasks, has shown promise in summarising scientific literature. BART, proposed by Lewis *et al.* (2020), is pre-trained by auto-encoding the text and has shown improved performance in abstractive summarisation tasks.
Work of Yasunaga *et al.* (2017) is worth a special mention for their interesting work in the multi-document summarisation of scientific articles. They developed a model that combined extractive and abstractive methods using graph neural networks. This provided a significant leap in dealing with the complexity of multiple scientific articles.

2.6.2 Gaps and potential areas for improvement in the current methodologies

Despite the considerable recent progress made in the summarisation of scientific literature, there still exists several prominent gaps and opportunities for further exploration and refinement in the methodologies employed. These gaps can be considered in the context of the research questions identified for this study.

While models such as BERT and BART have shown promise in the task of single-document summarisation, their potential in the context of multi-document summarisation still remains underexplored. Specifically, in the context of scientific papers, these models often struggle with the integration and synthesis of information from multiple documents and to reflect the relative importance of themes and topics across documents. Further research is required to understand how these state-of-the-art models can be effectively adapted and fine-tuned for the task of multi-document summarisation of scientific papers (Fabbri *et al.*, 2019).

The current body of research lacks a detailed exploration of the specific features and characteristics that make an abstractive multi-document summarisation framework efficient for scientific papers. Additionally, newer hybrid techniques, such as Retrieval Augmented Generation, have not been well-integrated into the existing frameworks. These techniques can identify sections of interest and intelligently incorporate them into the summaries, potentially greatly enhancing the quality of summaries by focusing on the most critical aspects of the documents (Huang *et al.*, 2020).

Finally, most existing studies have focused on proposing novel methodologies and demonstrating their effectiveness using limited datasets and evaluation metrics. There is still a lack of comprehensive comparative studies that analyse the performance of the frameworks against existing approaches, both extractive and abstractive. his gap prevents a complete understanding of the strengths and weaknesses of the different approaches and hinders the identification of best practices for the task. Evaluations need to be conducted on diverse scientific datasets to ensure the generalisability of the findings (Otterbacher, Erkan and Radev, 2005).

2.7 Application of Advanced NLP techniques to MDS

The ability to effectively summarise multiple documents is a complex problem. As such, and in terms of the neural network models themselves, researchers have employed a range of advanced techniques in Natural Language Processing (NLP) to deal with it. These techniques include transfer learning, attention mechanisms, graph-based methods, and domain adaptation. Considering each one in turn and its relevance to MDS:

The basic concept of **transfer learning** is quite straightforward: it involves taking what a model has learned from one problem and applying it to another related problem. This approach is particularly useful in natural language processing (NLP), where specialised training data can be scarce and computational resources limited. Transfer learning supports leveraging large pre-trained models like BERT, RoBERTa, or T5, which have already been trained on extensive corpora of text. These models are then fine-tuned for specific tasks, such as multi-document summarisation (MDS) (Devlin *et al.*, 2019; Liu *et al.*, 2019; Raffel *et al.*, 2020).

Transfer learning is useful because these pre-trained models already come with a rich understanding of language nuances, including syntax, semantics, and context, which significantly enhances the performance of downstream tasks. For instance, BERT (Bidirectional Encoder Representations from Transformers) uses a bidirectional approach to understand the context of a word based on all of its surroundings, both left and right (Devlin *et al.*, 2019). This deep contextual understanding helps in generating more accurate and coherent summaries when the model is fine-tuned on specific summarisation tasks.

RoBERTa (Robustly Optimized BERT Pretraining Approach) builds on BERT by optimising its training procedure, using more data and computational resources to improve performance (Liu *et al.*, 2019). This enhancement allows RoBERTa to achieve better results in various NLP tasks, including summarisation, by providing a more refined understanding of language structures and relationships.

T5 (Text-To-Text Transfer Transformer) takes a different approach by framing all NLP tasks as a text-to-text problem (Raffel *et al.*, 2020). This unified framework simplifies the training process and allows for more flexibility in applying the model to various tasks, including MDS. T5's ability to handle different tasks under a single framework makes it a powerful tool for generating high-quality summaries.

Additionally, the application of transfer learning to MDS benefits from the extensive pre-training of these models, which already includes exposure to diverse text forms and topics. This exposure enables the models to generalise better to new, unseen data, producing summaries that are not only more accurate but also contextually rich and relevant.

The **attention mechanism** was introduced in NLP as a part of the neural machine translation models to help the model *focus* on different parts of the input sequence when generating each word in the output sequence (Bahdanau, Cho and Bengio, 2016). The concept was inspired by the way humans pay selective attention to different parts of the input when processing information or making decisions. In the context of MDS, attention mechanisms have been used to determine the importance of different parts of the source documents when generating a summary. This is important in MDS because the documents often contain redundant information, and the goal of the summary is to present the key points concisely without extra information.

The attention mechanism works by assigning a weight to each part of the source documents. This weight can be determined by the context and the relationship between different parts of the documents. For instance, sentences or sections that are referred to frequently in the document or that contain key points are likely to be given higher attention weights.

The attention weights are then used to generate a *context vector* for each part of the summary. The context vector is a weighted sum of the source document embeddings, with the weights given by the attention mechanism. This context vector effectively represents the parts of the source documents that the model should 'pay attention to' when generating each part of the summary. See Appendix 1 for a more detailed explanation of the attention mechanism and the impact on summarisation.

The use of attention mechanisms has been demonstrated to significantly improve the quality of the generated summaries in MDS (Vaswani *et al.*, 2017). By focusing on the important parts of the source documents, the attention mechanism helps the model manage the complexity and redundancy of the information, leading to more concise and coherent summaries.

It is worth noting, however, that while attention mechanisms have proven beneficial in MDS, there are still challenges to be addressed. One of the main challenges is determining the optimal way to assign attention weights in the context of MDS. This remains an area of ongoing

research, with new techniques and approaches being developed to further improve the effectiveness of attention mechanisms in MDS.

Graph-based methods are a set of techniques that represent documents as graphs. Each node in the graph represents a sentence or a paragraph, and the edges denote the semantic relationship between them. Algorithms like PageRank can then be applied to these graphs to identify the most important nodes (i.e., the most important sentences or paragraphs) for inclusion in the summary (Erkan and Radev, 2004). Graph-based methods provide a useful way to visualise and manage the relationships between different parts of the source documents.

Domain adaptation involves adjusting a model that has been trained on one domain (or type of data) to perform well on a different, but related, domain. This technique is particularly relevant in MDS, where a model may need to summarise documents from various fields of study or topics. Domain adaptation methods can be used to fine-tune pre-trained models to better understand the specific language, concepts, and structures used in different domains, thereby improving the quality of the generated summaries (Gururangan *et al.*, 2020).

2.7.1 Review of Studies Applying These Techniques in MDS

As the wider domain of Natural Language Processing (NLP) has developed, the way that emerging techniques are applied in multi-document summarisation (MDS) has seen increased research attention. This section focuses on studies that have used transfer learning, attention mechanisms, graph-based methods, and domain adaptation in the context of MDS.

2.7.1.1 Transfer Learning in MDS

Transfer learning has revolutionised the application of deep learning models in NLP, and MDS has been no exception to this trend. With the release of transformer-based models, which have achieved state-of-the-art results across a wide range of NLP tasks, their application in MDS has been a natural extension.

Foundations of Transfer Learning

The fundamental principle behind transfer learning is not new. It derived from the cognitive sciences, where the idea is that learning in one context or task can be beneficial for performance in another context or task (this can be observed in our day to day lives- in the development of transferable skills). In the machine learning landscape, Pan and Yang (2009) provided an extensive overview of transfer learning, describing its significance in scenarios

where the target task has limited labelled data. The point is to leverage knowledge from a source task, where ample data might exist, and adapt it for the target task.
Consider the scenario of transferring sentiment analysis capabilities from product reviews to film reviews. In this case, a model trained on a large dataset of e-commerce product reviews serves as the starting point, having learnt to classify text as positive, negative or neutral. This source task provides the model with a robust understanding of sentiment-related vocabulary and sentence structures. To adapt this knowledge to the target task of analysing film reviews, the model undergoes fine-tuning using a limited set of labelled film reviews. During this process, the model adjusts its understanding to the new domain, learning film-specific vocabulary (e.g., "Oscar-worthy", "plot twist") and adapting to the different writing styles typically found in cinema critiques. The model leverages its foundational knowledge of sentiment analysis from product reviews, allowing it to quickly adapt to the nuances of film reviews without requiring a large, domain-specific dataset. This transfer of knowledge enables the creation of an effective film review sentiment analyser with significantly less labelled data than would be needed to train a model from scratch.

2.7.1.2 Transformers and Pre-trained Models

The domain of NLP changed with the development of transformer architectures. Vaswani et al. (2017) introduced the Transformer model, which applied self-attention mechanisms to process input data in parallel, instead of the sequential processing in previous models like RNNs and LSTMs (lack of attention, up until this point, had been a major disadvantage of the LSTM model which had been an original focus for the development of this work). The Transformer's architecture was particularly applicable to the task of language modelling, leading to the development of several large-scale pre-trained models.

BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) was one of the first transformer-based models that showcased the power of pre-training on vast corpora. BERT's design allowed it to understand context from both the left and right side of a token in any input sentence, making it particularly robust for many NLP tasks.

This robustness of BERT was explored for MDS. Liu et al. (2019) harnessed the capabilities of BERT for extractive summarisation. The model, after being pre-trained on a large corpus, was fine-tuned on the task of selecting relevant sentences from multiple documents. The results showed the ability of the model to discern relevant information across documents, capturing the essence required for a concise summary.

RoBERTa, proposed by Liu et al. (2019), further built upon BERT by refining the pre-training process. It demonstrated that with more data, larger batch sizes, and longer training, the performance across various NLP tasks, including MDS, could be enhanced. Several works have since fine-tuned RoBERTa for MDS, noting a marked improvement in the quality of generated summaries.

Another noteworthy model in this domain is T5 (Text-to-Text Transfer Transformer) by Raffel et al. (2019). Instead of designing a model specifically for each NLP task, T5 was designed to treat every problem as a text-to-text problem. Summarisation, translation, question-answering, and even classification can be framed in this paradigm. For MDS, researchers fine-tuned T5 to generate concise summaries from multiple input documents, benefiting from the vast knowledge the model gained during its pre-training phase.

Benefits and Challenges in MDS

The primary benefit of using transfer learning, especially with transformer-based models, in MDS is the ability to leverage vast linguistic knowledge without the need for extensive labelled data in the summarisation task. This is particularly useful, given that creating labelled datasets for summarisation can be laborious and resource intensive. Indeed in some domains that data may simply not exist.

However, transfer learning in MDS is not without its challenges. One of the key concerns is the domain discrepancy. A model pre-trained on a general corpus might not capture the nuances of specific domains, such as medical or legal documents. While fine-tuning can address this to some extent, it still requires access to domain-specific data.

Moreover, there is the computational challenge. Models like BERT and RoBERTa have hundreds of millions of parameters. Fine-tuning them requires substantial computational resources. While this might be feasible for large research institutions or corporations, it could be a barrier for individual researchers or smaller organisations. In this work, the author has been fortunate enough to have access to large-scale HPC (High Performance Computing) GPU resource both at his workplace and through the UK national JADE-2 and Bede research facilities.

Transfer learning, reinforced by the advent of transformer architectures, has provided a significant avenue for advancements in MDS. The ability to harness vast linguistic knowledge,

captured by models pre-trained on extensive corpora, and apply it to the summarisation task has been a game-changer. However, as with any technique, it is important to be mindful of its challenges and limitations. Future research will potentially continue to refine these approaches, making them even more effective and accessible for MDS.

2.7.1.3 Attention Mechanisms in MDS

The attention mechanism is probably one of the most influential advancements in recent NLP research. Originating in the domain of neural machine translation, *attention* sought to address the problem of long-sequence translations, enabling the model to "focus" on different parts of the input when generating the output. Its application to multi-document summarisation (MDS) has developed new methods to generate more coherent and contextually relevant summaries.

Origins and Evolution of Attention Mechanism

The initial idea of attention was introduced by Bahdanau et al. (2014) in the context of neural machine translation. Their model dynamically focused on different portions of the input sequence when producing the translation, so overcoming limitations of fixed-length context vectors that were used in earlier sequence-to-sequence models.

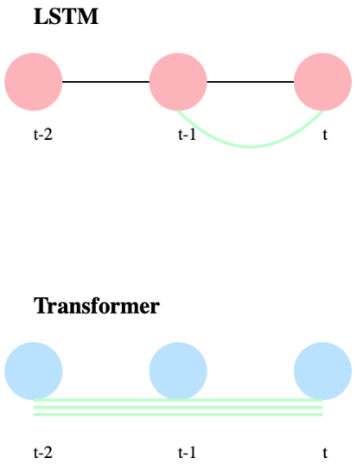

Figure 2: Attention Mechanism

Figure 2 (above) illustrates the key difference in attention mechanisms between LSTMs and Transformers. In LSTMs, represented by the pink circles, attention is limited and sequential. The current state (t) can only directly attend to the immediate previous state (t-1), as shown by the green curved line. Information from earlier states must pass through intermediary states, potentially losing information. In contrast, Transformers, depicted by blue circles, employ a more comprehensive attention mechanism. The multiple green lines connecting all states in the Transformer model represent its ability to attend directly to all parts of the input sequence simultaneously. This parallel processing allows Transformers to capture both short-term and

long-term dependencies more effectively, enabling them to maintain context over longer sequences and process information more efficiently than LSTMs.

Building upon this, Vaswani et al. (2017) introduced the Transformer architecture (described earlier), which was wholly based on self-attention mechanisms. Instead of relying on recurrent or convolutional layers, the Transformer used stacked self-attention layers to process input data, making it highly parallelisable and efficient. This architectural choice was foundational for many subsequent models in NLP, including those applied to MDS.

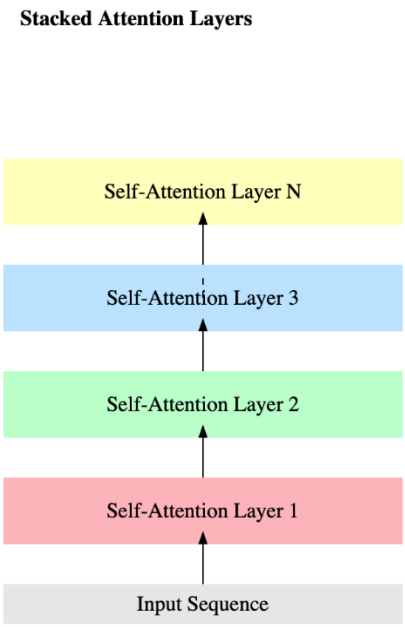

Figure 3: Stacked Attention Layers

As illustrated in the diagram above (figure 3), the process begins with an input sequence passing through multiple self-attention layers. Each layer calculates attention scores for every element in the sequence relative to all others, updating representations accordingly. This stacked structure allows the model to build increasingly abstract and complex representations of the input data. Unlike recurrent neural networks, all elements in a sequence can be processed in parallel within each layer, significantly improving computational efficiency. The arrows between layers represent information flow, with each element in a layer having access to all elements from the previous layer, enabling the capture of both local and global dependencies. This architecture can be scaled by adding more layers, allowing for deeper and more powerful models. The success of this stacked self-attention structure has led to its widespread adoption and adaptation in subsequent models.

Application to MDS

Given the ability of attention mechanisms to weigh the importance of different parts of the input data, it was a logical step to apply this to MDS, where the challenge lies in extracting salient information from multiple documents.

Zhou et al. (2018) made significant strides in this direction. Their model leveraged the Transformer's self-attention mechanism to weigh the relevance of sections across multiple documents. An essential aspect of their approach was the ability to capture cross-document relationships. This was required in those cases where documents had overlapping or complementary information. By effectively weighing these relationships, the model could generate summaries that were not just extracts of individual documents but a coherent fusion of information from all sources.

Zhang et al. (2019) took a novel approach by integrating the attention mechanism with reinforcement learning. In traditional extractive summarisation, the focus is on selecting the most relevant sentences. Zhang and colleagues added an additional layer of sophistication by also considering the sequence in which these sentences were presented in the summary. By doing so, their model could produce summaries that were not only information-rich but also had a logical flow, improving readability and comprehension.

The role of attention in handing redundancy

A unique challenge in MDS is the presence of redundant information. When multiple documents discuss similar topics or events, there's a high likelihood of repeated information. The attention mechanism, with its ability to weigh the importance of different sections, plays a pivotal role here. Models can assign lower weights to repeated or overlapping information, ensuring that the final summary is concise and devoid of unnecessary repetition.

Challenges and Future Directions

While the attention mechanism has significantly advanced the field of MDS, it is not without challenges. One of the main concerns is the interpretability of attention weights. While these weights provide a measure of importance, understanding why a model assigns a particular weight remains a complex task. Efforts towards making attention mechanisms more interpretable will be very important as these models find more real-world applications. Moreover, as the length, complexity and volume of documents increase, there's a growing need to develop more efficient and scalable attention mechanisms. Current models, especially

those based on the Transformer architecture, can be computationally intensive, posing challenges for real-time applications.

2.7.1.4 Graph-based Methods in MDS

Graph-based methods, which model documents as interconnected structures, present a compelling approach to multi-document summarisation (MDS). By representing documents as nodes within a graph and determining the relationships between these nodes based on content similarity, these methods offer a structured way to identify and extract pertinent information.

Foundations of Graph-based Approaches

The use of graph representations in MDS derives from the way information in documents is interrelated. Each node in the graph represents a sentence or a segment of text, and the edges (or connections) between these nodes indicate the degree of similarity or relatedness. The strength of these connections is typically determined using measures of semantic or content-based similarity.

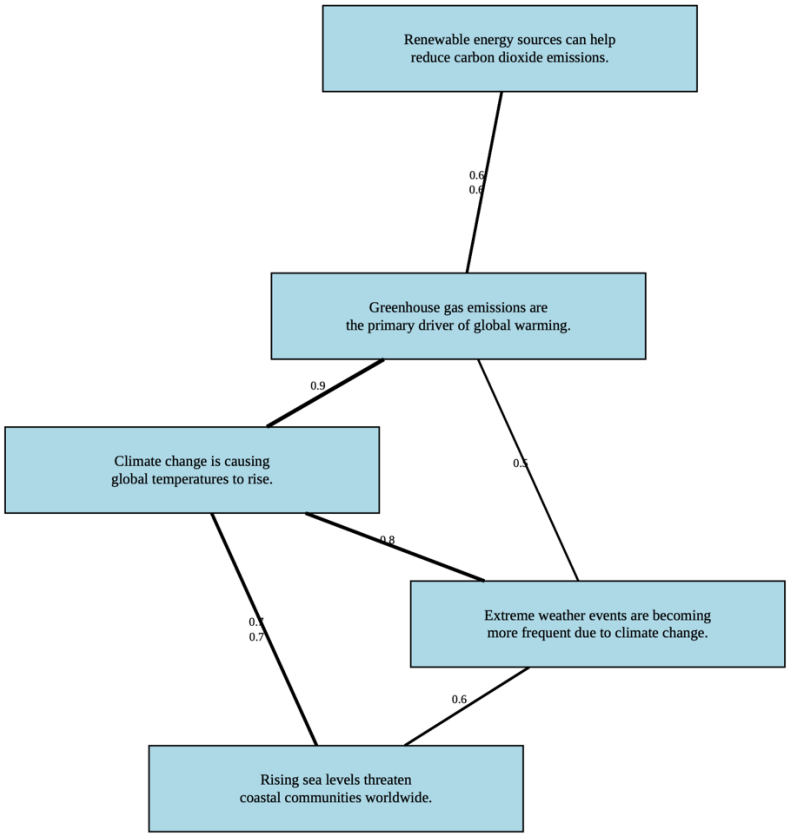

Figure 4: Graph based sentence relationships

This graph (figure 4) representation allows algorithms like LexRank or TextRank to calculate the centrality or importance of each sentence based on its connections to other sentences. Sentences with more and stronger connections (like the sentences *"Climate change is causing global temperatures to rise"* and *"Greenhouse gas emissions are the primary driver of global warming"* in this example) would likely be considered more important for the summary. In graph-based approaches to multi-document summarisation, sentences are represented as nodes in a graph structure, with edges between nodes indicating the relationships or similarities between sentences. These relationships are typically described using vector representations of the sentences, often derived from techniques like word embeddings or sentence encoders. The strength of the connection between two sentences (represented by the weight of the edge connecting their corresponding nodes – the number on the line in the example above) is usually determined by the cosine similarity or another similarity measure between their vector representations. This graph structure allows for the capturing of complex inter-sentence relationships across multiple documents, enabling algorithms to identify central or important sentences based on their connections within the broader context of the entire document set.

This approach helps identify key themes and central ideas across multiple documents on the same topic, making it effective for multi-document summarisation tasks.

Integration with Neural Methods

While the initial graph-based methods for MDS were not deep learning-based, the surge in neural network applications in NLP soon led to their integration with graph-based techniques. The idea was to combine the structured representation of graph-based methods with the powerful feature extraction capabilities of neural networks. Yasunaga et al. (2017) demonstrated a noteable integration of these two paradigms. Their model combined graph representations with recurrent neural networks (RNNs). At each iteration, the RNN updated the node (sentence) representations in the graph based on the information from its neighbours. This iterative process allowed the model to refine its understanding of each sentence's context, resulting in more accurate and coherent summaries.

Another significant advantage of this integration was the ability to handle large-scale document collections. Neural networks could efficiently process and update node representations, making it feasible to summarise extensive document sets.

Benefits and Challenges

Graph-based methods offer several benefits in MDS. They provide a structured way to analyse and summarise documents, ensuring that the extracted information is contextually relevant. The integration with neural methods further enhances their capability by enabling them to handle larger datasets and capture more intricate patterns in the data.

However, challenges remain. Determining the optimal graph structure and edge weights can be complex, especially when dealing with diverse document sets. Moreover, while neural integrations offer scalability, they also introduce significant computational overheads, especially with large graphs.

2.7.1.5 Domain Adaptation in MDS

Domain adaptation, within the scope of multi-document summarisation (MDS), addresses the challenges arising from the variability and diversity of document sources. Given that documents can span a wide range of topics, styles, and structures—from academic research papers to casual blog posts—the task of creating a summarisation model that works effectively across these domains presents a considerable challenge.

Understanding Domain Adaptation

Domain adaptation deals with the transfer of knowledge from a source domain, where lots of data is available, to a target domain, where data might be limited or inherently different. The primary goal is to use the knowledge acquired in the source domain to achieve better performance in the target domain, despite the differences between the two. Chu and Wang (2018) provide a comprehensive survey of domain adaptation techniques in the context of neural models for machine translation. Their work describes various strategies for adapting models, particularly when labelled data in the target domain is scarce. For MDS, this is especially pertinent. For instance, while lots of data might exist for summarising news articles, there may be limited data available for summarising specialised scientific papers. The techniques discussed by Chu and Wang suggest a framework to use what data is available to bridge this gap.

Multi-task Learning and Domain Adaptation

One of the more recent and promising approaches to domain adaptation is multi-task learning. The underlying idea is that by training a model on multiple tasks simultaneously, it can learn representations that are more general and less biased towards any single task. Raffel *et al.* (2020) made significant developments in this direction with the T5 (Text-to-Text Transfer

Transformer) model. T5 was trained on a large range of NLP tasks, including summarisation, translation, classification, and question answering, all contextualised as text-to-text problems. This unified approach allowed the model to learn rich, versatile representations that proved beneficial for various domain-specific tasks, including for MDS. When applied to summarisation tasks, T5 showed adaptability across different document domains, often requiring minimal fine-tuning. The performance of the model on MDS tasks showed the potential of multi-task learning in enhancing domain adaptation, particularly in scenarios where target domain data might be limited.

Implications for MDS

The introduction of domain adaptation techniques to MDS has made significant improvements to what is possible. As the volume and variety of digital content continue to grow, the ability to summarise content from diverse domains becomes increasingly important. Domain adaptation ensures that MDS models remain versatile and effective, regardless of the source of the documents. Furthermore, as MDS finds more real-world applications - from supporting researchers in literature reviews to helping journalists sift through enormous amounts of information - the importance of domain adaptation will only grow. By ensuring that summarisation models are adaptable and resilient to domain shifts, the quality and utility of generated summaries can be maintained.

2.8 Choosing and preparing data for MDS

2.8.1 Importance of dataset selection for MDS

Dataset selection is a key element of any machine learning or Natural Language Processing (NLP) task, but it has particular significance for multi-document summarisation (MDS). The quality, diversity, and representativeness of the dataset can significantly influence the success of summarisation models, demonstrating the critical importance of thorough consideration in dataset selection for MDS.

Variety and Complexity

The landscape of potential documents is enormous. From news articles and blogs discussing current events to intricate scientific papers delineating the latest research, each document type presents unique challenges and nuances.

News Articles: These documents are often time-sensitive, centred on current events, and can vary significantly in depth and perspective. While some articles provide a succinct account of events, others (and the source of the news article is important, consider the difference in style between a tabloid newspaper and a news agency) dive deeper, offering analyses or multiple viewpoints. Summarising these articles requires models to discern central events from ancillary details and potentially synthesise diverse viewpoints. As noted by Fabbri *et al.* (2019), news articles present unique challenges due to their temporal nature and the need to capture multiple perspectives.

Scientific Papers: These documents are typically dense, filled with domain-specific jargon, and structured to separate background information, methodologies, results, and conclusions. Summarising these documents demands a keen understanding of this structure and the ability to extract key findings without distorting their meaning. Cohan and Goharian (2015) highlighted the importance of considering citation contexts and discourse structure when summarising scientific articles.

Legal Documents: Characterised by formal language and stringent structures, legal documents like contracts or court rulings can be particularly challenging. Extracting the essence without misrepresenting legal clauses is therefore essential. Bhattacharya *et al.* (2019) discussed the unique challenges in summarising legal documents, emphasising the need for domain-specific approaches.

Literary Works: Fictional works, whether novels or short stories, present a different challenge. Here, the problem is not just about events but also themes, character developments, and narrative styles. As explored by Kazantseva and Szpakowicz (2010), summarising literary texts requires consideration of narrative structure and thematic elements.

Given this set of scenarios, choosing datasets that encompass this complexity and diversity becomes essential. For a model to be robust and adaptable in real-world scenarios, it needs exposure to varied document types during training, ensuring that it configured to handle diverse MDS tasks.

2.8.1.1 Domain-Specific Considerations

The need for domain specificity in MDS is therefore of paramount importance. Consider the difference between medical research papers and financial reports. While both are formal and structured, the former is full of medical terminologies, patient data, and experimental results, whereas the latter is full of economic indicators, financial jargon, and market analyses.

For MDS models to be effective then, they must be trained on domain-specific datasets that capture these nuances. A model trained predominantly on news articles might pe less performant when summarising a detailed medical study. On the other hand, a model well-trained in financial terminologies might misinterpret literary themes. Therefore, ensuring that the selected dataset aligns with the intended application domain of the MDS model is of critical importance. In the development of SciBERT as discussed earlier, Beltagy, Lo and Cohan (2019) demonstrated the importance of domain-specific language models for scientific text, showing the need for specialised datasets in technical domains.

Bias and Ethical Considerations

Dataset selection is not just a technical problem; it is also an ethical one. The data chosen to train models can inadvertently introduce biases, leading to skewed or prejudiced summaries. Potential issues with dataset bias are as follows:

1. **Representation Bias:** If a dataset predominantly contains articles from a particular region or represents a specific demographic, the resultant model might be biased towards that group's perspective, sidelining other voices. Mehrabi *et al.* (2021) give a comprehensive survey of bias and fairness in machine learning, showing the importance of having diverse and representative datasets.
2. **Content Bias:** Relying heavily on datasets with a particular stance (e.g., political or philosophical) can lead to models that echo that stance, reducing objectivity. As discussed by Olteanu *et al.* (2019), data collection strategies can significantly impact the biases present in datasets and, consequently, in the models trained on them.

3. **Temporal Bias:** Using outdated datasets might result in models that may not recognise contemporary terminologies or themes, reducing their effectiveness in current scenarios. Huang and Paul (2018) explored the impact of temporal effects on text classification, again showing the need for datasets that reflect current language and terminology usage.

2.8.2 Techniques for data pre-processing and formatting

Data pre-processing and formatting play a very important role in the success of multi-document summarisation (MDS) models. Before delving into model training, it's essential to ensure that the textual data is clean, structured, and represented in a way that maximises the model's understanding. Some of the main techniques involved in this preparatory phase are described below:

2.8.2.1 Text Cleaning

Textual data, especially when sourced from the web or user-generated platforms, can be messy. It may contain redundant or irrelevant information, errors, or inconsistencies that can hinder the performance of the model.

Removing Stop Words: Stop words like "and", "the", "is", etc., are frequent in text but often do not always have a lot of meaning in the context of MDS. Removing these can help reduce the dimensionality of the data without sacrificing important information. For instance, in the sentence "The cat sat on the mat", words like "the" and "on" can be omitted without compromising the primary message.

Punctuation Removal: Depending on the application, punctuations might not be necessary. For certain NLP tasks, removing punctuation can simplify the text and enhance processing speed.

Stemming and Lemmatisation: Both these techniques reduce words to their base or root form. For instance, "running", "runner", and "ran" might be reduced to the base "run". While stemming does this by heuristic chopping of word ends, lemmatisation uses vocabulary and morphological analysis.

Heuristic chopping is the process of reducing words down to their basic form by applying a set of predefined rules and heuristics. It involves removing common suffixes and prefixes based on patterns observed in the language, without considering the specific linguistic properties of the word or its context within a sentence. The intention is to simplify words to their stem, which

can help in a range of natural language processing (NLP) tasks by reducing the number of unique terms and focusing on the core meaning of words. For example, in English, heuristic chopping might involve rules such as:

- Removing common suffixes like "ing", "ed", "es", "s" (e.g., "running" becomes "run", "played" becomes "play").
- Trimming off prefixes like "un", "re", "in" (e.g., "unhappy" becomes "happy", "redo" becomes "do").

This process is called "heuristic" because it relies on general rules derived from experience and patterns rather than a deep linguistic analysis. It is often fast and computationally inexpensive but can be less accurate than more sophisticated methods.
In contrast, lemmatisation is a more advanced technique that reduces words to their base or dictionary form, known as a "lemma." Lemmatisation involves the use of vocabulary and morphological analysis to accurately identify the base form of a word based on its context and part of speech. This process requires a comprehensive understanding of the language, including its grammar and syntax.
For instance, the word "better" would be reduced to "good" by lemmatisation, considering its comparative form, whereas stemming might not handle this irregularity correctly. Similarly, "running" would be lemmatised to "run" by recognising it as a verb in its present participle form.

These normalisation processes help to treat different forms of a word as a single entity, improving model consistency. For example, the NLTK (Loper and Bird, 2002) and spaCy (Honnibal and Montani, 2017) libraries in Python provides both stemming and lemmatisation utilities for this purpose, amongst other standard NLP text cleaning routines such as case normalisation.

2.8.2.2 Segmentation and Tokenisation

Segmenting and tokenising are basic steps in converting raw text into structured data amenable to processing.
Sentence Segmentation: This involves breaking down a document into individual sentences. Tools like the Python Natural Language Toolkit (NLTK) or the Python library ‘spaCy’ can achieve this using pre-trained machine learning models and other computational techniques.
Tokenisation: Post-segmentation, sentences are further broken down into tokens, typically words, but in modern Large Language Models tokens are often smaller parts of speech. For the

sentence "Cats chase mice", tokenisation would yield ["Cats", "chase", "mice"]. Tokenisation not only helps to structure the data but is also a precursor step to more advanced pre-processing steps like embedding generation.

2.8.2.3 Feature Engineering

Turning raw text into a format that machine learning models can understand is often achieved through **feature engineering**. Together with TF-IDF (described earlier), other commonly used feature engineering processes are:

Word Embeddings: These are dense vector representations of words that capture their semantic meaning. Models such as Word2Vec (Mikolov *et al.*, 2013) or GloVe (Pennington, Socher and Manning, 2014) convert words into vectors based on their context within large corpora, resulting in semantically similar words having close vector representations. For instance, the vectors for "king" and "queen" might be closer than those for "king" and "apple", reflecting their semantic similarity.

Handling Multilingual Data: As global data increases, the need to process documents in multiple languages becomes increasingly important. Particularly with reference to the focus of this research (multi-document summarisation of scientific papers- in English), it is appreciated that knowledge exists in other languages and other cultures although that is not a focus for this study.

2.8.3 Quality and characteristics of suitable datasets for MDS

The quality and nature of the datasets used in multi-document summarisation (MDS) are very important elements that can determine the ultimate effectiveness of summarisation models. Some of the issues that impact dataset selection form MDS are discussed below:

2.8.3.1 Size and Representativeness

Dataset size and diversity are instrumental in shaping the robustness and versatility of MDS models.

Dataset Size: The benefits of large datasets in machine learning are well documented. A comprehensive dataset offers diverse examples, allowing models to discern intricate patterns. Banko and Brill (2001) provide a thorough examination of this, illustrating that model performance often improves with increased data, especially for tasks with high variability like MDS.

Representativeness: Beyond sheer volume, the representativeness of a dataset - its ability to mirror real-world scenarios - is of great importance. For instance, if an MDS model is designed for summarising academic papers across disciplines, a dataset restricted to physics articles would be insufficient. This imbalance can lead to models that are overly specialised and lack versatility.

Real-world Application: In practice, datasets that are both large and representative are ideal. For instance, summarising news articles from diverse global sources would necessitate a dataset that covers varied topics, regions, and writing styles. Such diversity ensures that the model is equipped to handle real-world summarisation tasks effectively.

2.8.3.2 Annotation Quality

Annotated data plays a very important role in supervised learning, serving as the core method for model training.

Importance of Quality Annotations: The accuracy and consistency of annotations directly influence model performance. Anglin, Boguslav and Hall (2022) provide a thorough examination of this, highlighting the pitfalls of poor annotations and the resultant challenges in model training and evaluation. Their work emphasises that high-quality annotations are critical

for developing robust and reliable machine learning models, as inaccuracies in the training data can lead to errors in model predictions and evaluations.

Challenges in Annotation: Acquiring quality annotations is a substantial effort. While crowdsourcing platforms like Amazon's Mechanical Turk offer scalable annotation solutions, they often necessitate rigorous post-annotation quality checks. For domain-specific tasks, expert annotators are indispensable, which can escalate costs and time commitments. These challenges are compounded by the need for consistent and accurate annotations across potentially large datasets.

Potential Solutions: Active learning, where models are initially trained on a small set of annotated data and then iteratively improved using their own predictions, can alleviate some annotation challenges. Such approaches, however, require careful monitoring to prevent model drift. This iterative process helps to optimise the annotation effort by focusing on the most informative examples, thereby improving the efficiency and effectiveness of the model training process.

2.8.3.2 Accessibility and Privacy Concerns

As data becomes the new oil (or perhaps the new renewable energy, might be a better analogy), its acquisition and usage come alongside significant new challenges and responsibilities.

Data Accessibility: While vast quantities of data are available online, acquiring them in usable formats can be challenging. Many valuable data sources, such as academic journals, are behind paywalls or have other IP (intellectual property) protection; data is just not all ‘free’. Even open-access sources can pose challenges, with anti-scraping mechanisms in place.

Privacy Concerns: The ethical and legal dimensions of data usage are gaining prominence. Regulations like GDPR underscore the importance of user consent and data anonymisation. But as Narayanan and Shmatikov (2008) point out, true anonymisation is challenging, and there is always a risk of de-anonymising data, so leading to privacy breaches. They demonstrated that supposedly anonymised datasets could often be de-anonymised, revealing the identities of individuals in the data. This study highlights the challenges of true anonymisation, showing that even when direct identifiers are removed, other indirect data points can be used to re-

identify individuals.

Implications for MDS: Given these constraints, MDS researchers need to tread carefully. Using proprietary or private data without requisite permissions can lead to legal complications and damage the credibility of the research.

2.8.3.3 Benchmark Datasets

Benchmark datasets serve as the gold standard, offering a consistent platform for model comparison and validation. Some of the datasets used in initial work in this domain are discussed below. As this is a rapidly emerging field, additional datasets suitable for MDS of scientific papers are further discussed in chapter 3.

DUC and TAC: Over the past several years, these conferences have set the benchmark (literally) for MDS research. The datasets released cater to varied MDS challenges, ranging from topic-focused summarisation to cross-document analysis. Dang's (2006) overview of DUC provides insights into the evolution of these challenges and the significance of the dataset.

CNN/Daily Mail Dataset: This dataset, with its real-world news articles, has become a standard in MDS research. Hermann's *et al.* (2015) introduction of this dataset marked a move towards more realistic summarisation tasks, moving away from artificially curated texts.

PubMed: The medical domain, with its vast repository of research articles, presents unique challenges for MDS. Lu (2011) explored these challenges, demonstrating the nuances of medical text summarisation and the importance of domain-specific datasets like PubMed.

2.9 Choice of evaluation metrics: assessing summarisation quality

As MDS models will attempt to distil large amounts of information into concise, coherent, and informative summaries, assessing the quality of the summaries is very important, ensuring that the essence of the original content is retained and conveyed. Without a robust evaluation mechanism, it becomes challenging to ascertain the effectiveness of an MDS model and, consequently, its applicability in real-world scenarios (Nenkova and McKeown, 2012).

However, the evaluation of text summaries itself presents a unique set of challenges. Unlike tasks with clear right or wrong answers, the subjectivity of summarisation means that multiple summaries might be equally valid for a given set of documents (and for different readers). This variability, combined with the nuances of language and the many ways information can be conveyed, makes the evaluation of summaries complex (C. Lin, 2004). The goal, therefore, is to strike a balance, using metrics and methodologies that capture both the objective and subjective aspects of summarisation quality.

2.9.1 Automated Metrics

In this field, the need for consistent, scalable, and rapid evaluation has driven the development and popularity of automated metrics. These metrics, often derived from quantifiable linguistic comparisons, provide researchers with objective tools to gauge model performance against benchmarks.

Two of the most prominently utilised metrics (each with their own family of sub metrics) in the domain of summarisation are **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) and **BLEU** (Bilingual Evaluation Understudy).

ROUGE: Introduced by Lin (2004), ROUGE primarily focuses on recall, evaluating the extent to which elements (like n-grams, word sequences, and word pairs) in the reference summaries are captured by the generated summary. Variants such as ROUGE-N (n-gram recall), ROUGE-L (longest common subsequence), and ROUGE-S (skip-bigram) offer different granularities of evaluation.

BLEU: Originally designed for machine translation by Papineni *et al.* (2002), BLEU evaluates the precision of the generated text compared to one or more reference texts. It examines the co-occurrence of n-grams in the generated text with those in the reference texts. Despite its origin in translation, its applicability has been extended to summarisation tasks.

This short example demonstrates ROUGE, BLEU, Precision, Recall, and F1 score. It uses a simple original text and a summary, then calculates these metrics.

Original Text:

"The quick brown fox jumps over the lazy dog. The dog was too tired to chase the fox. Meanwhile, a cat watched from a nearby tree."

Summary:

"A fox jumps over a dog. A cat watches."

Calculating the metrics:

ROUGE-1 (unigram overlap):

Precision = (overlapping words) / (total words in summary) = 6 / 8 = 0.75

Recall = (overlapping words) / (total words in original text) = 6 / 19 = 0.316

F1 = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.75 * 0.316) / (0.75 + 0.316) ≈ 0.444

BLEU (using unigrams only for simplicity):

Precision = (overlapping words) / (total words in summary) = 6 / 8 = 0.75

BLEU Score = Precision * Brevity Penalty = 0.75

(Assuming no brevity penalty for this short example)

Precision, Recall, and F1 Score: (Using word overlap as a simple metric)

Precision = (overlapping words) / (total words in summary) = 6 / 8 = 0.75

Recall = (overlapping words) / (total words in original text) = 6 / 19 = 0.316

F1 = 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.75 * 0.316) / (0.75 + 0.316) ≈ 0.444

Interpretation:

1. **ROUGE-1:** The F1 score of 0.444 indicates a moderate overlap between the summary and the original text at the unigram level.
2. **BLEU:** The score of 0.75 suggests a good precision of the summary words, but it doesn't account for recall.
3. **Precision, Recall, and F1:**
    - The high precision (0.75) indicates that most words in the summary are relevant.

- The lower recall (0.316) shows that the summary doesn't capture all the information from the original text.
- The F1 score (0.444) balances precision and recall, giving an overall measure of the summary's quality.

Both ROUGE and BLEU have been widely adopted in the research community due to their computational efficiency and ease of use. Their deterministic nature ensures consistent results, making them valuable for tracking model improvements and comparing different approaches.

2.9.2 Strengths and Limitations of Automated Metrics

2.9.2.1 Strengths

Scalability: Automated metrics can rapidly evaluate large quantities of text, a feat challenging for human evaluators. This scalability is especially vital given the ever-growing datasets in NLP.
Consistency: While human judgement can be influenced by numerous factors, automated metrics ensure consistent evaluations across different models and datasets, making them indispensable for comparative studies (Nenkova and McKeown, 2012).
Ease of Use: Most automated metrics, given their algorithmic nature, are straightforward to implement and integrate into research pipelines.

2.9.2.2 Limitations

Potential Misalignment with Human Judgement: While metrics like ROUGE and BLEU provide quantitative evaluations, they might not always align with human perceptions of quality. For instance, a summary could achieve a high BLEU score by replicating sentences from the source but might lack coherence or novelty when assessed by human evaluators. Callison-Burch, Osborne and Koehn (2006) discuss the limitations of BLEU in evaluating translation quality, showing that optimizing for BLEU does not necessarily result in better translations according to human judgment. This discrepancy underscores the need for evaluation metrics that better capture the nuances of human language understanding (Callison-Burch, Osborne and Koehn, 2006).

Granularity: Automated metrics, particularly when focusing on n-gram overlaps, might not capture the nuances and richness of language. They can sometimes overlook the semantic essence of summaries, prioritising matches based on syntax. This limitation is important to note because a good summary should not only match the source text in terms of word usage

but also convey the underlying meaning effectively. For example, the nuances of medical literature summarisation are often lost in simple n-gram matching metrics, necessitating more sophisticated evaluation methods that account for deeper semantic understanding (L. L. Wang *et al.*, 2023).

Specificity: While automated metrics provide a general measure of summary quality, they might not cater to domain-specific requirements or the unique characteristics of certain summarisation tasks. Louis and Nenkova (2009) emphasised that standard metrics may fail to capture the full spectrum of quality aspects relevant to different domains. They suggested that for fields such as biomedical research, metrics must be tailored to reflect the specific needs and challenges of summarising scientific literature, such as handling contradicting evidence and synthesising comprehensive reviews from multiple sources (Louis and Nenkova, 2009).

2.9.3 Choosing the Right Metric: A Research Perspective

The evaluation of multi-document summarisation (MDS) outcomes is as much a reflection of the underlying objectives of the research as it is a technical necessity. The nature of the summarisation task often determines the choice of the evaluation metric. For tasks aiming to generate abstractive summaries, which may introduce novel phrasings or synthesise information across documents, metrics that purely focus on word overlap, such as ROUGE or BLEU, may fall short. The depth of abstraction demands a comprehensive evaluation metric that can simultaneously assess semantic coherence and novelty (Cohan *et al.*, 2018).

Going further, the subject domain of the summarisation task plays an important role in metric selection. For instance, in critical domains like medicine or law, where the precision of information and factual accuracy are paramount, metrics that accentuate exact matches or domain-specific terminology are particularly relevant. Additionally, when working with datasets that are full of stylistic variations or with diverse summary lengths (yet another reason to focus on structured scientific papers), it becomes very important to use metrics that consider a balance of **recall** and **precision** and potentially factor in length variations.

2.9.4 Human Evaluation: The Gold Standard

Moving beyond automated metrics is the possibility of using human evaluation. Summarisation developed for human understanding benefits greatly from a human-centric evaluation approach. As Grusky, Naaman and Artzi (2018) point out, human readers have the ability to

discern subtleties like readability, coherence, and the overarching 'feel' of a summary, nuances that might be missed by automated metrics.

Several methodologies underpin human evaluations. One is to employ pairwise comparisons, where evaluators compare two or more summaries, ranking them on perceived (subjective) quality. While this approach gives relatively higher quality assessments, it might not give explicit quality scores. Alternatively, evaluators could rank multiple summaries in order of preference, a strategy that helps to understand the relative standing of different summarisation techniques. A more granular mode of evaluation could Likert-scale ratings, where evaluators can rate summaries on a set of predetermined criteria, capturing the multifaceted nature of summary quality (Mohtarami *et al.*, 2018).

However, human evaluations are not without their challenges. Their subjectivity implies that different evaluators may have different opinions on summary quality. Furthermore, the scalability of human evaluations is simply not comparable to that of automated metrics. The slower pace of human assessment makes it a difficult task to evaluate large datasets or summaries from multiple models or techniques. Additionally, consistency remains an elusive goal, with repeated evaluations by the same individual sometimes yielding very different results, influenced by factors ranging from boredom to evolving perceptions or just a different viewpoint on a different day.

The contradiction between automated metrics and human evaluations is a key area of discussion in MDS research evaluation. While automated metrics promise scalability and consistency, the depth and nuance offered by human evaluations remain unparalleled. As chapter 3 will discuss further though, a relatively new methodology called LLM-as-a-judge (Zheng *et al.*, 2023) leverages the latest LLMs to emulate an expert in the field to judge summary quality.

2.9.5 Comparative studies – automated metrics compared to human judgement

Finding the ideal evaluation metric in multi-document summarisation (MDS) often leads researchers down a twin path: the algorithmic rigour of automated metrics and the nuanced discernment of human evaluators. The relationship between these modes has been the subject of numerous studies, each attempting to unravel the interplay of their relationship.

An important study by Callison-Burch, Osborne and Koehn (2006) investigated machine

translation, comparing the BLEU metric with human judgements. The findings, while centred on translation, offered insights transferrable to MDS. The study revealed that while BLEU scores often correlated with human evaluations, discrepancies arose, particularly when the generated text deviated syntactically from the reference yet retained semantic fidelity. Such small differences underscore the limitations of automated metrics that rely heavily on surface-level textual overlaps.

In the specific context of MDS, Liu and Liu (2008) did an extensive study, comparing ROUGE scores with human evaluations across multiple datasets. Their findings were interesting. While there was a general positive correlation between ROUGE scores and human judgements, the strength of this correlation varied across datasets. This variability highlights the challenge of seeking a universally applicable automated metric. The study further underscored the fact that high ROUGE scores did not necessarily guarantee human-perceived quality, especially in cases where summaries, though linguistically coherent, missed pivotal information.

The potential gaps between automated metrics and human evaluations are not just statistical discrepancies but often mirror deeper linguistic and cognitive differences. As described further by Dang (2006), automated metrics, by their very design, can sometimes reward verbosity or penalise succinctness, deviating from human evaluators who might appreciate concise and to-the-point summaries (and different evaluators might themselves have different preferences). Similarly, the granularity of evaluation differs. While human evaluators might judge a summary based on its overall coherence, informativeness, and fluency, automated metrics often split summaries into n-grams, potentially overlooking the summary's full narrative.

This two-way split between automated and human evaluations has major implications for MDS research. It demonstrates the need for a holistic evaluation strategy, one that combines the scalability of automated metrics with the depth of human evaluations. The quest for the 'perfect' evaluation metric, therefore, is not about choosing between humans and algorithms but about harnessing the relative strengths of each.

Such an approach, as proposed by Grusky, Naaman and Artzi (2018), involves iterative evaluations. Initial evaluations using automated metrics can sift through vast datasets or model variants, narrowing down top-performing models. Subsequent in-depth evaluations using human evaluators can then fine-tune model selection, ensuring both algorithmic and human-centric quality. This multi-tiered evaluation strategy, while resource-intensive, promises

robustness and depth, guiding the MDS community towards more impactful and human-aligned research outcomes.

2.10 Chapter Conclusion

This chapter reviewed key concepts in multi-document summarisation (MDS) of scientific literature, tracing its evolution from early extractive methods to modern abstractive approaches. It defined fundamental terminology and compared extractive versus abstractive techniques, justifying this research's focus on abstractive methods for scientific texts.

The review examined MDS challenges including redundancy management and coherence maintenance across sources and discussed applications of pre-trained language models and advanced NLP techniques. It addressed dataset selection considerations, evaluation metrics, and the specific challenges of summarising scientific papers with their specialized terminology and complex concepts.

Key research gaps identified include: limited exploration of pre-trained language models for scientific MDS; lack of domain-specific approaches that handle specialised terminology; integration challenges with newer hybrid techniques like Retrieval Augmented Generation (RAG); and the need for more sophisticated evaluation metrics beyond n-gram matching.

Throughout the review, current gaps and future directions in the field were identified, setting the stage for the research questions addressed in this dissertation. The key gaps identified in the current literature that will inform the research process include:

1. Limited exploration of pre-trained language models for multi-document summarisation: While models like BERT and BART have shown promise in single-document summarisation, their potential in MDS, particularly for scientific papers, remains underexplored.
2. Lack of domain-specific approaches: There is a need for summarisation techniques tailored specifically to scientific literature, which can handle specialised terminology, complex concepts, and maintain factual accuracy.
3. Integration of advanced techniques: The chapter suggests that newer hybrid techniques, such as Retrieval Augmented Generation, have not been well-integrated into existing MDS frameworks for scientific papers.

4. Comprehensive comparative studies: There is a lack of extensive studies that analyse the performance of different MDS frameworks against existing approaches, both extractive and abstractive, especially in the context of scientific literature.
5. Evaluation metrics for scientific summarisation: The review indicates a need for more sophisticated evaluation metrics that can capture the nuances of summarising scientific papers, going beyond simple n-gram matching.
6. Handling of multi-document complexity: There is room for improvement in techniques that effectively manage information redundancy, contradictions, and maintain coherence across multiple scientific documents.
7. Scalability and efficiency: The review suggests that there's a need for more efficient methods to handle large-scale summarisation tasks, particularly for extensive collections of scientific papers.
8. Addressing bias and ethical considerations: The chapter highlights the need for more research into mitigating biases in summarisation models and datasets, especially when dealing with scientific literature from diverse sources.

Chapter 3: Modern LLM tools and techniques and their applications to Multi Document Summarisation

3.1 Introduction

As described in earlier chapters, the growing complexity and volume of information being produced has necessitated the development of advanced techniques for extracting and summarising content across multiple documents. Multi-document summarisation aims to condense information from various sources into a cohesive and concise summary, aiding in knowledge discovery and decision-making processes. This chapter explores the advancements in large language models (LLMs) since 2021, their associated tools and techniques, and their application to multi-document summarisation. In terms of LLMs themselves, this chapter will focus on current state-of-the-art models such as GPT-4, Google Gemini and Gemma, MetaAI's LLaMA, and models from Mistral.

The development of so-called large language models (LLMs) since 2021 has facilitated rapid advancements in the field of natural language processing (NLP) and has enabled more sophisticated and accurate text generation and understanding. Despite their impact, these models exhibit inherent limitations (Bommasani *et al.*, 2024). Consequently, it is important to gain an understanding of where they can be effective, their constraints and the natural language tasks where they can be most effective.

Observers of the field will have noticed that 2021/2, significant moves have been made in the development of LLMs, marked by the release of models like GPT-4 by OpenAI, which builds on the success of its predecessors (GPT2, GPT3.5) with improved capabilities and a larger parameter count (T. B. Brown *et al.*, 2020). Google's Gemma and Gemini (the Open Source and proprietary variants respectively) represents a significant improvement in *multimodal* learning, combining textual and visual data to enhance comprehension and generation tasks. The Gemma models also perform well at low parameter counts. MetaAI's LLaMA models focus on efficient scaling and maintaining high performance, while models from Mistral prioritise computational efficiency without compromising on capability. In isolation, the models may seem similar (and indeed their performance may seem similar, but they each have advantages and disadvantages.

As described earlier, the importance of effective tokenisation in LLMs cannot be overstated as tokens are the fundamental units these models process. Effective tokenisation impacts the model's performance significantly, influencing its ability to understand and generate coherent text. Pre-training and fine-tuning are very important phases in the development lifecycle of an LLM, where **pre-training** involves learning linguistic structures from very large corpora, and **fine-tuning** adapts the model to specific tasks using supervised learning (T. B. Brown *et al.*, 2020).

Beyond traditional training methodologies, modern LLM systems can be augmented with external data sources, enhancing their performance in specific tasks. Retrieval-Augmented Generation (RAG) is one such technique that integrates external knowledge from a 'knowledge base' to improve the accuracy and contextual relevance of the output that the LLM generates. Lewis *et al.* (2021) describe the advantages of this technique in that it can help prevent LLM 'hallucinations' (where an LLM arbitrarily makes up new facts) by allowing the LLM to expand and augment their knowledge. Additionally, the incorporation of knowledge graphs into LLM systems can enrich their reasoning capabilities and contextual understanding, so making them invaluable tools in complex information synthesis tasks such as single and multi-document summarisation.

The subsequent sections of this chapter will explore the specifics of these tools, providing a detailed overview of the latest techniques and their applications to multi-document summarisation.

The chapter will also examine the strengths and weaknesses of various new LLM models, describing how they can be fine-tuned and discusses the integration of external data and knowledge graphs to develop end-to-end solutions.

3.2 Advances in LLM Architectures: GPT-4, Google Gemini, LLaMA, and Mistral

3.2.1 Overview of LLM Architectures

Large Language Models (LLMs) have significantly improved in their architectural design leading to improved capabilities in natural language understanding and generation. The foundational architecture these models is the **Transformer**, introduced by Vaswani *et al.*, (2017), which enabled efficient training of deep neural networks by leveraging mechanisms such as self-attention (self-attention in large language models (LLMs) is a mechanism that allows each word in a sentence to weigh the importance of every other word when generating a representation, enabling the model to capture context and relationships between words more effectively).

The following section explores the architectures of state-of-the-art models such as GPT-4, Google Gemini, LLaMA, and Mistral, comparing them with earlier transformer-based architectures like BERT and T5. Figure 5 (below) shows a timeline of some of the key architectural enhancements leading to the development of the modern LLM.

Figure 5: Timeline of Transformer-based model releases

The timeline shows a number of significant developments in transformer-based models relevant to MDS. One trend is the progression towards handling longer sequences (Longformer, BigBird, and LongT5, for example) which address the challenge of processing extensive document sets. Another key development is the emergence of models specifically designed for summarisation tasks, such as PEGASUS, which uses novel pre-training techniques to improve abstractive summarisation performance. The evolution from BERT to more efficient variants like ALBERT and DistilBERT shows the research effort in reducing computational demand while maintaining performance, making these models more accessible for practical applications in MDS. In addition, the introduction of T5 and its variants marked a shift towards more versatile "text-to-text" frameworks, which had the impact of simplifying the application of these models to various NLP tasks, including document summarisation.

3.2.1.1 Transformer-Based Architectures: BERT and T5

***BERT (Bidirectional Encoder Representations from Transformers)**:*

Architecture: BERT is a language representation model developed by a research team at Google AI Language (Devlin *et al.* (2018) which employs a bidirectional transformer encoder to read text in both directions (left-to-right and right-to-left) to understand the context of words. The diagram below (fig 6) highlights how the bidirectional nature of BERT is implemented through a self-attention mechanism. This allows each token to interact with all other tokens in the sequence and so allows the model to capture context from both directions simultaneously.

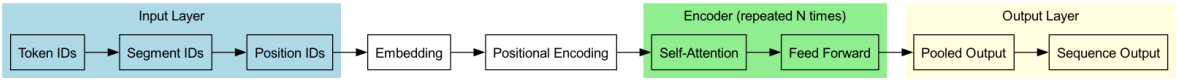

Figure 6: BERT architecture

Objective: It uses two main pre-training tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). Masked language modelling (MLM) is a pre-training task used in NLP where certain words in a sentence are replaced with a mask token. The model is then trained to predict these masked words based on the surrounding context of words, helping it to learn bidirectional (words before and after a particular word) representations of text. Next Sentence Prediction (NSP) is a pre-training task where the model is given a pair of sentences and trained to predict whether the second sentence is the *actual* next sentence in the context of the first. This task helps the model to understand the relationship between sentences and improves its ability to comprehend and generate coherent text. In the case of BERT, doing both pre-training tasks allows the model to learn both the detail of contextual relationships at the word level and the higher-level inter sentence relationships.

Strengths: BERT is strong in tasks requiring deep contextual understanding due to its bidirectional nature. It set new benchmarks in various NLP tasks such as question answering and text classification. BERT is still used as a core component of Google Search.

T5 (Text-To-Text Transfer Transformer):

Architecture: T5 frames every NLP task as a text-to-text problem, converting inputs into a text format that the model then generates a text output for. It uses the standard transformer architecture with both encoder and decoder.

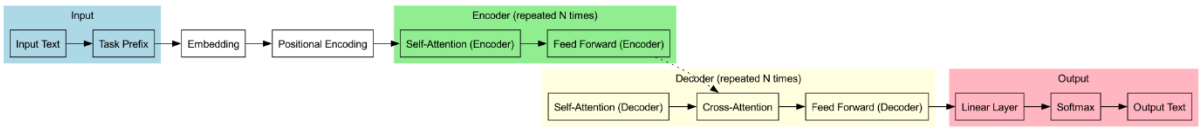

Figure 7: T5 architecture

As illustrated in Figure 7, T5's architecture implements this text-to-text approach through a complete encoder-decoder transformer structure. The diagram shows how task-specific prefixes are incorporated directly into the input layer, enabling the model to process diverse NLP tasks through the same unified framework.

Objective: T5 uses what is termed a 'multi-task learning objective' during pre-training, which includes masked language modelling but in a text-to-text format. This means that each NLP task, whether it is translation, summarisation, question answering, or classification, is framed as feeding text as input and generating text as output. This approach makes the model architecture and training process simpler, and allows T5 to learn from a wide variety of tasks simultaneously, leveraging shared knowledge across tasks to improve the overall performance.

Strengths: T5 has a unified text-to-text approach which permits it to handle a wide range of tasks (including text summarisation) within a single framework, therefore making it versatile and powerful across different NLP benchmarks.

3.2.2 Modern LLM Architectures: GPT-4, Google Gemini, LLaMA, and Mistral

GPT-4 (Generative Pre-trained Transformer 4):

Architecture: GPT-4 follows the architecture of the OpenAI predecessor models but with significant scaling in terms of parameters. It is a unidirectional transformer, generating text by predicting the next token in a sequence.

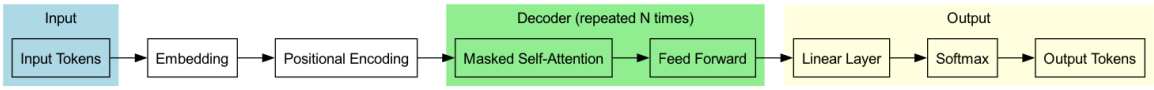

Figure 8: GPT architecture

Figure 8 shows that GPT's decoder-only architecture consists of a streamlined flow from input tokens through to embedding and positional encoding, followed by repeated layers of masked self-attention and feed-forward networks. This shows the unidirectional nature of GPT, where the masked self-attention mechanism ensures each token only processes information from previous tokens in the sequence.

It introduces several key improvements over the original Transformer architecture. Most obviously, GPT simplifies the structure by using only the **decoder** portion of the Transformer, removing the encoder stack entirely. This streamlined approach focuses on the autoregressive nature of language modelling. The "Masked Self-Attention" block in the decoder ensures that each token can only attend to its preceding tokens, maintaining the left-to-right generation capability important for text prediction tasks. Unlike the original Transformer or models such as BERT, there is no bidirectional context or cross-attention mechanism. This unidirectional approach, combined with the deep stack of decoder layers (represented by the "repeated N times" notation in the diagram), allows GPT to generate coherent and contextually relevant text by building upon its own previous outputs. The simplification to a decoder-only model, whilst maintaining the core components of embedding, positional encoding, and the transformer block (self-attention followed by feed-forward layers), results in a more focused and potentially more efficient architecture for generative tasks.

Objective: GPT-4 is trained using a generative pre-training objective called causal language modelling (CLM), focusing on predicting the next word in a sentence. This involves predicting the next word in a sequence based on the preceding words. This *autoregressive* approach (predicting future token sequences based on what has come before it) means the model generates text one word at a time, always considering only the previously generated words and not future ones.

This type of training helps GPT-4 learn to generate coherent and contextually appropriate text, making it effective for tasks like text completion, summarisation, and conversational agents. CLM has a number of benefits over MLM and NSP: it has a simpler training process as its goal is to straightforwardly predict the next word rather than MLM requiring masking of tokens and the prediction of those masked tokens; autoregressive models such as those trained with CLM are more easily scaled to larger datasets and model sizes so improving performance; CLM handles input sequences of variable lengths without requiring and special handling to mark sentence boundaries etc. so making it more flexible across different types of text generation tasks.

Strengths: Its massive scale and extensive training data enable GPT-4 to generate highly coherent and contextually relevant text, making it effective in a wide range of generative tasks.

3.2.3 Development of Large Language Models (LLMs) from 2022 Onwards

The landscape of so-called Large Language Models (LLMs) has evolved rapidly since 2022 with significant advancements in model architectures, training techniques, and applications. This section explores some of the most relevant developments and models that have supported significant developments in this field.

3.2.3.1 Advances in LLM Architectures

GPT-3 and GPT-4

OpenAI's GPT-4 (released in 2023) and before that GPT-3 in 2020, marked a significant development in LLM capabilities. While the exact architecture details remain proprietary an outline is described in the previous section. GPT-4 demonstrated substantial improvements over its predecessors, GPT-3.5 and GPT3, in areas such as reasoning, task complexity, and multimodal inputs (OpenAI *et al.*, 2024).

GPT-4 features improved context understanding and retention, an enhanced ability to follow complex instructions, fewer hallucinations, and increased factual accuracy. With the recent release of the GPT-4o model, it also introduced the capability to process both text and image inputs, expanding its potential applications across various domains. One of the most significant developments, and one that is likely to have driven uptake, is not the model itself but the parallel development of a Chat interface which makes the models much easier to use and to interact with.

Google Gemini

Google's Gemini models, introduced in late 2023, represents a multimodal approach to AI, designed to understand and generate text, image, video, and audio content (Gemma Team *et al.*, 2024). Gemini was trained to be multimodal from the outset, allowing for understanding and synthesis of different forms of data (text, video, images etc.). The model's architecture is scalable across different sizes (Ultra, Pro, and Nano), catering to various computational requirements and use cases. Gemini has demonstrated state-of-the-art performance on various benchmarks (Gemma Team *et al.*, 2024), including language understanding and mathematical reasoning, thus demonstrating suitability for a range of language tasks. The multimodality features of Gemini are worth special mention. Although the types of document (academic papers) this research seeks to summarise are mainly text, papers do of course contain tables, charts and images and it may prove to be a useful feature to be able to enhance context with information drawn from these non-text sources.

Meta's LLaMA

Meta AI's LLaMA (Large Language Model Meta AI) series were first released in 2023 and updated with the LLaMA 2 models later that year. They focus on efficient scaling and open-source accessibility (Touvron *et al.*, 2023). The LLaMA series offers various model sizes, ranging from 7B (7 billion) to 70B parameters, to suit different computational resources. In the context of Large Language Models (LLMs), parameters are the adjustable components within the model that are learned during training. These parameters, typically represented as numerical values, capture patterns and relationships in the training data, enabling the model to understand and generate human-like text. The number of parameters in an LLM is often used as a measure of its complexity and potential capability. Take for example a model with two variants, say a '2B' model and a '7B' model; the 2B and 7B variants have approximately 2 billion and 7 billion parameters, respectively. The number of parameters significantly impacts the model's performance and capabilities. Generally, models with more parameters, like the 7B variant, have the potential to capture more intricate patterns and nuances in language potentially leading to better performance on complex tasks. However, they also require more computational resources for training and inference. The 2B variant, with fewer parameters, may be more efficient in terms of computational requirements and memory usage, making it more suitable for deployment in resource-constrained environments, but potentially at the cost of some performance on more complex tasks compared to its larger counterpart.

Meta's approach to model design again emphasises improved training efficiency, allowing for higher performance with fewer parameters compared to some other models. The open-source release of LLaMA has been particularly useful, as it has driven community-driven development and research and enabling wider experimentation in a range of language fields such as multi-document summarisation.

Mistral AI Models

Mistral AI models (Jiang *et al.*, 2023) have recently gained some attention for their focus on efficient, high-performance, open source models. Their approach emphasises computational efficiency without compromising on capability. Mistral's models incorporate innovative architecture designs, such as sliding window attention and sparse attention mechanisms, which contribute to their efficiency. In a similar way to LLaMA, Mistral's decision to release open-source versions of their models has enabled wider adoption and experimentation, further supporting access to advanced LLM technology.

3.2.4 Key Milestones and Breakthroughs

3.2.4.1 Parameter-Efficient Fine-Tuning (PEFT)

One of the most significant developments in LLM technology has been the introduction of Parameter-Efficient Fine-Tuning techniques (Xu *et al.*, 2023). These methods allow for the adaptation to new knowledge domains and fine-tuning of large pre-trained models to specific tasks by using minimal computational resources. Being able to fine-tune an existing model, rather than training from scratch, is an enormous benefit and again has supported the democratisation of LLM development.

LoRA (Low-Rank Adaptation) introduces trainable rank decomposition matrices to the model weights, allowing for efficient fine-tuning (Hu *et al.*, 2021). By introducing these matrices, LoRA enables the model to adapt to new tasks with reduced computational overhead, enhancing the fine-tuning process's effectiveness and efficiency.

QLoRA (Quantized Low-Rank Adaptation) (Dettmers *et al.*, 2023) extends the concept of Low-Rank Adaptation by incorporating quantisation techniques, which significantly reduce memory requirements. By applying these techniques, QLoRA achieves efficient fine-tuning with even lower computational and memory overhead, making the adaptation process even more resource efficient. Quantisation techniques themselves involve reducing the precision of the numbers used to represent the parameters of a model, typically from floating-point to lower-bit representations such as 8-bit integers. This process decreases the memory footprint and computational requirements of the model, so enabling faster processing and reduced resource consumption while maintaining an acceptable level of performance.

There are many benefits of these PEFT techniques, but most significantly they significantly reduce the computational requirements for fine-tuning, making it possible to adapt large models on consumer-grade hardware. PEFT methods preserve much of the pre-trained model's knowledge while allowing for task-specific adaptations, striking a good balance between generalisation and specialisation.

3.2.4.2 Advancements in Embeddings

Embedding techniques have seen substantial improvements, enhancing the ability of LLMs to understand and represent textual data. Models like SBERT (Sentence-BERT) have improved the creation of semantically meaningful sentence representations (Reimers and Gurevych, 2019).

These advancements allow for more nuanced understanding of text, which is particularly valuable in tasks like document summarisation where capturing semantic relationships is very important. Additionally, progress in multilingual embeddings has enabled better representation of text across multiple languages, expanding the potential applications of LLMs in global and multilingual contexts.

Chunking and Vector Databases

The development of efficient chunking techniques was first described by Reimers and Gurevych (2019) in their development of Sentence-BERT and, together with the use of vector databases, has been fundamental for managing and retrieving information from large text corpora. Advanced algorithms for breaking down large documents into semantically coherent chunks have improved the handling of long-form content, a critical aspect of multi-document summarisation. These chunking methods are complemented by techniques that maintain context across chunks, ensuring that the broader narrative or thematic elements of a document are not lost in the process. This process is important when creating embeddings and vector representations for longer documents. By dividing a document into chunks, each chunk can be processed independently, allowing for efficient handling of large texts and facilitating more accurate and meaningful vector representations.

The process of creating embeddings with chunking begins by dividing the text into smaller segments. These chunks could be sentences, paragraphs, or fixed-length segments. Each chunk is then converted into a vector representation using an embedding model, such as BERT or GPT-3. This results in a set of vectors that represent the chunks. The vectors for each chunk can then be aggregated to form a single vector representation for the entire document. This aggregation can be done through averaging, concatenation, or more sophisticated pooling techniques.

Chunking offers several benefits, particularly in handling long texts. Traditional embedding models often have limitations on the maximum input length they can process. Chunking allows for handling texts that exceed these limits. Additionally, by processing chunks separately, models can focus on the local context within each chunk, potentially leading to better overall understanding when the chunks are combined. This approach also enhances scalability, enabling the processing of large documents in parallel and improving efficiency.

There are six commonly used chunking techniques used in RAG systems shown in figure 9:

1. **Fixed-Size Chunking**: This method splits the text into chunks of a predetermined size, such as 100 words or 500 characters. It is simple to implement but may break semantic units.
2. **Sentence-Based Chunking**: This approach creates chunks based on sentence boundaries, which helps maintain the semantic unity of individual sentences.
3. **Paragraph-Based Chunking**: This method uses paragraph breaks as natural chunk boundaries, which can help preserve more context within each chunk.
4. **Semantic Chunking**: This technique employs Natural Language Processing (NLP) methods to create chunks based on topic or semantic similarity, potentially improving the relevance of retrieved chunks.
5. **Sliding Window Chunking**: This approach creates overlapping chunks by sliding a fixed-size window over the text, which can help maintain context across chunk boundaries.
6. **Hybrid Chunking**: This method combines multiple techniques to balance context preservation and retrieval efficiency, adapting to the specific needs of the RAG system.

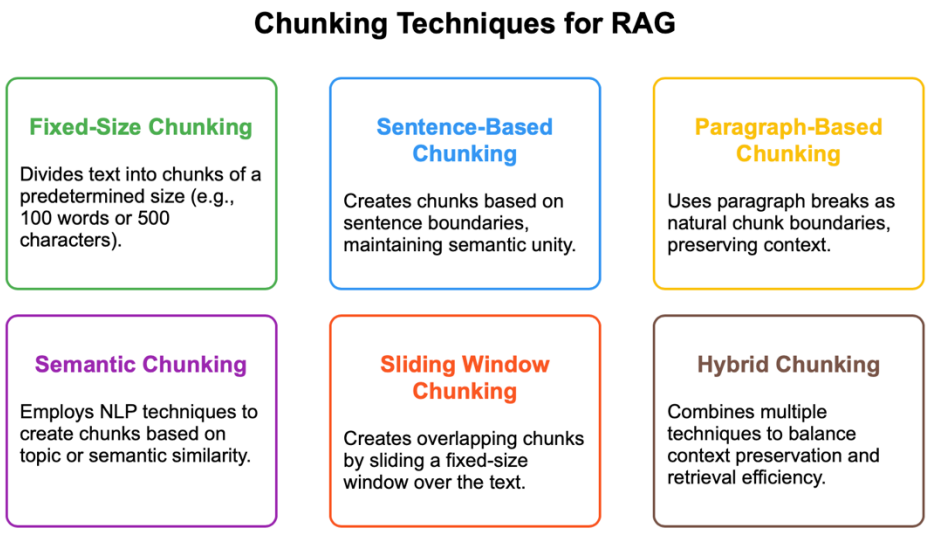

Figure 9: Chunking techniques for RAG

Figure 10 below illustrates four different chunking techniques applied to the same sample text:

1. **Fixed-Size Chunking**: The text is divided into chunks of 50 characters each. This method is simple but can break words and sentences arbitrarily.
2. **Sentence-Based Chunking**: Each sentence becomes a separate chunk. This preserves the integrity of sentences but results in chunks of varying sizes.
3. **Semantic Chunking**: The text is divided based on topic or meaning. In this example, we've grouped related sentences about AI and its applications.

4. **Sliding Window Chunking**: This creates overlapping chunks of 100 characters each, with a 50% overlap. This method helps maintain context between chunks but results in some repetition.

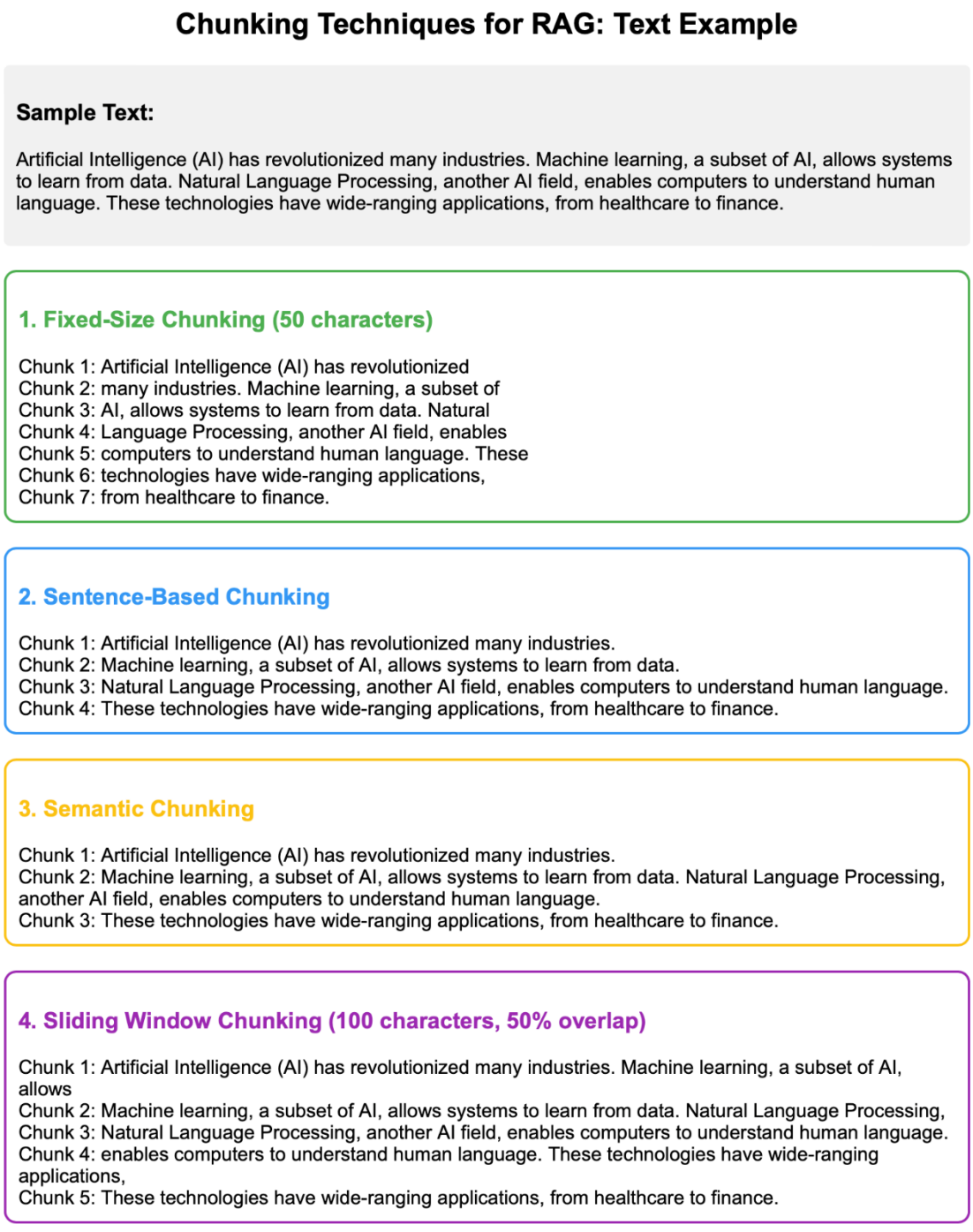

Figure 10: Example of chunking techniques applied to text

As described in the review by Han, Liu and Wang (2023), vector databases, such as Chroma DB, have emerged as efficient solutions for storing and retrieving high-dimensional vector

representations of text. These databases offer scalable solutions for similarity search in large document collections, a fundamental operation in many NLP tasks, including document retrieval for summarisation. The combination of advanced chunking techniques and efficient vector storage and retrieval has significantly enhanced the ability of LLM-based systems to handle large-scale document processing tasks.

Vector databases store data as high-dimensional vectors, which are usually derived from machine learning models like word embeddings (e.g., Word2Vec, GloVe) or more advanced sentence and document embeddings (e.g., BERT, GPT embeddings). These vectors capture semantic meaning in a dense (i.e. non-sparse) numerical format.

Insertion in a vector database involves converting the input (e.g., a document or sentence) into its vector representation and then storing this vector along with any associated metadata. The database indexes these vectors to facilitate fast similarity searches.

Retrieval is primarily based on similarity search. When querying, the input is converted to a vector, and the database returns the most similar vectors based on distance metrics like cosine similarity or Euclidean distance. This process is often optimized using techniques such as Approximate Nearest Neighbor (ANN) search algorithms, which trade a small amount of accuracy for significant speed improvements.

While TF-IDF can be used to create vector representations, modern vector databases typically use more sophisticated embedding techniques that capture deeper semantic relationships. These embeddings are often produced by neural networks trained on large corpora, allowing them to capture context and meaning more effectively than traditional statistical methods like TF-IDF.

Figure 11 illustrates the chunking process workflow, showing how raw text documents are first processed through chunking, then transformed into vector representations using embedding models such as BERT or GPT. These vector representations are aggregated into document vectors, which are then stored in vector databases like Chroma DB for efficient similarity search and retrieval operations.

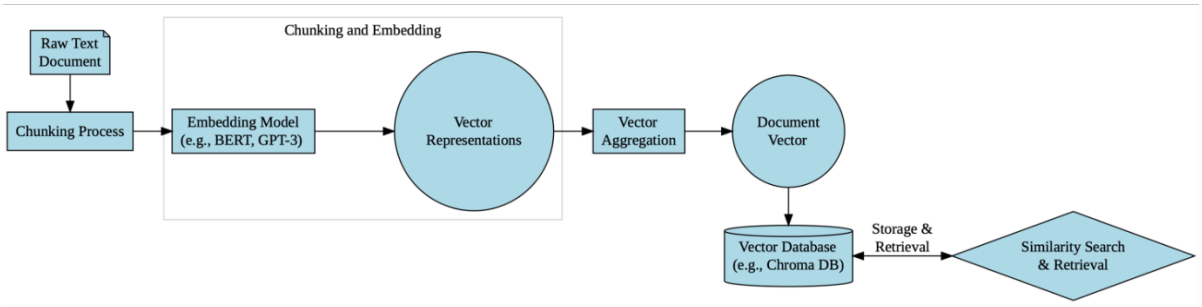

Figure 11: The chunking process

3.2.5 Comparative Analysis of Notable Models

Table 2: Comparison of models

Model	Key Features	Strengths	Limitations
GPT-4	Multimodal input, improved reasoning	Versatile, high accuracy	Closed-source, high computational requirements
Gemini	Native multimodal training, scalable	State-of-the-art performance, efficient	Limited availability of smaller variants
LLaMA 2	Open-source, efficient scaling	Community-driven development, adaptable	Requires fine-tuning for specific tasks
Mistral	Computational efficiency, innovative attention mechanisms	High performance with lower resource requirements	Newer, with less extensive testing in production environments

3.3 Understanding LLM Tokens

The concept of tokens and tokenisation is fundamental to the functioning of Large Language Models (LLMs). This section explores the definition and role of tokens, various tokenisation techniques, and how tokenisation impacts model performance.

3.3.1 Definition and Role of Tokens in LLMs

In the context of LLMs, tokens are the basic units of text that the model processes. A token can be a word, part of a word, or even a single character, depending on the tokenisation method used. The role of tokens is fundamental: they serve as the input and output units for the model, bridging the gap between human-readable text and the numerical representations that neural networks can process.

Tokens act as the building blocks of language understanding for LLMs. When an LLM processes text, it does not work with raw characters or words directly. Instead, it operates on sequences of these tokens. Each token is usually associated with a unique numerical identifier in the model's vocabulary. This tokenisation process enables the model to handle a wide range of linguistic items, from common words to rare terms, punctuation and even sub-word units.

Tokens also influence the model's **context window**—the amount of text an LLM can process at once. Most modern LLMs have a fixed context window, often measured in tokens rather than words or characters. For instance, GPT-3 has a context window of 2048 tokens, while GPT-4 can handle up to 8192 tokens in its standard configuration. Having a token-based approach allows for more precise control over the input size and computational requirements.

3.3.2 Tokenisation Techniques

Tokenisation techniques have evolved significantly, reflecting the need for more efficient and effective ways to represent text for LLMs. Several key approaches have emerged:

Word-based Tokenisation: This straightforward method splits text into words, typically using spaces and punctuation as delimiters. While easy to create tokens in this way, it can lead to large vocabularies and struggle with out-of-vocabulary words.

Example:
Input: "The quick brown fox jumps over the lazy dog."
Tokens: ["The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog", "."]

Character-based Tokenisation: At the other extreme, this approach treats each character as a token. It results in a small vocabulary but requires longer sequences to represent text, and so it potentially can miss higher-level patterns in the inputs.

Example:
Input: "Hello, world!"
Tokens: ["H", "e", "l", "l", "o", ",", " ", "w", "o", "r", "l", "d", "!"]

Subword Tokenisation: This method strikes a balance between word and character-based approaches and is the approach most often taken by modern LLMs. Commonly used algorithms include:

Byte Pair Encoding (BPE): BPE iteratively merges the most frequent pairs of bytes or characters to form new tokens. It efficiently handles common subwords while maintaining the ability to tokenise any string of characters (Gage, 1994).

Example using a simplified BPE vocabulary:
Input: "uncomfortable"
Tokens: ["un", "comfort", "able"]

WordPiece: Similar to BPE, WordPiece (Wu *et al.*, 2016) builds a vocabulary by selecting subword units that maximise the likelihood of the training data. It is particularly effective for languages with compound words. It is a subword tokenisation algorithm developed by Google and widely used in models like BERT. It helps in handling the out-of-vocabulary problem by splitting words into subwords, ensuring that even rare or unseen words can be represented through their component subwords. This method is particularly effective in creating embeddings and vector representations for words, allowing models to efficiently manage large vocabularies and improve performance on various natural language processing (NLP) tasks.

Example using a hypothetical WordPiece vocabulary:
Input: "tokenisation"
Tokens: ["token", "##isation"]

SentencePiece: This algorithm performs subword tokenisation directly from raw sentences, without requiring pre-tokenisation. It can handle various languages uniformly, including those without clear word boundaries.

Example using a hypothetical SentencePiece model:
Input: "LLMs are powerful tools for NLP tasks."
Tokens: ["▁LLM", "s", "▁are", "▁powerful", "▁tool", "s", "▁for", "▁NLP", "▁task", "s", "."]
(Note: "▁" represents a space in SentencePiece)

Hybrid Approaches: Some modern LLMs use combinations of these techniques. For example, GPT-2 and its successors use a variant of BPE that operates on bytes rather than Unicode characters, allowing for a fixed-size vocabulary that can tokenise any Unicode string.

Example using a simplified GPT-2 style tokenisation:

Input: " 🤗 Learning is fun!"

Tokens: [" 🤗", "Learning", "is", "fun", "!"]

(Note: The emoji is treated as a single token in this byte-level approach)

3.3.3 Impact of Tokenisation on Model Performance

Context Window Utilisation: The efficiency of tokenisation affects how much semantic content can fit within the model's context window. More efficient tokenisation allows for more information to be processed in a single forward pass, which is particularly important for tasks involving long documents or multiple inputs.

Example:

Consider a model with a 1024 token context window:

Input 1 (word-based): "The quick brown fox jumps over the lazy dog." (9 tokens)
Input 2 (subword-based): "The quick brown fox jumps over the lazy dog." (7 tokens if "quick" and "brown" are common subwords)

In this simple example,, (shown diagrammatically below in figure 12) the subword tokenisation approach uses fewer tokens to represent the same text, potentially allowing more content to fit within the context window.

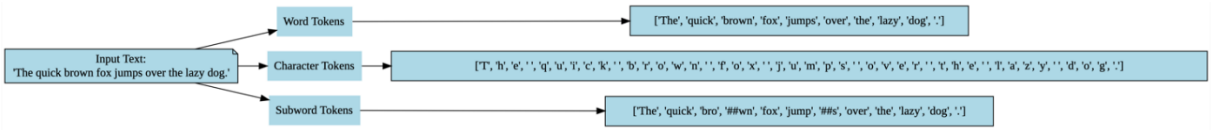

Figure 12: Word-based, character-based, and subword-based embeddings

3.4 Pre-Training and Fine-Tuning of LLMs

The development of Large Language Models (LLMs) typically involves two main stages: pre-training and fine-tuning. This section explores these processes, their objectives and techniques, with a particular focus on their relevance to multi-document summarisation tasks.

3.4.1 Overview of Pre-Training: Objectives and Techniques

Pre-training is the initial phase of LLM development where the model learns general language understanding and generation capabilities from a very large corpus of text data. The primary objectives of pre-training are:

- To develop a wide understanding of language patterns, semantics, and world knowledge.
- To create a versatile foundation that can be adapted to various downstream tasks.
- To learn effective representations of text that capture complex linguistic items.

Common pre-training techniques include:

Self-Supervised Learning: This approach uses the in-built structure of the data to create supervised learning tasks. To recap, the most prevalent methods are:

a) Masked Language Modelling (MLM): Originally used by BERT, this technique randomly masks tokens in the input and trains the model to predict these masked tokens. It supports the development of bidirectional understanding of context.

b) Causal Language Modelling (CLM): Used in models like GPT, this method predicts the next token given the previous tokens, supporting the model's ability to generate coherent text.

Contrastive Learning: This technique trains the model to differentiate between similar and dissimilar pieces of text, helping it learn more sturdy representations. An example is **SimCSE** (Simple Contrastive Sentence Embeddings) (Jiang, Zhang and Wang, 2022), which has shown potential in improving sentence embeddings.

Multi-task Learning: Some models are pre-trained on multiple objectives simultaneously. For instance, T5 (Text-to-Text Transfer Transformer) frames various NLP tasks as text-to-text problems during pre-training.

As Multi-task learning involves training a model on multiple related tasks simultaneously. The T5 (Text-to-Text Transfer Transformer) model (Raffel *et al.*, 2020) is a prime example of this approach.

Example tasks in T5's pre-training:

- **Translation**:

  Input: "translate English to German: The weather is nice today."

  Output: "Das Wetter ist heute schön."

- **Summarisation**:

  Input: "summarise: The article discusses climate change impacts..."

  Output: "Climate change affects global temperatures and weather patterns."
- **Question Answering**:

  Input: "question: What is the capital of France? context: Paris is the capital and most populous city of France."

  Output: "Paris"

The choice of pre-training data is also important. Modern LLMs often use web-crawled data, books, and academic papers ensuring exposure to diverse writing styles, topics, and domains. This diversity is particularly beneficial for tasks like multi-document summarisation where the model may encounter a wide range of document types and subjects.

3.4.2 Fine-Tuning Methods for Specific Tasks

As discussed earlier in this chapter, fine-tuning can be used to adapt a pre-trained model to specific tasks or domains. For multi-document summarisation, fine-tuning is essential to teach the model the specialised task of identifying key information across multiple sources and generating coherent summaries.

Traditional Fine-Tuning: This involves further training of the entire pre-trained model on a task-specific dataset. While effective, it can be computationally expensive and may lead to catastrophic forgetting of pre-trained knowledge.

Parameter-Efficient Fine-Tuning (PEFT): As discussed previously, PEFT methods like LoRA and QLoRA are more efficient alternatives. These techniques are particularly valuable for multi-document summarisation, as they allow for adaptation to specific document types or summarisation styles without the need for extensive computational resources.

Prompt Tuning: This (Qiu *et al.*, 2024) method involves learning continuous prompt embeddings while keeping the pre-trained model frozen. It can be an effective approach for tailoring an LLM to specific summarisation tasks without modifying the base model.

Instruction Tuning: This technique (Ghosh *et al.*, 2024) involves fine-tuning the model on a diverse set of tasks framed as instructions. For multi-document summarisation, this could

involve training on various summarisation instructions (e.g., "Summarise these scientific papers" or "Provide a comparative summary of these news articles").

When fine-tuning for multi-document summarisation, several specific considerations should be considered:

- **Dataset Preparation**: Creating high-quality datasets that include multiple source documents and their corresponding summaries. These datasets often need to span various domains and document types.
- **Handling Long Inputs**: Developing strategies to deal with the combined length of multiple documents, which often exceeds the model's context window. This is where techniques like efficient chunking and Retrieval-Augmented Generation (RAG) become relevant.
- **Cross-Document Understanding**: Encouraging the model to identify and reconcile information across multiple documents, including potential contradictions or varying perspectives.
- **Output Control**: Fine-tuning the model to generate summaries of appropriate length and style, which may vary depending on the specific use case.

3.4.3 Case Studies of Successful LLM Implementations

To illustrate the practical application of these concepts, a few case studies relevant to multi-document summarisation can be examined:

PEGASUS: Google's PEGASUS model (J. Zhang *et al.*, 2020), specifically designed for abstractive summarisation, demonstrates the power of task-specific pre-training. Its pre-training objective, gap-sentence generation, involves masking whole sentences and generating them from the remaining document. This approach is very close to the summarisation task, leading to strong performance even with limited fine-tuning data.

PRIMERA: This (Xiao *et al.*, 2022) model, developed for multi-document summarisation, showcases the benefits of task-specific architecture design and pre-training. PRIMERA uses a hierarchical encoder to efficiently process multiple documents and employs a pre-training strategy that explicitly encourages cross-document understanding.

LongT5: An extension of the T5 model, LongT5 (Guo *et al.*, 2022) addresses the challenge of processing long inputs, a common issue in multi-document summarisation. It incorporates

efficient attention mechanisms that allow it to handle much longer sequences than traditional transformers, making it well-suited for summarising multiple or lengthy documents.

GPT-3 with RAG: While not specifically designed for summarisation, the combination of GPT-3 with Retrieval-Augmented Generation has shown promising results in multi-document tasks. This approach uses a retrieval system to fetch relevant information from a large corpus, which is then fed into GPT-3 along with the query. For example, in work by Lewis *et al.*, (2021), the researchers introduced the concept of Retrieval-Augmented Generation (RAG), where a retrieval system is combined with a generative model to enhance the model's performance on knowledge-intensive tasks. While not exclusively focused on summarisation, the principles discussed are applicable to multi-document summarisation tasks.

These case studies highlight several key trends in LLM development for multi-document summarisation:

- The importance of aligning pre-training objectives with the target task.
- The need for efficient architectures capable of handling multiple long documents.
- The potential of combining LLMs with retrieval systems to enhance performance.
- The benefit of domain-specific fine-tuning, even for large, general-purpose models.

As research in this field develops, it might be expected to see further innovations in pre-training and fine-tuning techniques specifically tailored to the challenges of multi-document summarisation and similar language tasks. The integration of RAG (Retrieval Augmented Generation) techniques in particular, is promising for enhancing the accuracy and relevance of generated summaries by grounding them in information retrieved from the source documents.

3.5 Augmenting LLMs with External Knowledge

3.5.1 Introduction to Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a technique that enhances the capabilities of Large Language Models (LLMs) by incorporating external knowledge during the generation process. This approach aims to combine the language understanding of LLMs with the accuracy and up-to-date information from external sources.

RAG typically involves two main components:

1. A **retriever**: Responsible for finding relevant information from an external knowledge base.
2. A **generator**: Usually an LLM that uses the retrieved information to produce the final output.

As illustrated in Figure 13, the RAG architecture consists of a dual-component system where user input is processed by both a retriever, which queries an external knowledge base, and a generator (typically an LLM) that incorporates the retrieved information to produce a final output. This shows how RAG effectively bridges the gap between the knowledge stored within LLMs and external, up-to-date information sources.

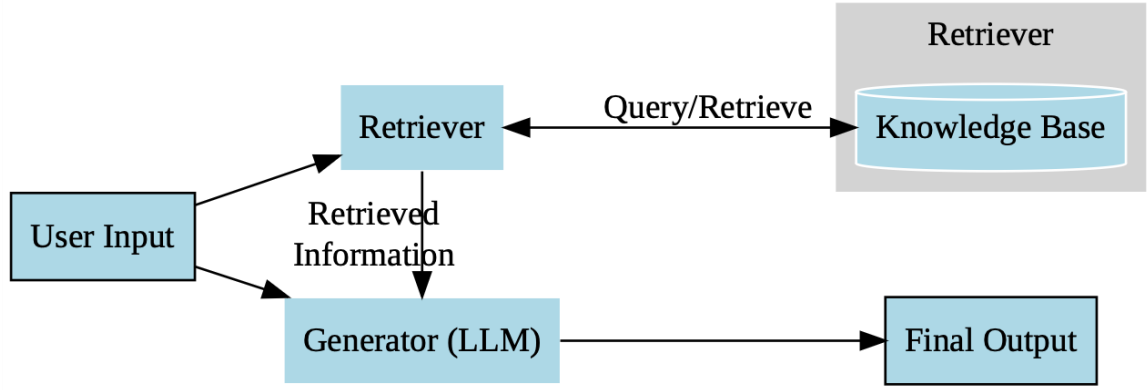

Figure 13: RAG retriever-generator

The RAG paradigm was introduced by Lewis *et al.* (2021) in their 2021 paper, demonstrating significant improvements in tasks requiring access to specific knowledge.

The versatility of RAG techniques extends beyond text-based applications, as explored in a presentation (Callaghan, 2024b) on introducing multimodality into the RAG pipeline to enhance information retrieval for customer support systems (many Enterprise documents contain images, graphs and charts as well as text). This work demonstrates the potential for integrating diverse data types in retrieval-augmented systems, opening avenues for future research in scientific literature summarisation.

3.5.2 Techniques for Integrating External Knowledge

Several techniques can be employed to integrate external knowledge into LLMs:

- **Dense Retrieval**: This method uses dense vector representations of both the query and the documents in the knowledge base. Similarity search is then performed to find the most relevant documents.
- **Sparse Retrieval**: This approach uses traditional information retrieval methods like TF-IDF or BM25 to find relevant documents based on keyword matching.

- **Hybrid Retrieval**: Combines both dense and sparse retrieval methods to leverage the strengths of both approaches.
- **Reranking**: After initial retrieval, a separate model can be used to re-rank the retrieved documents based on their relevance to the query.

3.5.3 Benefits and Challenges

Retrieval-Augmented Generation (RAG) offers several several benefits in this field. First of these is improved accuracy, as RAG can provide more precise and current information, particularly for queries that demand specific or up-to-date knowledge. This approach also helps in reducing hallucinations, a common issue with LLMs, by grounding responses in retrieved information thus decreasing the likelihood of generating false or inconsistent content. Another advantage is the high degree of customisability; the external knowledge base can be easily modified or tailored for specific domains or use cases so allowing for greater flexibility. Moreover, RAG systems enhance transparency by providing the sources of information used in generating responses, which in turn increases trustworthiness and explainability of the model's outputs. It is even possible to create RAG-based system that will cite their sources to documents within the knowledge base.

However, RAG is not without its challenges. Their performance is heavily dependent on the quality and relevance of the retrieved information making retrieval quality a very important factor. Additionally, there is considerable complexity involved in effectively integrating the retrieved information with the LLM's inherent knowledge. Latency is another concern, as the retrieval step introduces additional processing time, which can impact applications requiring real-time responses. Lastly, the management and upkeep of large external knowledge bases can be resource-intensive, posing challenges in terms of storage and maintenance. But despite these hurdles, the benefits of RAG make it a promising approach in enhancing the capabilities of language models and thus multi-document summarisation.

3.5.4 Case Study: Multi-Document Summarisation with RAG

Consider a scenario where an LLM is tasked with summarising multiple research papers on climate change. RAG could be applied using Chroma DB (a commonly used vector database) to solve this problem in this fashion:

1. Document Ingestion:

```
# Assuming 'papers' is a list of research paper texts
for i, paper in enumerate(papers):
    collection.add(
        documents=[paper],
        ids=[f"paper_{i}"],
        metadatas=[{"topic": "climate change"}]
    )
```

Code Listing 1: Document Ingestion

2. Query and Retrieval:

```
query = "Summarize the main findings on sea level rise from recent
climate change research"
retrieved_docs = collection.query(
    query_texts=[query],
    n_results=3
)
```

Code Listing 2: Query and retrieval

The summarisation process begins with document ingestion (Listing 1), where research papers are added to the vector database with appropriate metadata. This is followed by the query and retrieval phase (Listing 2), which demonstrates how specific information about sea level rise (in this case) can be efficiently retrieved from the collection using semantic search capabilities, limiting the results to the most relevant documents for summarisation.

3. LLM Integration:

The retrieved documents are then passed to the LLM along with the original query to generate a comprehensive summary. Figure 14 illustrates the complete RAG-LLM retrieval process. It shows the workflow from paper ingestion to summary generation, where papers are first added to a vector database (Chroma DB) with metadata, then retrieved based on user queries. The retrieved documents, along with the original query, are passed to a Large Language Model which generates the final comprehensive summary.

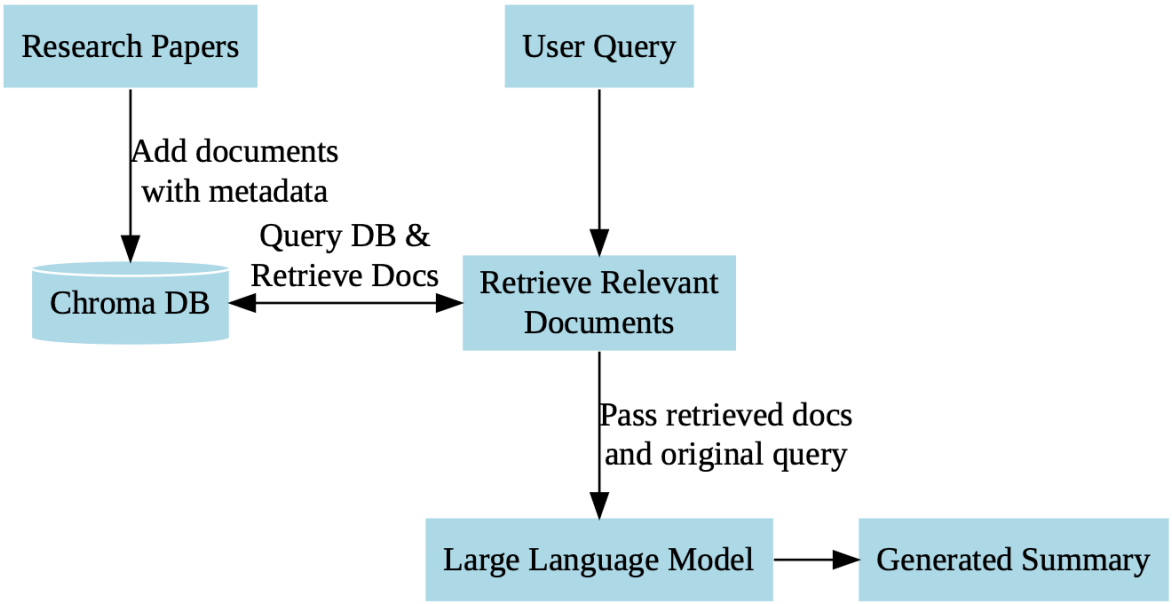

Figure 14: The RAG-LLM retrieval process

3.6 Incorporating Knowledge Graphs in LLMs

3.6.1 Overview of Knowledge Graphs

Knowledge Graphs (KGs) are structured representations of information that capture entities, their attributes, and relationships. They provide a structure for organising and querying complex, interconnected data in a machine-readable format.

Key components of a Knowledge Graph:

- **Nodes**: Represent entities (e.g., people, places, concepts)
- **Edges**: Represent relationships between entities
- **Properties**: Attributes of entities or relationships

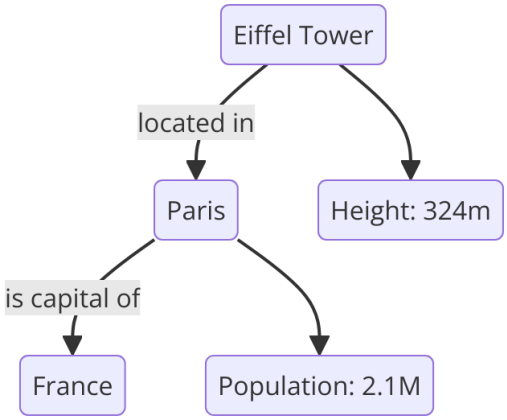

Figure 15: Knowledge graph

Knowledge Graphs have become used in various applications, from enhancing search engines to powering question-answering systems (Punjani and Tsalapati, 2023). The example shown in Figure 15 is a knowledge graph which represents information as interconnected entities with defined relationships and properties. This example demonstrates how factual knowledge about the Eiffel Tower is structured, with nodes representing entities (Eiffel Tower, Paris, France), edges showing relationships ('located in', 'is capital of'), and properties providing specific attributes (height, population).

3.6.2 Techniques for Integrating Knowledge Graphs with LLMs

Integrating Knowledge Graphs with Large Language Models and RAG could significantly enhance their performance especially in tasks requiring factual knowledge and logical reasoning. Some key techniques are:

Knowledge-Enhanced Pre-training:

This approach incorporates KG information during the LLM's pre-training phase. For example, the **ERNIE** model (Enhanced Representation through kNowledge IntEgration) (Wang and Feng, 2022) aligns textual contexts with their corresponding entities in a KG .

Example: Given the sentence "Paris is the capital of France", ERNIE would align "Paris" and "France" with their corresponding entities in a KG, allowing the model to learn both textual and structured knowledge simultaneously.

Knowledge-Guided Attention:

This technique modifies the attention mechanism in transformer-based models to incorporate KG information. The **K-BERT** model uses this approach, injecting relevant knowledge into the input sequence and adapting the attention mask to accommodate the added information (Bai *et al.*, 2022). This paper discusses the EK-BERT model, an enhanced version of K-BERT, which uses a sentiment knowledge graph to improve sentiment analysis tasks. The study shows the incorporation of knowledge-guided attention mechanisms to better integrate external knowledge into the model, demonstrating its effectiveness in handling complex NLP tasks.

Knowledge-Aware Output Generation:

In this method, the LLM's output layer is modified to consider KG information when generating text. The **KALM** (Knowledge-Aware Language Model) (S. Feng *et al.*, 2023) uses this technique, incorporating a knowledge selection module that chooses relevant KG triples to guide text

generation. It integrates knowledge-aware contexts for long document understanding, demonstrating the application of knowledge-aware output generation.

Graph Neural Networks (GNNs) for Knowledge Integration:

As illustrated in Figure 16, the knowledge graph integration approach uses Graph Neural Networks (GNNs) to process knowledge graphs and incorporate their structured information into LLMs. The diagram shows how information flows from a knowledge graph through a GNN into various components of the language model's architecture, including the input embedding layer, attention mechanism, and output layer. The GreaseLM model exemplifies this technique, using a GNN to encode KG information and incorporating it into a pre-trained language model (Zhang *et al.*, 2022).

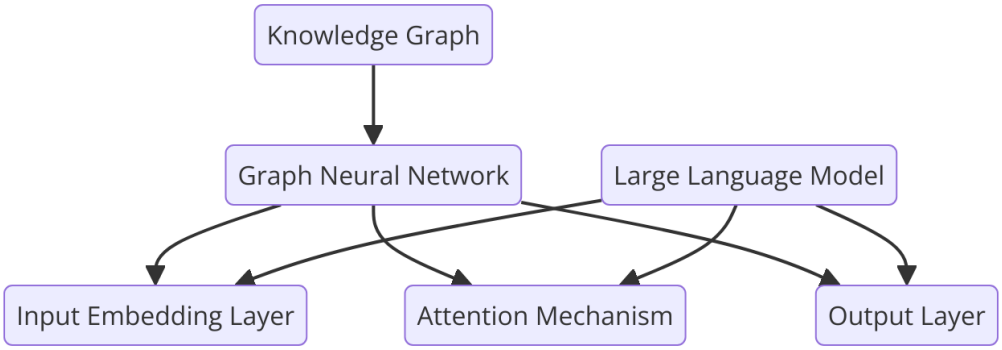

Figure 16: Knowledge graph integration

3.6.3 Applications in Enhancing Summarisation Tasks

Incorporating Knowledge Graphs into LLM/ RAG systems could significantly improve summarisation tasks, particularly for multi-document summarisation. Some potential applications are as follows:

Entity-centric Summarisation:

KGs can help identify and link important entities across multiple documents, ensuring that the summary captures key information about central entities.

For example: consider summarising multiple news articles about a specific company. A KG-enhanced LLM could identify the company, its key personnel, and major events, ensuring these important elements are included in the summary.

Fact Consistency Checking:

KGs can serve as a 'factual backbone', helping the LLM verify information across documents and reduce inconsistencies in the generated summary.

Contextual Enrichment:

KGs can provide additional context that might not be explicitly stated in the source documents, leading to more informative summaries.

Example implementation using Python and a hypothetical KG-enhanced LLM:

```
from kg_enhanced_llm import KGEnhancedLLM
from knowledge_graph import KnowledgeGraph

# Initialize the KG and LLM
kg = KnowledgeGraph.load("company_kg.json")
model = KGEnhancedLLM.from_pretrained("kg_llm_model")

# Input documents
docs = [
    "TechCorp announced a new AI chip.",
    "The CEO of TechCorp spoke about renewable energy.",
    "TechCorp's stock price rose by 5% yesterday."
]

# Generate KG-enhanced summary
summary = model.summarize(docs, knowledge_graph=kg)
print(summary)
```

Code Listing 3: Summary Generation Pseudocode

The pseudocode above (Code Listing 3) suggests how a KG-enhanced LLM might be used to generate a summary that incorporates both the information from the input documents and relevant knowledge from the KG. The same process can be represented diagrammatically. Figure 17 shows a proposed KG integration model for document summarisation. It shows the workflow described in the pseudocode, indicating how knowledge graphs interact with the LLM during the summarisation process. The flow begins with loading both the knowledge graph and the KG-enhanced LLM as separate components which are then used in the document processing stage. Input documents feed into this processing step, finally generating a KG-enhanced summary as output.

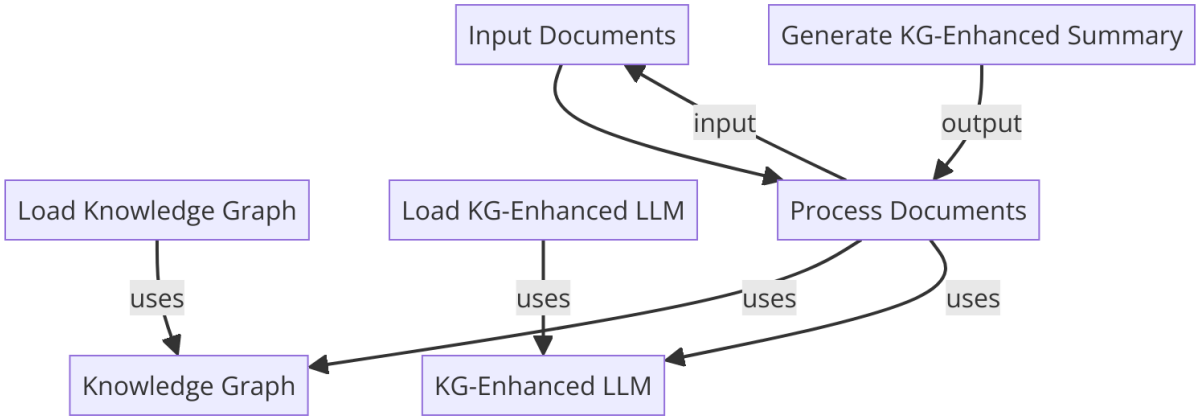

Figure 17: A proposed KG integration model

3.7. Modern LLM Techniques in Multi-Document Summarisation

3.7.1 Comparative Analysis of Traditional vs. Modern Approaches

As discussed in Chapter 2, traditional approaches to multi-document summarisation often relied on extractive methods, selecting and combining the most relevant sentences from source documents. These methods typically used statistical techniques, such as term frequency-inverse document frequency (TF-IDF), or graph-based algorithms like TextRank (Mihalcea and Tarau, 2004). As illustrated in Figure 18, these traditional approaches follow a multi-step pipeline where documents undergo sentence scoring before selection and then producing extractive summaries. In contrast, modern LLM-based approaches use the power of large-scale language models to generate more coherent and abstractive summaries through a more direct processing path. These models can understand context, paraphrase content and even infer (although that inference needs to be treated with some care to ensure it is correct) information not explicitly stated in the source documents.

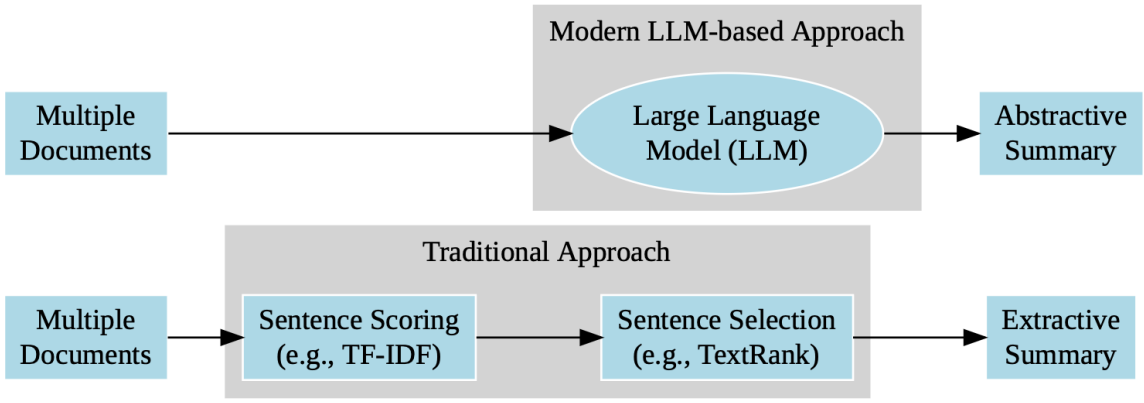

Figure 18: A comparison of approaches to summarisation

Key differences:

Abstractive vs. Extractive: While traditional methods mainly extracted and rearranged existing sentences, LLMs can generate new sentences that capture the essence of the input documents.

Coherence: LLM-generated summaries often exhibit better narrative flow and coherence compared to the sometimes disjointed output of extractive methods.

Background Knowledge: LLMs can leverage their pre-trained knowledge to provide context and fill gaps in the source documents (especially if used in combination with techniques such as RAG, described earlier).

Language Understanding: Modern LLMs demonstrate superior understanding of nuances, context, and implicit information in the source texts.

Scalability: LLMs can handle larger volumes of input text more effectively than many traditional methods.

3.7.2 Comparison of models, features and examples

Table 3: Summary features and applications of Transformer-based models

Model/Dataset	Key Features	Example Application	Implementation Steps
LONGFORMER	- Linear attention scaling - Pre-trained on long documents	Summarising multiple scientific papers on climate change	1. Concatenate full texts with special tokens 2. Apply LONGFORMER tokeniser 3. Feed tokenised input to model 4. Generate comprehensive summary

Model/Dataset	Key Features	Example Application	Implementation Steps
LED (Longformer Encoder-Decoder)	- Long input processing - Decoder for coherent long-form text - Global attention mechanism	Summarising multiple news articles on complex geopolitical events	(Not specified in the given information)
PRIMER	- Multi-task learning approach - Auxiliary tasks (e.g., next sentence prediction) - Hierarchical attention mechanism	Summarising multiple customer reviews for a product	1. Preprocess reviews, maintain document boundaries 2. Apply hierarchical encoding 3. Generate integrated summary 4. Post-process for coherence and readability
HIBERT (Hierarchical BERT)	- Two-level hierarchy: sentence and document - Captures document structure and inter-document relationships	Summarising multiple legal documents for a complex case	1. Segment documents into sentences 2. Encode sentences 3. Encode sentence representations at document level 4. Generate summary based on hierarchical representation
Multi-News (Dataset)	- 56,000+ articles in 45,000 subjects - Human-written summaries - Diverse news sources	Fine-tuning models (e.g., BART, T5) for multi-document summarisation	1. Prepare data (cluster articles, pair with summaries) 2. Initialise pre-trained model. 3. Train on Multi-News data 4. Evaluate using metrics and human evaluation

Table 3 gives a comparison of a number of Transformer-based models suitable for multi-document summarisation, showing their key architectural features, practical applications, and implementation workflows. This shows how different architectural models have addressed the challenges of processing and synthesising information from multiple documents.

3.7.3 Performance Metrics and Evaluation

Evaluating the quality of multi-document summaries can be a complex but important task. Both automatic metrics and human evaluation can play important roles in assessing summarisation performance. In this section, some of the most appropriate evaluation mechanisms for multi-document summarisation are described:

3.7.3.1 Automatic Metrics:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

- ROUGE-N: Measures overlap of n-grams between the generated summary and reference summaries.
- ROUGE-L: Considers the longest common subsequence.
- ROUGE-W: Weighted longest common subsequence.

Consider a reference summary: "The cat sat on the mat."
And a generated summary: "A cat was sitting on a mat."

ROUGE-1 (unigram overlap):
Reference: {the, cat, sat, on, the, mat}
Generated: {a, cat, was, sitting, on, a, mat}
Overlap: {cat, on, mat}
ROUGE-1 Precision = 3/7 = 0.429
ROUGE-1 Recall = 3/6 = 0.5
ROUGE-1 F1-score = 2 * (0.429 * 0.5) / (0.429 + 0.5) = 0.462

ROUGE-2 (bigram overlap):
Reference: {the cat, cat sat, sat on, on the, the mat}
Generated: {a cat, cat was, was sitting, sitting on, on a, a mat}
Overlap: {on a}
ROUGE-2 Precision = 1/6 = 0.167
ROUGE-2 Recall = 1/5 = 0.2
ROUGE-2 F1-score = 2 * (0.167 * 0.2) / (0.167 + 0.2) = 0.182

BLEU (Bilingual Evaluation Understudy): Originally designed for machine translation, BLEU can also be used for summarisation evaluation.

Using the same summaries as above:

Reference: "The cat sat on the mat."
Generated: "A cat was sitting on a mat."

1-gram precision: 5/7 (5 matching words out of 7 in the generated summary)
2-gram precision: 2/6 (2 matching bigrams: "cat was", "on a")
3-gram precision: 1/5 (1 matching trigram: "on a mat")
4-gram precision: 0/4 (no matching 4-grams)

BLEU score calculation involves multiplying these precisions and applying a **brevity penalty**, resulting in a final score between 0 and 1.

BERTScore: Uses contextual embeddings from BERT to compute similarity scores between generated and reference summaries (T. Zhang *et al.*, 2020).

BERTScore calculates the **cosine similarity** between BERT embeddings of words in the reference and generated summaries. It produces precision, recall, and F1 scores, typically ranging from 0 to 1, with 1 indicating perfect similarity.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considers synonyms and paraphrases, providing a more flexible evaluation.

METEOR aligns words between the reference and generated summaries, considering exact matches, stemmed matches, synonym matches, and paraphrase matches. It then calculates precision, recall, and a final score that penalises chunk fragmentation.

3.7.3.2 Human Evaluation:

Human evaluation is very important for assessing aspects that automatic metrics might miss. Some of the key dimensions include:

- **Coherence**: How well the summary flows and maintains a logical structure.
- **Consistency**: Whether the summary contains any contradictions or factual errors.
- **Relevance**: How well the summary captures the main points of the source documents.
- **Informativeness**: The amount of important information conveyed in the summary.
- **Readability**: The clarity and ease of understanding the summary.

Evaluation Process:

- **Prepare** a diverse set of multi-document inputs and their corresponding human-written reference summaries.
- **Generate** summaries using the LLM-based system being evaluated.
- **Calculate** automatic metrics (ROUGE, BLEU, BERTScore, etc.) comparing generated summaries to references.
- **Conduct** human evaluation using a Likert scale (e.g., 1-5) for each dimension.
- **Analyse** results, considering both automatic metrics and human judgments.

Example of Human Evaluation:

Consider the following generated summary:

"The new healthcare bill, proposed by the ruling party, aims to provide universal coverage. Critics argue it may increase taxes, while supporters claim it will reduce long-term costs. The bill is set for a vote next month."

Human evaluators might rate this summary as follows:

Coherence: 4/5 (The summary flows well and presents information logically)

Consistency: 5/5 (No contradictions or factual errors are apparent)

Relevance: 4/5 (Captures main points but might miss some details)

Informativeness: 3/5 (Provides key information but could include more specifics)

Readability: 5/5 (Clear and easy to understand)

Challenges in Evaluation:

Evaluation metrics have some potential challenges:

- **Reference Bias**: Automatic metrics rely heavily on reference summaries, which may not capture all valid summary variations
- **Length Sensitivity**: Many metrics are sensitive to summary length, potentially penalising concise but accurate summaries.
- **Lack of Semantic Understanding**: Most automatic metrics struggle to capture deeper semantic similarities.
- **Inter-Annotator Agreement**: Human evaluations can suffer from subjectivity and disagreement between annotators.

To address these challenges, researchers often use a combination of multiple automatic metrics and human evaluation, sometimes employing techniques like inter-annotator agreement scores (e.g., Cohen's Kappa) to ensure reliability in human judgments (Sanchez-Velazquez and Sierra, 2016).

3.8. Challenges and Future Directions

As Large Language Models (LLMs) continue to advance, their application to multi-document summarisation presents both opportunities and significant challenges. This section explores the current limitations of LLMs in summarisation tasks, potential improvements and innovations on the horizon, and the critical ethical considerations that must guide responsible AI development in this field.

While LLMs have greatly advanced the MDS field, significant challenges remain. Addressing these limitations through ongoing innovation, while simultaneously prioritising ethical considerations will be very important in order to develop the full potential of LLMs in this area.

3.8.1 Current Limitations of LLMs in Summarisation

Despite their capabilities, LLMs face several limitations when applied to multi-document summarisation:

1. Context Length Constraints: Most LLMs have a maximum input length, typically ranging from 2048 to 8192 tokens (Song *et al.*, 2024). This limitation can be problematic when summarising multiple long documents, as important information may be lost if it falls outside the context window.
2. Factual Accuracy and Hallucinations: LLMs can sometimes generate plausible-sounding but incorrect information, a phenomenon known as "hallucination" (Ji *et al.*, 2023). In summarisation tasks, this can lead to the inclusion of inaccurate or non-existent information in the summary.
3. Lack of Domain-Specific Knowledge: While LLMs have broad general knowledge, they may struggle with highly specialized or technical content without additional fine-tuning or external knowledge integration (Gururangan *et al.*, 2020).
4. Coherence Across Multiple Documents: Maintaining coherence and capturing key themes across multiple, potentially diverse documents remains challenging for current LLM-based summarisation systems (Liu and Lapata, 2019b).
5. Bias and Fairness: LLMs can perpetuate or amplify biases present in their training data, potentially leading to unfair or skewed summaries (Bender *et al.*, 2021).

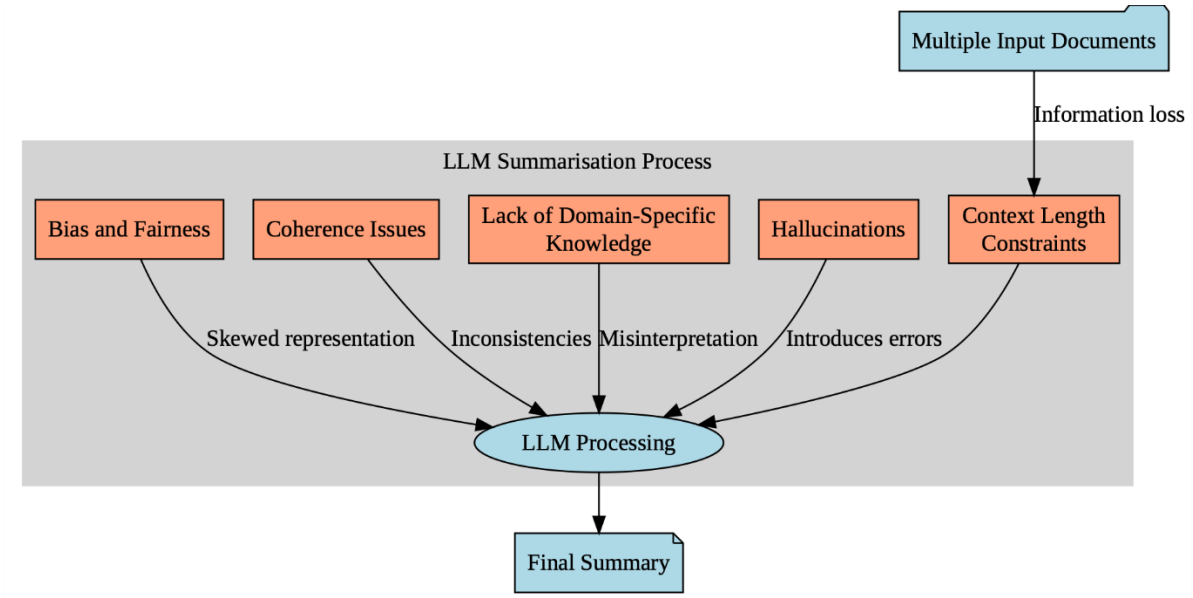

Figure 19: Impact of limitations on the summarisation process

Figure 19 shows how these limitations impact the multi-document summarisation process. The diagram describes the pathways through which each constraint affects LLM processing, resulting in various forms of information degradation. Context length constraints cause direct information loss from input documents while issues such as hallucinations introduce errors into the final summary. Coherence issues lead to inconsistencies, domain knowledge gaps cause misinterpretations, and inherent biases produce skewed representations. Understanding these limitation pathways is important for developing more robust multi-document summarisation systems that can mitigate these challenges.

3.8.2 Potential Improvements and Innovations

Multi-document summarisation using Large Language Models (LLMs) has many opportunities for innovation, with several promising avenues for improvement. One significant development is the creation of long-context models, such as Anthropic's Claude, which can process up to 100,000 tokens (Anthropic, no date). This expanded context window could enhance the ability to summarise multiple lengthy documents simultaneously, providing a more comprehensive and coherent output.

Another interesting approach is Retrieval-Augmented Generation (RAG), which integrates external knowledge bases with LLMs to enhance summarisation tasks (Lewis *et al.*, 2021). As described earlier, RAG allows models to access and incorporate relevant information beyond their training data, significantly improving their performance and adaptability.

The RAG process typically involves several key steps: document chunking, where input texts are broken into manageable pieces; embedding generation, creating vector representations of these chunks; retrieval, using similarity search to find the most relevant chunks for a given query or context; and generation, incorporating the retrieved information into the LLM's output process.

This approach offers numerous benefits for MDS. It improves factual accuracy by grounding the model's output in retrieved facts, so reducing hallucinations. RAG also enables rapid domain adaptation, allowing LLMs to quickly adjust to new subject areas by accessing relevant external knowledge without extensive fine-tuning. Furthermore, the retrieval step enhances transparency by providing a clear link between the generated summary and the source documents, improving interpretability.

However, implementing RAG for summarisation is not without challenges. Developing efficient retrieval mechanisms that can swiftly and accurately identify the most relevant information from large document sets is very important. Seamlessly integrating retrieved information into the generation process while maintaining coherence and relevance presents another set of difficulties. Additionally, managing and maintaining up-to-date, high-quality external knowledge bases requires significant effort and resources.

Despite these challenges, the potential of RAG and long-context models to revolutionise MDS is a key area for exploration. As these technologies continue to evolve, it may be expected to observe significant improvements in the accuracy, comprehensiveness, and adaptability of AI-generated summaries.

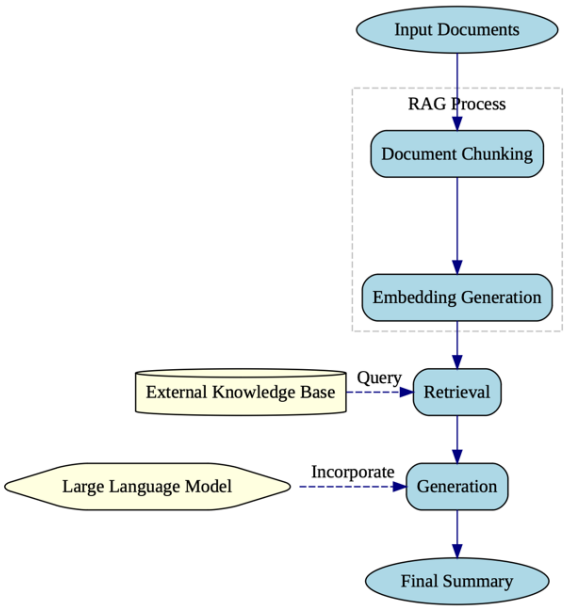

Figure 20: Flowchart of the Retrieval Augmented Generation (RAG) process

Figure 20 shows this Retrieval Augmented Generation (RAG) process. As shown in the flowchart, RAG integrates external knowledge retrieval with LLM generation capabilities through a structured pipeline. The process begins with document chunking and embedding generation, followed by a retrieval phase that queries an external knowledge base. This retrieved information is then incorporated into the generation process, resulting in a final summary that benefits from both the model's inherent knowledge and language generation capacity and external factual knowledge.

Building on the potential of RAG and long-context models, another promising way to improve multi-document summarisation (MDS) is the implementation of **multi-stage summarisation pipelines**. This approach addresses the challenges of maintaining coherence and capturing overarching themes when dealing with multiple documents.

A multi-stage summarisation pipeline typically consists of several key steps, each designed to enhance the overall quality and coherence of the final summary. The process begins with individual document summarisation, where each document is independently condensed to capture its key points and main ideas. This is followed by cross-document theme extraction, which analyses the individual summaries to identify common themes, contradictions, and unique points across all documents. The third stage involves synthesis and coherence building, where a cohesive summary is generated that integrates the identified themes and important points, ensuring logical flow and connection between ideas. Finally, a refinement and fact-

checking stage can be implemented, cross-referencing the generated summary with the original documents to ensure accuracy and completeness.

Liu and Lapata (2019) made significant developments in this area. They proposed a hierarchical transformer architecture for multi-document summarisation. Their work demonstrated the effectiveness of a multi-stage approach, showing how it could handle the complexities of summarising multiple documents whilst maintaining coherence and capturing key themes. Their research underscored the potential of structured, hierarchical approaches in tackling the unique challenges posed by MDS tasks.

The multi-stage pipeline approach complements the benefits of RAG and long-context models by providing a structured framework for handling complex, multi-document inputs. As illustrated in Figure 21, this pipeline progresses through distinct stages: beginning with individual document summarisation, moving to cross-document theme extraction, followed by synthesis and coherence building, and concluding with refinement and fact-checking to produce a final coherent summary. It deals with the challenges of information overload and thematic diversity that often arise in MDS tasks, offering a systematic method for distilling and synthesising large volumes of information into a coherent, comprehensive summary. As research in this area continues to evolve, it's likely that further refinements and innovations in multi-stage summarisation techniques will occur.

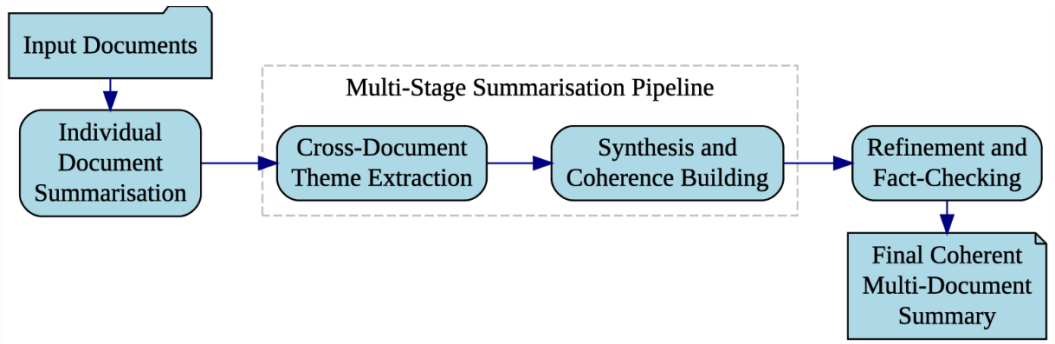

Figure 21: Document summarisation pipeline

Further avenues for improvement in multi-document summarisation include advancements in fine-tuning techniques and the development of explainable AI systems. Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA (Low-Rank Adaptation) and QLoRA (Quantized Low-Rank Adaptation), offer promising approaches for adapting Large Language Models to specific summarisation tasks or domains more efficiently (Hu *et al.*, 2021). These techniques allow for

the fine-tuning of models with significantly fewer parameters, reducing computational costs and potentially improving performance on specialised summarisation tasks.

In parallel, the development of explainable AI techniques for summarisation is gaining importance. These approaches aim to make the summarisation process more transparent and interpretable, helping users understand and trust the generated summaries (Danilevsky *et al.*, 2020). Explainable AI in summarisation could involve highlighting the source sentences that contributed most to the summary, providing confidence scores for different parts of the summary, or offering alternative phrasings or viewpoints.

3.8.3 Ethical Considerations and Responsible AI

As Large Language Models (LLMs) become increasingly prevalent in summarisation tasks, addressing ethical concerns and promoting responsible AI development is paramount. Figure 22 presents an ethical framework for LLM-based summarisation, illustrating the interconnected relationships between key ethical dimensions including fairness, transparency, accuracy, privacy, environmental responsibility, human oversight, accessibility, and ethical use. This framework serves as a roadmap for the following discussion exploring ethical guidelines and considerations for the application of LLMs in multi-document summarisation.

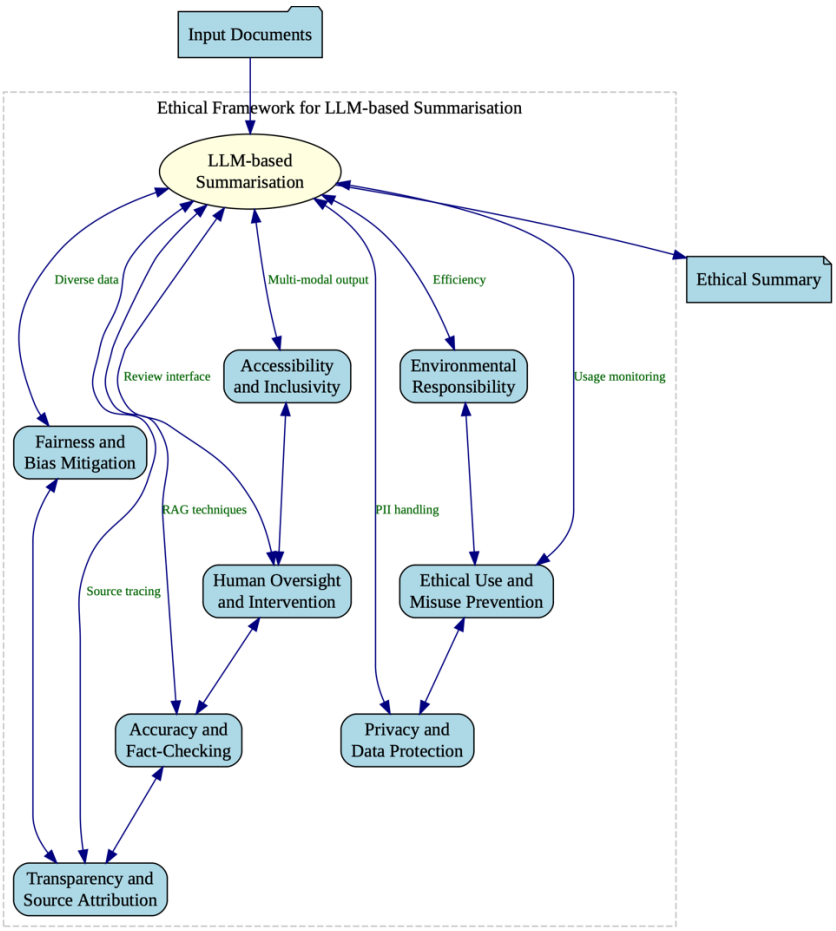

Figure 22: Ethical framework

Fairness and bias mitigation are essential aspects of ethical AI development. Bender *et al.* (2021) highlight the potential dangers of large language models, using the term "stochastic parrots" to describe how these models can reproduce and amplify biases present in their training data. Their work reinforces the importance of regularly auditing summaries for bias using diverse training data and implementing bias detection and correction mechanisms in the summarisation pipeline. This ensures that the generated summaries do not perpetuate or amplify existing biases present in the source documents or the model itself.

Transparency and source attribution are equally important. Doshi-Velez and Kim (2017) argue for the necessity of interpretable machine learning, particularly in high-stakes decisions. Their position paper gives a framework for evaluating the interpretability of machine learning models which is directly applicable to summarisation systems. Implementing systems that can trace information in summaries back to specific parts of source documents and including

metadata about the summarisation process with the output can significantly enhance transparency, as suggested by their research.

Accuracy and fact-checking are critical in preventing the spread of misinformation. As discussed previously, Lewis *et al.*(2021) introduced the concept of Retrieval-Augmented Generation (RAG) combining pre-trained language models with a retrieval mechanism to access external knowledge. This approach can be adapted for fact-checking in summarisation tasks, allowing models to verify information against reliable sources and significantly improve the reliability of generated summaries.

Privacy and data protection concerns must be addressed, especially when dealing with sensitive information in source documents. Lyu, Xu and Wang (2020) discuss collaborative fairness in federated learning and emphasise the importance of privacy-preserving techniques in machine learning. While their work focuses on distributed learning, the principles they discuss could be extended to developing summarisation models that can recognise and appropriately handle personally identifiable information (PII) and other sensitive data. This approach could help to ensure that summarisation systems can process documents containing sensitive information without compromising individual privacy or violating data protection regulations.

Environmental responsibility is an often-overlooked aspect of AI development. Strubell, Ganesh and McCallum (2019) provide an analysis of the environmental impact of training large language models. Their work highlights the need to optimise models for efficiency, use green computing resources, and carefully consider the trade-offs between model size and performance in summarisation tasks. oversight and intervention should be integral to the design of summarisation systems. Rong *et al.* (2023) have surveyed various methods for human-centred explainable AI and have emphasised the importance of human-AI collaboration. Their findings support the creation of interfaces that allow human users to review, edit, and approve machine-generated summaries, especially for critical applications.

Accessibility and inclusivity are also essential considerations in the development of summarisation tools. Trewin *et al.* (2019) discuss AI fairness specifically for people with disabilities, highlighting the need for inclusive design in AI systems. Their work supports the importance of designing user interfaces that comply with accessibility standards and

considering multi-modal summary outputs to improve the inclusivity of summarisation technologies.

Finally, preventing misuse of summarisation technology is very important. Floridi and Cowls (2019) suggested a *'unified framework of five principles for AI in society'*: beneficence, non-maleficence, autonomy, justice, and explicability. These principles can guide the development of ethical guidelines and safeguards for summarisation tools, including the implementation of user agreements, usage monitoring, and potential restrictions on capabilities for sensitive or high-risk applications.

3.9. Chapter Conclusion

3.9.1 Key Points

The recent developments of Large Language Models (LLMs) has significantly advanced the field of natural language processing, with particular implications for multi-document summarisation. Key points from this chapter include:

1. **Architectural Developments**: Models like GPT-4, Google Gemini, LLaMA, and Mistral have extended the capabilities of LLMs, each developing certain capacities:
    i. GPT-4: Enhanced reasoning and multimodal input processing
    ii. Gemini: Native multimodal training and scalability
    iii. LLaMA: Efficient scaling and open-source accessibility
    iv. Mistral: Computational efficiency with innovative attention mechanisms
2. **Tokenisation and Embeddings**: Improved tokenisation techniques, particularly sub-word methods like BPE and WordPiece, have developed LLMs' ability to process and understand text. Advancements in embeddings, including sentence-level and multilingual representations, have further refined semantic understanding.
3. **Parameter-Efficient Fine-Tuning (PEFT):** Techniques like LoRA and QLoRA have made it possible to adapt large pre-trained models to specific tasks with minimal computational resources, supporting easier LLM development and application.
4. **Retrieval-Augmented Generation (RAG):** This technique has significantly improved LLMs' ability to incorporate external knowledge, enhancing accuracy and reducing hallucinations in tasks like summarisation.

5. **Knowledge Graph Integration**: Incorporating **structured** knowledge into LLMs has improved their reasoning capabilities and contextual understanding, particularly beneficial for complex information synthesis tasks.
6. **Chunking and Vector Databases**: These developments have enhanced the handling of large document collections, essential for MDS tasks.

These developments are summarised in table 4 below.

Table 4: Summary of LLM developments and potential impact on MDS

Development	**Impact on Multi-Document Summarisation**
Architectural Advancements	Improved coherence and context understanding in summaries
PEFT Techniques	Easier adaptation of LLMs to specific summarisation tasks
RAG	Enhanced factual accuracy and relevance in summaries
Knowledge Graph Integration	Improved handling of complex, interconnected information
Chunking and Vector Databases	Better processing of large document sets

The next chapter presents the experimental design, implementation details, and evaluation framework which are all directly informed by the theoretical foundations and technical advancements reviewed here. By applying these developments to the specific challenge of scientific paper summarisation, the methodology will address the limitations identified in current approaches while using the strengths of modern LLM architectures and vector retrieval techniques.

3.9.2 The Future of MDS with LLMs

Looking ahead, several important developments could shape the future of multi-document summarisation using LLMs. While these potential developments are exciting, they will most likely come with challenges related to computational resources, ethical considerations, and the need for robust evaluation metrics.

One of the most promising areas is the development of hyper-contextual models. Recent research in this direction includes work by Beltagy, Peters and Cohan (2020) who introduced the **Longformer**, an efficient transformer model that can handle sequences of length up to

32,768 tokens. This suggests that future LLMs might dynamically adjust their context window based on the complexity and length of input documents, allowing for more nuanced summarisation of lengthy or intricate text collections.

Multi-modal summarisation is another area for future development. The potential in this area is highlighted by the work of Feng *et al.* (2023), who proposed a multi-modal pretraining approach for document understanding and generation. Their UniDoc model can process textual, visual, and layout information from documents. This research suggests that future systems could integrate text, images, and even audio/video content in generating comprehensive summaries, building on the multimodal capabilities of models like Gemini.

Real-time adaptive summarisation is an area with significant potential, especially for summarising dynamic content like ongoing news events or research developments. While current research in this specific area is limited, the concept builds on work in stream-based summarisation, such as that by Kedzie, McKeown and Daume III (2019). Their approach to update summaries as new information arrives could be extended and enhanced with more advanced LLM capabilities.

Personalised summarisation is another promising development and one with applications in education. (Zhao, Wang and Rios, 2024) have made developments in this area with their work on preference-grounded summarisation for radiology reports. Their approach incorporates user preferences into the summarisation process, suggesting that future models might be able to tailor summaries to individual needs through advanced fine-tuning techniques.

The development of explainable AI in summarisation is essential for enhancing trust and interpretability. DeYoung *et al.* (2020) have made some progress in this direction with their work on extractive rationales for natural language inference. While this work isn't specifically on summarisation, the principles could potentially be applied to develop LLMs that not only produce summaries but also provide clear reasoning for their content selection and phrasing choices.

Cross-lingual summarisation is also an area of active research that could lead to advanced models capable of seamlessly summarising documents in multiple languages. The work of Wang *et al.* (2023) on zero-shot cross-lingual summarisation shows the potential in this area,

suggesting future developments could produce coherent summaries in any target language from multi-lingual source documents.

While quantum LLMs are still largely theoretical (and maybe still in the realms of Science Fiction), research in quantum natural language processing, such as that by Meichanetzidis *et al.* (2023), suggests potential applications to language models. As quantum computing advances, it could potentially be applied to LLMs, increasing their processing power and enabling more sophisticated summarisation techniques.

Lastly, bias detection and mitigation in summarisation systems is a critical area for future development. The work of Pryzant *et al.* (2020) on automatically neutralising subjective bias in text, while not specific to summarisation, provides a foundation for developing summarisation systems that incorporate advanced bias detection algorithms, ensuring that summaries present balanced viewpoints from diverse document sources.

As LLM technology continues to evolve, MDS stands to benefit significantly from these advancements, potentially revolutionising the processing and synthesis of large volumes of information across various fields and applications. However, it is important to note that these developments will likely come with their own set of challenges, particularly in terms of computational resources, ethical considerations, and the need for robust evaluation metrics.

Chapter 4: Methodology

4.1 Introduction

4.1.1 Research Objectives and Questions

As described in earlier chapters, the enormous growth of scientific literature has created an urgent need for efficient methods to distil and synthesise information from multiple related documents. This research aims to address this challenge by developing and evaluating novel approaches to multi-document summarisation (MDS) of scientific papers, with a particular focus on hybrid techniques that leverage the strengths of both extractive and abstractive methods. As described in the introductory chapter, the study is guided by the following research questions:

RQ1: What are the key features and characteristics of an efficient hybrid multi-document summarisation framework for scientific papers, and how can Retrieval-Augmented Generation (RAG) techniques be effectively incorporated to identify and use sections of interest?

RQ2: How can state-of-the-art language models be adapted and fine-tuned for the task of multi-document summarisation of scientific papers, and what advantages do newer LLMs (such as Gemma 2B/7B) offer over earlier models (like BERT and BART)?

RQ3: How does the performance of the proposed hybrid framework compare to existing approaches, both extractive and abstractive, when evaluated using standard metrics (e.g., ROUGE, BLEU) and on diverse scientific datasets?

To address these questions, the latter phases of the study pursued the following objectives:

1. Develop and implement hybrid multi-document summarisation frameworks based on advanced natural language processing techniques, with a focus on integrating RAG methods with state-of-the-art (as of 2023-24) LLMs such as Gemma 2B and 7B (Gemma Team *et al.*, 2024).
2. Investigate the potential benefits of RAG techniques in identifying and summarising relevant sections of scientific papers effectively, exploring various retrieval and generation strategies.
3. Evaluate the performance of the proposed framework using established evaluation metrics, such as ROUGE and BLEU, using a framework similar to that suggested by Yang *et al.* (2018) for machine reading comprehension tasks (an NLP task similar to

summarisation) as well as novel LLM-based evaluation methods, and compare it to existing state-of-the-art methods on a variety of scientific datasets.

4.1.2 Theoretical Framework

The underpinnings of this research are framed in the fields of natural language processing, information retrieval, with an emphasis on:

1. **Transformer-based architectures**: The research leverages the power of transformer models, which have revolutionised natural language processing tasks through their ability to capture long-range dependencies in text. Amatriain *et al.* (2023) catalogue and discuss many applications of Transformers with slightly modified architectures to attune them to various NLP tasks and it is this paper which gives the rationale for continuing with the Transformer architecture.
2. **Retrieval-Augmented Generation (RAG):** This study explores the theoretical intersection of information retrieval and text generation, aiming to harness the strengths of both paradigms to improve summarisation quality and relevance. In 2020, the Retrieval Augmented Generation technique (Lewis *et al.*, 2021) was introduced as a core technique to augment fine-tuned models with up-to-date and trusted knowledge.
3. **Transfer learning and few-shot learning**: The research also uses the concept of transfer learning, where pre-trained language models are fine-tuned for the specific task of multi-document summarisation. Additionally, it investigates few-shot learning capabilities of modern LLMs to potentially reduce the need for extensive fine-tuning. Gupta, Thadani and O'Hare (2020) discuss this concept of few-shot learning as a technique to steer the output of an LLM where little training data exists; in effect, the LLM is given a few examples to direct text generation towards the required domain.
4. **Evaluation theory**: The study incorporates established theories of summarisation evaluation, including both automated metrics and human judgement, as well as innovative LLM-based evaluation approaches, to ensure a comprehensive assessment of the proposed frameworks. One of the more interesting discussions in this area is by Fabbri *et al.* (2021), who evaluated a set of fourteen automated metrics and evaluated them against human and automatically created summaries. They found that ROUGE and related metrics are still seen as valid methods to evaluate summarisation tools although their research found that more work was needed in this field to devise a definitive set of evaluation metrics.

4.1.3 Overview of Methodological Approach and Evolution of Models

This research adopts a mixed-methods approach, combining quantitative and qualitative methodologies to provide a comprehensive understanding of the multi-document summarisation task for scientific papers. The methodological approach encompasses the following key components:

1. **Data collection and preprocessing:** Careful selection and preparation of diverse scientific datasets to ensure robustness and generalisability of the findings.
2. **Model development and fine-tuning**: Adaptation and fine-tuning of state-of-the-art language models for the specific task of multi-document summarisation of scientific papers.
3. **RAG implementation**: Development of retrieval mechanisms to identify relevant sections of scientific papers, and integration with generative models to produce coherent summaries.
4. **Experimental design**: Implementation of a series of experiments to evaluate different aspects of the proposed frameworks, including ablation studies and cross-domain applicability tests.
5. **Multi-faceted evaluation**: Using automated metrics, human evaluation, and recently developed “LLM-as-a-judge” approaches to evaluate the quality and effectiveness of the generated summaries.
6. **Comparative analysis**: Rigorous comparison of the proposed frameworks against existing state-of-the-art methods to establish the contribution to the field.
7. **Error analysis and iterative refinement**: Systematic analysis of errors and limitations, informing iterative improvements to the proposed frameworks.

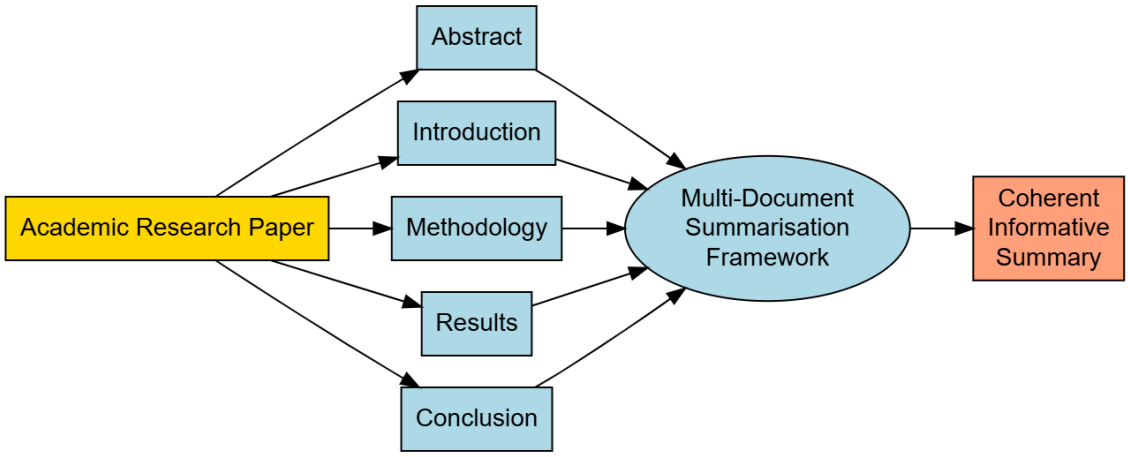

Figure 23: MDS framework

This diagram (figure 23) illustrates how the Multi-Document Summarisation (MDS) framework processes academic research papers. The diagram shows a research paper's structured components (Abstract, Introduction, Methodology, Results, and Conclusion) feeding into the MDS framework, which then generates a coherent and informative summary.

4.1.3.1 Evolution of Models

This research project began with the exploration of earlier transformer-based models such as BERT and BART (and prior to that the LSTM – Long Short-Term Memory model), which gave valuable insights into the capabilities and limitations of these architectures for multi-document summarisation. These models demonstrated good performance in capturing contextual information and generating coherent text. However, several limitations became apparent:

1. **Limited context window:** BERT and BART models typically have a maximum input length of 512 tokens (and LSTMs even less), which is insufficient for processing multiple scientific documents simultaneously or even multiple extracts thereof.
2. **Lack of domain-specific knowledge**: While these models were pre-trained on large corpora, they often lacked specialised scientific knowledge necessary for accurate summarisation of research papers.
3. **Computational efficiency**: Fine-tuning these models for multi-document summarisation tasks proved computationally expensive, particularly for resource-constrained environments.

The progression to more modern LLMs, specifically Google's Gemma 2B and 7B parameter models, was motivated by several factors:

1. **Increased context window**: These models can handle much longer input sequences, allowing for more comprehensive processing of multiple documents. The Gemma models have a context window of 8192 tokens.
2. **Enhanced few-shot learning capabilities**: Modern LLMs demonstrate effective performance in few-shot learning scenarios, potentially reducing the need for extensive fine-tuning on domain-specific data.
3. **Improved efficiency**: Despite their larger size, these models often exhibit better inference efficiency (i.e. they give better and more understandable summaries), making them more suitable for real-world applications.
4. **Advanced reasoning capabilities**: The newer models show improved abilities in logical reasoning and coherence, which are required to produce high-quality scientific summaries. In comparison, older models are relatively 'dumb' and this can be seen in the quality of the summaries.

Although this gives a step-change in improvement to the quality of summaries, context window is still a challenge (8192 tokens is still substantially less than the length of a typical paper) so by incorporating RAG techniques with these advanced LLMs, this research aims to leverage the strengths of both **extractive** (through retrieval) and **abstractive** (through generation) approaches. This hybrid methodology allows for more precise identification of relevant information across multiple documents while maintaining the flexibility to generate novel, coherent summaries. This is of course a very high-level overview and there are multiple lower level techniques that are then applied to this approach to iteratively fine-tune and improve the quality of summaries. The relationship and flow between these stages of the research, from data collection through to creation of the final model, is shown in the flowchart (figure 24) below:

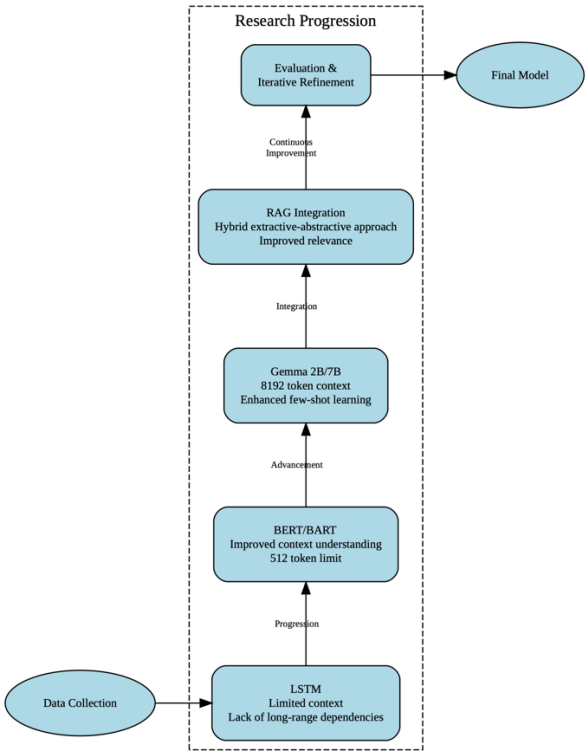

Figure 24: Progression of research

4.2 Mixed Methods Approach

4.2.1 Rationale for Mixed Methods

The complex nature of multi-document summarisation (MDS) for scientific papers needs a correspondingly nuanced and multifaceted research approach. This study therefore adopts a mixed methods design, integrating both quantitative and qualitative methodologies to provide a comprehensive understanding of the issues explored. The rationale for this approach is based on the following considerations:

1. **Complementarity**: Quantitative methods give precision and generalisability, whilst qualitative methods give depth and context. By combining these approaches, the weaknesses of each method and their respective strengths can be capitalised upon. This approach was validated by Johnson and Onwuegbuzie (2004) who described mixed methods approaches as a natural extension to qualitative and quantitative approaches.
2. **Triangulation**: The use of multiple methods allows for the corroboration of findings, improving the validity and reliability of the research outcomes. As Denzin (2017) argues, there is no single approach to validate research findings so several methods used together will be the approach taken here.

3. **Managing the complexity of MDS**: The multifaceted nature of summarising scientific documents requires both objective performance metrics and subjective quality assessments, which are best captured through a mixed methods approach. With this in mind, Mani (2001) discusses the need for both quantitative and qualitative evaluation methods in assessing summarisation systems, again confirming this approach.
4. **Holistic evaluation:** Integrating quantitative and qualitative data provides a more comprehensive evaluation of the proposed RAG-based hybrid summarisation framework, addressing both its texchnical performance and its practical utility in scientific context. In several works, Creswell and Plano Clark (eg. Creswell *et al.*, 2006) discuss how using both quantitative and qualitative data can give a more comprehensive understanding of complex and inter-related phenomena, again providing some validation for this approach.

In conclusion then, this mixture of mixed-methods approaches follows recent trends in NLP research, where researchers have increasingly recognised the limitations of purely quantitative evaluations. For instance, Belz *et al.* (2021) argue for a more nuanced evaluation paradigm in text generation tasks, emphasising the importance of human judgement alongside automated metrics and this is an approach that underpins the selection of the mixed methods approach. Additionally, Reiter (2018) provides a critical analysis of BLEU (Bilingual Evaluation Understudy), one of the most widely used automated metrics in NLP for tasks like machine translation and summarisation. Reiter argues that while BLEU is useful, it has significant limitations and should not be used as the sole evaluation metric. A key quote from the abstract of this paper states:

> *"I conclude that BLEU is not a valid measure of performance in machine translation and natural language generation, and should not be used as the primary evaluation technique in these fields."*

4.2.2 Quantitative Components

The quantitative components of this study focus on measurable aspects of summarisation performance and model efficiency. These include:

1. **Automated Evaluation Metrics**:
    - ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) to assess content overlap between generated summaries and reference summaries (C.-Y. Lin, 2004).
    - BLEU scores to evaluate the fluency and adequacy of generated summaries (Papineni *et al.*, 2002).

- BERTScore to capture semantic similarity beyond exact word matches (T. Zhang *et al.*, 2020).

2. **Model Performance Metrics**:
    - Perplexity measures to assess the language model's predictive performance.
    - Inference time and computational resource usage to evaluate efficiency.
3. **Retrieval Performance** (for RAG components):
    - Precision, Recall, and F1 scores to assess the accuracy of retrieved passages.
    - Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) to evaluate the ranking quality of retrieved passages (Manning, Raghavan and Schütze, 2008).
4. **Statistical Analysis**:
    - Significance testing (e.g., t-tests, ANOVA) to compare the performance of different models and configurations.
    - Effect-size calculations to quantify the magnitude of performance differences.

These quantitative measures provide a robust foundation for comparing the experimental approaches with existing state-of-the-art methods. However, as there limitations of the automated metrics in capturing some of the nuanced aspects of summary quality, these are complemented with qualitative assessments, described below.

4.2.3 Qualitative assessment components

The associated qualitative components of the study aim to capture the nuanced aspects of summary quality and usefulness that are not shown in quantitative metrics. These qualitative approaches can therefore provide richer and more contextual data that complement the quantitative metrics, so giving additional insight into the practical usefulness and perceived quality of the summaries. Qualitive assessments include:

1. **Human Evaluation**:
    - Using domain experts (e.g., researchers or other experienced individuals in relevant scientific fields) to assess the quality, coherence, and factual accuracy of generated summaries.

- Use of rubrics and Likert scales to evaluate aspects such as information coverage, conciseness, and scientific accuracy (van der Zee *et al.*, 2017) (Although the paper focuses on educational videos rather than text summaries, its methodology for human evaluation is applicable and widely respected in the field of content assessment).

2. **Qualitative Content Analysis**:
    - Systematic analysis of generated summaries to identify patterns, themes, and potential areas for improvement (such as through Topic Modelling) in the summarisation process, In Blei, Ng and Jordan (2003), Latent Dirichlet Allocation (LDA) is first introduced, one of the most popular methods for topic modelling. While it doesn't specifically focus on summary analysis, it does provide the foundational technique that is widely used for identifying patterns and themes in text data, including summaries.
3. **LLM-as-Judge Evaluation**:
    - Using large language models (importantly, not the same models that would be used in the summarisation process itself) as judges to provide detailed qualitative feedback on summary quality, with prompts designed to elicit specific critiques and suggestions for improvement (Zheng *et al.*, 2023).

4.2.4 Integration of Quantitative and Qualitative Approaches

The integration of quantitative and qualitative approaches is very important for developing a comprehensive understanding of the summarisation framework's performance. This integration occurs at multiple levels. By triangulating findings from multiple sources and methods, the aim is to provide a robust and comprehensive assessment of the proposed RAG-based hybrid summarisation approach. The levels at which these are applied are:

1. **Design Level**:
    - The study employs a concurrent triangulation design, where quantitative and qualitative data are collected in parallel and used to inform development which is integrated during the analysis and interpretation phases. This is based on the design methodology described by (Creswell and Plano-Clark, 2017)
2. **Data Collection Level**:
    - Quantitative metrics are collected for each generated summary, while qualitative assessments are conducted on a subset of summaries, ensuring a balanced representation across different document types and model configurations.
3. **Analysis Level**:
    - Quantitative results are used to identify high-performing and low-performing summaries, which are then subjected to in-depth qualitative analysis to understand the factors contributing to their performance.
    - Qualitative findings inform the refinement of quantitative metrics and the development of new, more nuanced evaluation measures.
4. **Interpretation Level**:
    - The study employs a 'following a thread' technique described by O'Cathain, Murphy and Nicholl, (2010), where initial quantitative results are used to identify key themes, which are then explored in depth through qualitative analysis.
    - Contradictions between quantitative and qualitative findings are then investigated to uncover potential limitations in the evaluation methods or areas for model improvement.

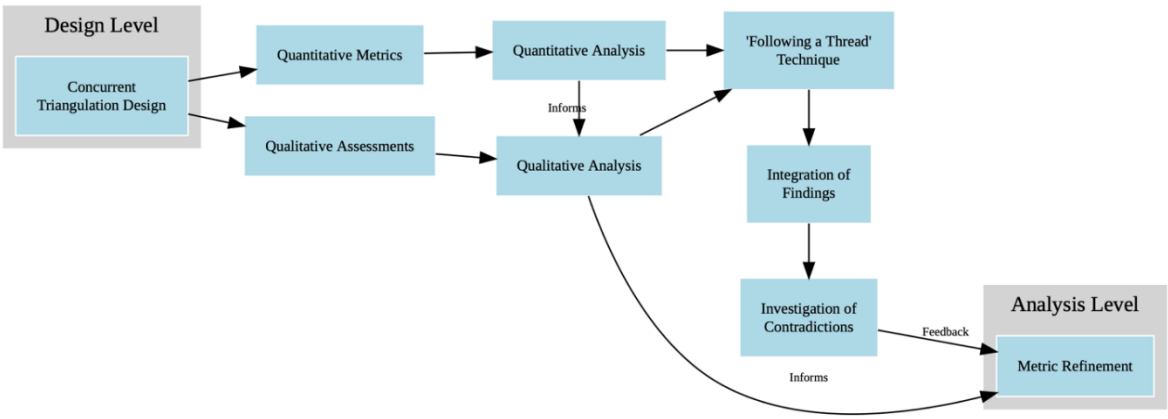

Figure 25: Inter-relationship of mixed methods approaches

The mixed methods design approach used in this study builds upon best practices in NLP research, such as those advocated by Läubli *et al.* (2020), who emphasise the importance of human evaluation alongside automated metrics in machine translation and other text generation tasks. By adapting these principles to the specific context of multi-document summarisation for scientific papers, this study aims to apply this methodology for comprehensive evaluation in this domain. Figure 25 above shows the inter-relationship of the mixed methods approaches employed in this study, indicating how concurrent triangulation design operates at the design level, with both quantitative and qualitative streams informing each other. The diagram shows how these parallel analyses converge through the 'Following a Thread' technique, leading to integration of findings and investigation of contradictions ultimately informing metric refinement at the analysis level.

4.3 Data Collection and Preprocessing

4.3.1 Dataset Selection and Justification

As described in Chapter 2 (Literature Review) and Chapter 3 (Modern LLM techniques), selection of appropriate datasets is very important for developing and evaluating robust summarisation models for scientific papers. This study uses two distinct datasets: **SciTLDR** for training embedding models and fine-tuning large language models (LLMs), and **SciSummNet** for testing and summarisation tasks.

1. **SciTLDR Dataset:**
    - **Description**: SciTLDR is a large-scale dataset containing over 3.9 million scientific papers with their corresponding abstracts and so-called TL;DR ("Too Long, Didn't Read") summaries (Cachola *et al.*, 2020).
    - **Justification**: The vast size and diversity of SciTLDR make it ideal for training embedding models and fine-tuning LLMs, ensuring broad coverage of scientific domains and writing styles.
2. **SciSummNet Dataset:**
    - **Description**: SciSummNet (Yasunaga *et al.*, 2019) is a manually-curated dataset of 1,000 scientific papers in the computer science domain, each accompanied by expert-written summaries.
    - **Justification**: The high-quality, expert-crafted summaries in SciSummNet provide an appropriate benchmark for evaluating the performance of the summarisation frameworks.

The use of two independent datasets for training and testing is important for several reasons, as illustrated in Figure 26. This diagram demonstrates the dataset selection strategy, showing how the SciTLDR dataset (containing 3.9M+ papers) and the SciSummNet dataset (with 1,000 papers) are used in different training and testing pathways to achieve multiple benefits. As shown in the figure, this dual-dataset approach serves several key purposes:

1. **Preventing Data Leakage**: By using separate datasets, the intention is to limit the opportunity that the models are inadvertently exposed to test data during training, which could lead to overfitting and inflated performance metrics. Kaufman *et al.* (2012) describe a similar approach.
2. **Improving Generalisability**: Training on SciTLDR and testing on SciSummNet potentially allows the evaluation of how well the models generalise to unseen data and different summarisation styles. Beltagy, Lo and Cohan (2019), in their SciBERT

paper, discuss this approach to generalisability in the development of language modelling techniques.

3. **Mitigating Dataset Bias**: Each dataset may have inherent biases in terms of writing style, domain coverage, or summarisation approach. Using two datasets helps to identify and mitigate these biases (Bianchi and Hovy, 2021).
4. **Improving Robustness**: The diverse nature of scientific literature needs models that can perform well across various domains and document types. Training on a broad dataset (SciTLDR) and testing on a focused, high-quality dataset (SciSummNet) helps ensure robustness – a methodology supported by Wang *et al.* (2020) in developing a general language understanding benchmark.

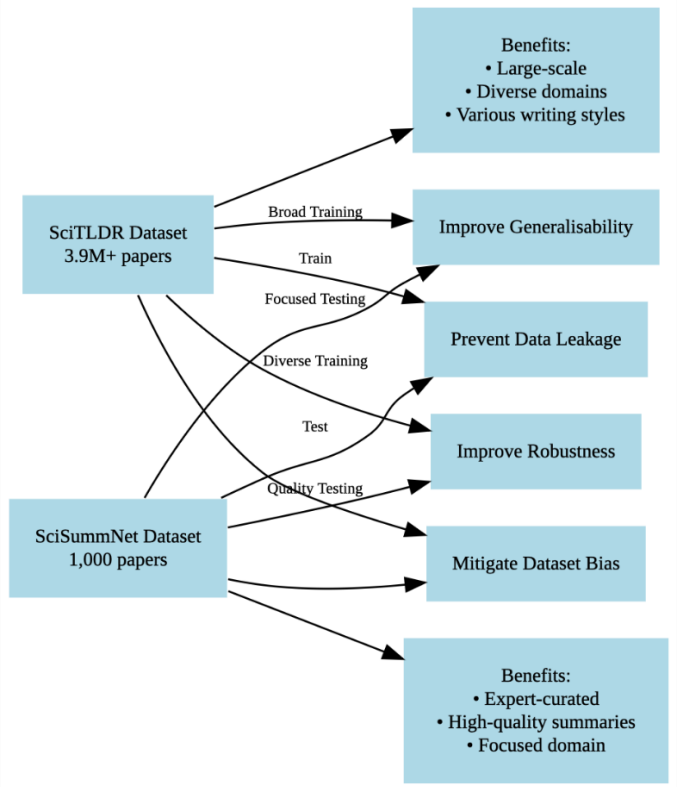

Figure 26: Dataset selection strategy

4.3.2 Data Preprocessing Techniques

Effective preprocessing is essential to ensure the quality and consistency of the input data for the summarisation models. The preprocessing pipeline therefore includes the following steps, as illustrated in Figure 27, which provides a visual representation of the complete workflow from raw data to preprocessed output ready for model input.

1. Text Extraction and Cleaning:
    - For SciTLDR: Extract full text, abstracts, and TL;DR summaries from the provided JSON files.
    - For SciSummNet: Extract full text and expert-written summaries from the XML files.
    - Remove non-ASCII characters, LaTeX commands, and other artefacts common in scientific papers.
2. Tokenisation and Sentence Segmentation:
    - Explore the application of the SciSpacy (Neumann *et al.*, 2019) library, which is specifically designed for processing scientific text
    - Implement careful handling of scientific notation, chemical formulae, and mathematical expressions to preserve their integrity.
3. Normalisation:
    - Convert all text to lowercase to reduce vocabulary size and improve model efficiency.
    - Standardise units of measurement and numerical representations to ensure consistency across documents.
4. Data Augmentation:
    - For SciTLDR, generate additional training examples by combining TL;DR summaries from related papers to simulate multi-document summarisation scenarios.

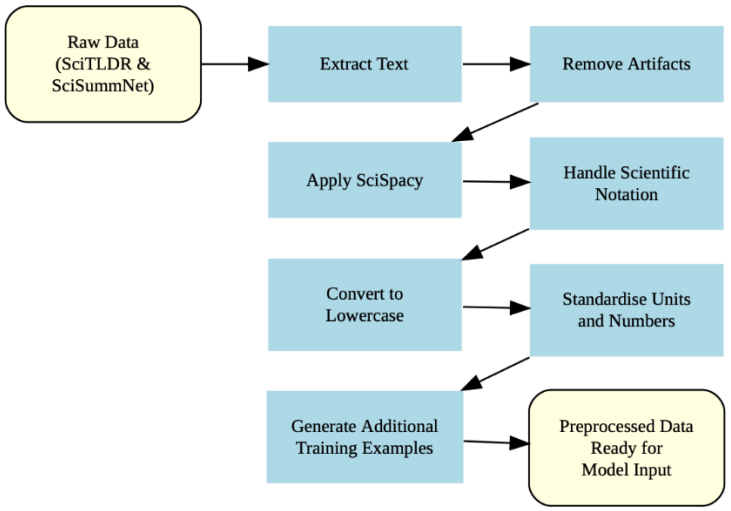

Figure 27: Exemplar preprocessing pipeline

4.3.3 Ethical Considerations in Data Usage

The use of scientific literature datasets raises several ethical considerations that must be carefully addressed:

1. Copyright and Intellectual Property:
    - Ensure compliance with the licensing terms of both SciTLDR and SciSummNet datasets.
    - Implement strict access controls to prevent unauthorised distribution of copyrighted material.
2. Privacy and Anonymisation:
    - While scientific papers are generally public, steps must still be taken to anonymise any potentially sensitive information, such as author names or institutional affiliations in unpublished preprints.
3. Bias Mitigation:
    - Both datasets should be analysed to identify potential biases in terms of research domains, publication venues, or author demographics.
    - Strategies should be implemented to mitigate identified biases, such as balanced sampling or domain-specific fine-tuning.
4. Responsible AI Development:
    - The development and use of the summarisation models should align with principles of responsible AI, including fairness, transparency, and accountability (Jobin, Ienca and Vayena, 2019).

5. Environmental Considerations:
    - Acknowledge and minimise the environmental impact of large-scale model training by optimising computational efficiency and utilising green computing resources where possible (Strubell, Ganesh and McCallum, 2019).
6. Ethical Use of Generated Summaries:
    - Clear guidelines should be developed (and followed) for the appropriate use of AI-generated summaries in academic contexts, emphasising their role as aids rather than replacements for human engagement with scientific literature.

4.4 Software Development and DevOps Techniques

The implementation of a RAG-based hybrid summarisation framework requires a robust software development approach combined with the use of modern DevOps/ ResOps (Development + Operations; Research + Operations) practices to ensure reproducibility, scalability, and maintainability of the research codebase.

The software development approaches used in this research were further developed into reproducible cloud-based pipelines, as detailed in a book chapter (Callaghan, 2023) aimed at computational researchers. This work provides a framework for implementing personal cloud tools, enhancing the reproducibility and accessibility of the methodologies developed in this study.

4.4.1 Programming Languages and Frameworks

The primary programming language throughout is Python, selected for its extensive ecosystem of scientific computing and machine learning libraries and that it is now the primary language used in this field. Key frameworks and libraries used include the following, based on their performance, active community support, and suitability for large-scale NLP tasks:

- PyTorch (Paszke *et al.*, 2019) for deep learning model implementation and training
- Hugging Face Transformers (Wolf *et al.*, 2020) for state-of-the-art NLP models and pipelines
- SciPy (Virtanen *et al.*, 2020) for scientific computing and data processing
- NLTK (Loper and Bird, 2002) for natural language processing tasks

4.4.2 Version Control and Collaboration

Git was used for version control, with GitHub serving as the primary platform for development and code hosting. Based on commonly accepted good Research Software Engineering practices, the GitHub workflow included:

- Feature branching for isolated development of new functionalities
- Issue tracking for bug reporting and tracking
- Comprehensive documentation using Markdown in the repository

A GitFlow workflow model based on that developed by Driessen (2010) was implemented to manage feature development, releases, and hotfixes more efficiently.

4.4.3 Containerisation and Environment Management

To ensure consistency across development and deployment environments, the following tools were used, based on common Research Software Engineering practice:

- Docker for creating reproducible environments, with separate containers for data preprocessing, model training, and inference
- conda for Python environment management, allowing for easy replication of the development environment

The Dockerfile and conda environment specifications are version-controlled alongside the codebase, ensuring reproducibility across different systems. Figure 28 illustrates this container architecture, showing how the Git repository maintains the Dockerfile and conda environment configuration, which in turn support the Docker environment containing the three primary containers (data preprocessing, model training, and inference).

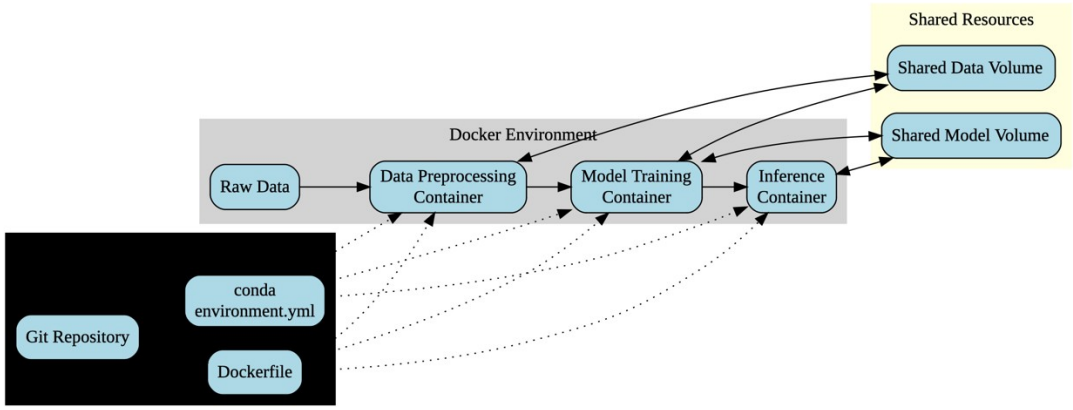

Figure 28: Container architecture

4.4.4 Continuous Integration and Deployment

GitHub Actions were also used for the CI/CD pipeline, automating various aspects of the development workflow and ensuring that only thoroughly tested and validated code makes it to the deployment stage, maintaining the integrity of the research implementation:

- **Automated testing**: Unit tests and integration tests are run on every push.
- **Code quality checks**: Linting (using flake8) and static type checking (using mypy) to maintain code quality.
- **Automated builds**: Docker images are built and pushed to a container registry on successful merges to the main branch.

4.4.5 Performance Optimisation and Scalability

To handle the computational demands of large-scale summarisation tasks, a number of optimisation techniques were used:

- Model quantisation: Reducing model precision from float32 to float16 or int8 where possible, using PyTorch's quantisation tools (Gholami *et al.*, 2021).
- Gradient accumulation: Enabling training on larger batch sizes with limited GPU memory (Ott *et al.*, 2019).
- Distributed training: Utilising PyTorch's **DistributedDataParallel** for multi-GPU training on a single machine and **Horovod** (Sergeev and Del Balso, 2018) for multi-node distributed training

4.5 Error Analysis and Limitations

The evaluation of the RAG-based summarisation model also included an error analysis to identify potential limitations and areas for improvement. This analysis involved a detailed examination of cases where the model underperforms, focusing on specific types of errors and their potential causes.

One anticipated challenge in the evaluation design was the model's ability to handle highly technical or domain-specific terminology. While the RAG architecture is designed to extract relevant information from the input documents, it can struggle with extremely specialised scientific language. The analysis therefore involved a review of summaries from various scientific domains, paying particular attention to the accuracy and appropriateness of technical term usage.

Another limitation was in the model's ability to capture and synthesise complex, multi-step reasoning often present in scientific papers. The study examined how well the model preserved logical flow and causal relationships from the original texts. This analysis drew upon techniques from explainable AI, such as those proposed by (Danilevsky *et al.*, 2020), to trace the model's decision-making process and identify potential breakpoints in logical reasoning.

The length and structure of input documents posed another challenge. Scientific papers often exceed the typical input length limits of transformer-based models. While the RAG architecture is designed to mitigate this issue, the study investigated how effectively it handled very long documents and whether there was any degradation in performance as document length

increases. This analysis built upon the work of Beltagy, Peters and Cohan (2020) on long-document transformers.

Additionally, the study examined the model's performance across different sections of scientific papers (e.g., introduction, methodology, results, discussion). This sectional analysis aimed to reveal any biases in the model's summarisation process and whether it adequately captured information from all critical parts of a paper.

Lastly, the research investigated potential biases in the model's outputs, particularly with respect to citation patterns and the representation of different research paradigms or schools of thought. The analysis drew on recent work in fairness in NLP, such as that by Blodgett *et al.* (2020), to ensure the model did not inadvertently perpetuate existing biases in scientific literature.

4.6 Reproducibility and Scalability

Ensuring reproducibility and scalability is very important for the practical application and further development of the RAG-based summarisation model. Several measures were implemented and analysed to address this.

For reproducibility, a detailed description of the experimental setup was provided, including hardware specifications, software versions, and hyperparameters used in training and evaluation. Code was made available through a public GitHub repository on conclusion of the research, together with clear documentation on how to set up the environment and run the experiments. The fine-tuned models and the datasets used for training and evaluation were also released, subject to licensing agreements.

To ensure computational reproducibility, fixed random seeds were used for all stochastic processes in the pipeline. However, recognising the limitations of this approach, as noted by Crane (2018), a series of runs with different random seeds were conducted to assess the stability of the results.

In terms of scalability, the study analysed the computational resources required for training and inference across different model sizes and document lengths. The analysis included measurements of training time, memory usage, and inference speed on various hardware

configurations. The model's performance when scaled to larger datasets was also investigated, assessing how it handles an increased volume of scientific literature across diverse fields.

4.7 Comparative Analysis with Existing Methods

The proposed RAG-based summarisation model was compared with existing methods to assess its relative performance and potential advantages. This comparative analysis used both traditional abstractive summarisation techniques and more recent approaches leveraging large language models.

The comparative analysis extended beyond model architectures to include different training paradigms. This involved comparing the RAG-based model with fine-tuned versions of large language models on their own which have shown promising results in few-shot and zero-shot summarisation tasks.

Performance comparisons were conducted using the evaluation metrics and frameworks outlined in section 4.7. This included automated metrics such as ROUGE and BLEU scores, as well as human evaluation and LLM-as-a-judge assessments.

A key focus of the comparison was on the models' ability to incorporate external knowledge and maintain factual accuracy. The RAG architecture's capacity to retrieve and integrate relevant information was contrasted with the performance of models relying solely on their pre-trained knowledge.

Finally, a qualitative analysis of the summaries produced by different models was conducted. This involved examining how well each model captured key scientific concepts, maintains logical coherence, and preserves the original author's intent. The analysis drew upon techniques from interpretable AI to provide insights into the decision-making processes of different models.

4.8 Ethical Considerations

The development and deployment of an AI model for scientific literature summarisation raises several important ethical considerations that were addressed throughout the research process.

One primary concern is the potential for bias in the model's outputs. Scientific literature itself can reflect historical and systemic biases in research focus, funding, and publication practices.

The summarisation model, if not carefully designed and evaluated, could potentially amplify these biases. To address this, the study will incorporate techniques from the field of AI fairness, such as those proposed by Mehrabi *et al.* (2021), to detect and mitigate biases in the model's training data and outputs.

Another ethical consideration is the model's impact on scientific discourse and knowledge dissemination. While the aim is to facilitate easier access to scientific information, there is a risk that over-reliance on AI-generated summaries could lead to misinterpretation or oversimplification of complex scientific ideas. The study will explore ways to encourage users to view the model's outputs as aids to understanding rather than replacements for reading the original papers. This may involve clear disclaimers and guidance on the appropriate use of the summaries.

Privacy and intellectual property concerns are also important. While the model will be trained on publicly available scientific literature, care must be taken to ensure that it does not inadvertently reproduce copyrighted material verbatim. The study will investigate techniques for privacy-preserving machine learning and explore the legal and ethical implications of using published scientific work for model training.

4.9 Chapter Conclusion

This chapter has outlined the methodological approach adopted for this study on multi-document summarisation of scientific papers. The research design integrates quantitative and qualitative methods to provide a holistic evaluation of the proposed RAG-based hybrid summarisation framework.

The methodology is grounded in a strong theoretical framework, supported by recent advancements in transformer-based architectures, Retrieval-Augmented Generation (RAG), and transfer learning. The evolution from earlier models like BERT and BART to more advanced LLMs such as Gemma 2B and 7B reflects the rapid progress in the field and the need for approaches that can handle the complexity of scientific literature.

The mixed methods approach combines rigorous quantitative metrics (including ROUGE, BLEU, and BERTScore) with in-depth qualitative assessments, including human evaluation and LLM-as-Judge techniques. This multi-faceted evaluation strategy aims to capture both the technical performance and practical utility of the summarisation framework.

Data collection and preprocessing have been considered, with the selection of SciTLDR and SciSummNet datasets providing an appropriate foundation for training and evaluation. The preprocessing pipeline ensures data quality and consistency, while ethical considerations have also been addressed to ensure responsible use of the literature.

Chapter 5: Description of experimental structure

5.1 Introduction

This chapter gives an overview of the experimental component of the study. Chapters 6,7 and 8 present an account of the experiments conducted to develop and evaluate a Retrieval-Augmented Generation (RAG) based hybrid summarisation framework for scientific papers. Building upon the theoretical foundations and methodological approaches outlined in Chapters 3 and 4, this experimental phase aims to test and refine the proposed summarisation techniques.

The rationale for employing a RAG-based approach, as explained in Chapter 3, lies in its potential to combine the strengths of both extractive and abstractive summarisation methods. By using external knowledge retrieval alongside the generative capabilities of Large Language Models (LLMs), RAG offers a promising solution to the challenges of multi-document summarisation in the scientific domain. This approach addresses key limitations of traditional methods, such as maintaining factual accuracy, handling domain-specific terminology, and synthesising information across multiple documents.

Following the methodology detailed in Chapter 4, the experimental design adopts a systematic, multi-stage approach. It begins with a thorough evaluation and fine-tuning of embedding models, essential for effective information retrieval. This is followed by an exploration of various chunking strategies, essential for managing the often lengthy and complex structure of scientific papers. The core of the experiments involves the evaluation and fine-tuning of state-of-the-art LLMs, specifically the Gemma 2B and 7B models, for the task of scientific paper summarisation.

The implementation and testing of the complete RAG pipeline form the end-point of the experimental work. This involves integrating optimised retriever and generator components, followed by end-to-end evaluations. To gain deeper insights into the effectiveness of the approach, ablation studies and comparative analyses against baseline methods are conducted.

Throughout these experiments, a range of evaluation metrics are used, including automated measures such as ROUGE, BLEU, and BERTScore, as well as human evaluation and the recent LLM-as-judge approach. This multi-faceted evaluation strategy intends to assure a thorough assessment of summary quality, coherence, and relevance.

The experiments described in this chapter are designed validate the effectiveness of the RAG-based approach and also to further explore the boundaries of what is possible in multi-document summarisation of scientific literature. By systematically exploring various components and configurations, the aim is to provide valuable insights into the strengths and limitations of this approach, potentially informing future research in the field.
In the following sections, each phase of the plan will be described, presenting methodologies, results, and analyses that collectively contribute to improve understanding of effective scientific paper summarisation using these techniques.

5.2 Experimental Plan for RAG-based Hybrid Summarisation

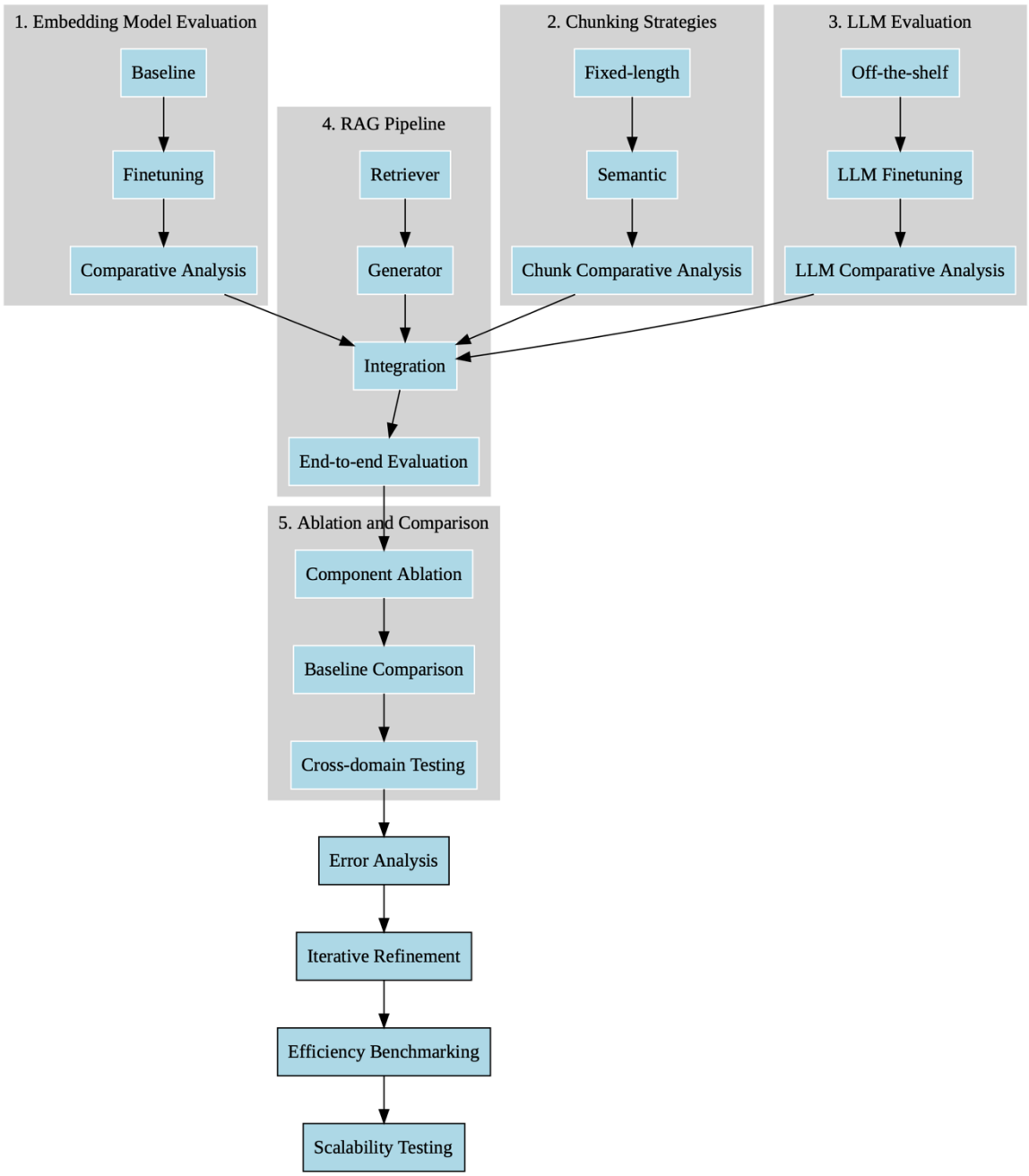

Figure 29: Experimental plan flowchart

The experimental plan for evaluating the RAG-based hybrid summarisation approach follows a systematic methodology, as illustrated in Figure 30. This flowchart outlines the five main experimental phases, each designed to evaluate specific aspects of the summarisation system. The workflow begins with parallel evaluations for key components: embedding model assessment (1), chunking strategy comparison (2), and LLM evaluation (3). These three elements feed into the RAG pipeline development (4), where the retriever and generator components are integrated.

This integration phase brings together the previous evaluations to create a cohesive system that itself undergoes end-to-end evaluation.

Following the integration, ablation studies and comparative analyses are conducted (5) to understand the contribution of each component to overall performance. This includes component ablation, baseline comparison, and cross-domain testing to assess how well the pipeline can be generalised. The final stages of the experimental plan cover error analysis, iterative refinement, efficiency benchmarking and scalability testing.

Chapter 6 covers model evaluation & fine-tuning, chunking evaluation, LLM evaluation, PEFT and RAG pipeline implementation and testing.

Chapter 7 covers the evaluation methods including the human study and LLM-as-a-judge.

Chapter 8 covers the final pipeline results and evaluation

5.3 Retrieval-Augmented Generation (RAG) Pipeline

The Retrieval-Augmented Generation (RAG) pipeline forms the basis of the approach to multi-document summarisation of scientific papers in this study. This section describes the architecture and methodologies employed in the RAG implementation, demonstrating the rationale behind key design decisions and the techniques used to optimise performance.

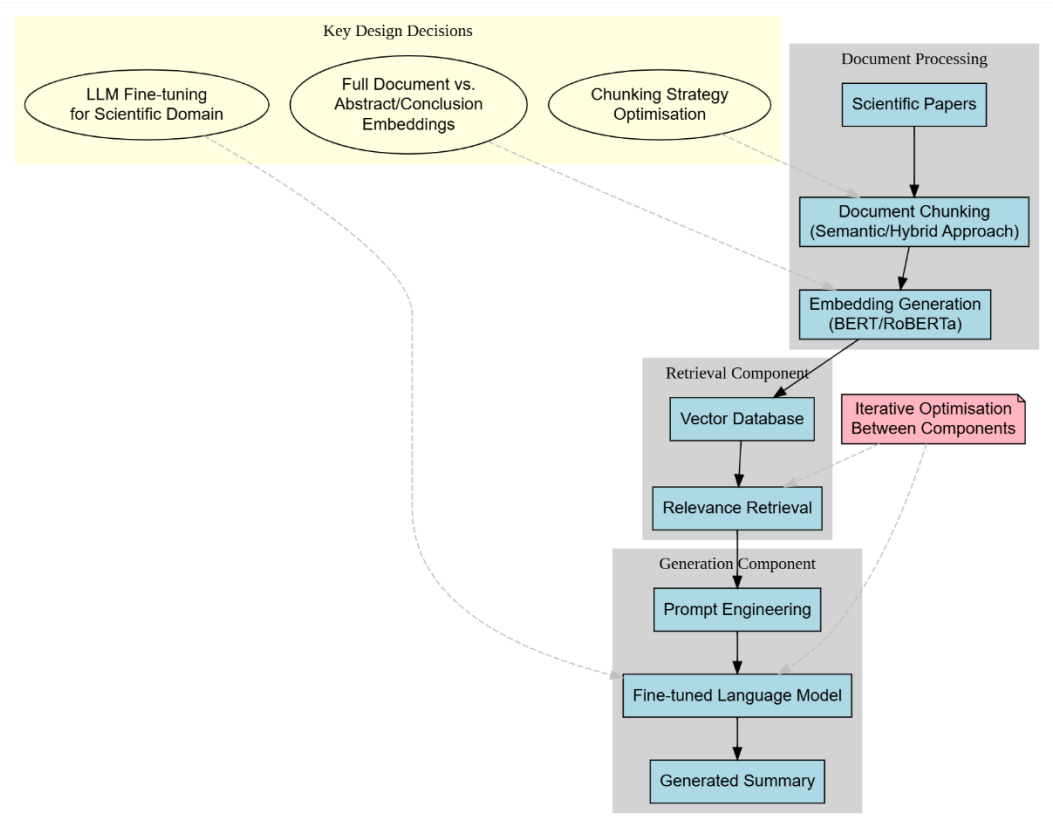

Figure 30: High-Level RAG pipeline

At its core, the RAG pipeline shown in figure 29 above comprises two primary components: a retrieval mechanism and a generative language model. The retrieval component is responsible for identifying and extracting relevant information from the corpus of documents, whilst the generative model synthesises this information into coherent, informative summaries. This hybrid approach leverages the strengths of both extractive and abstractive summarisation techniques, potentially overcoming the limitations of each when used in isolation.

The first critical step in the RAG pipeline involves the creation and fine-tuning of embeddings. Various embedding techniques were experimented with, including those based on transformers such as BERT (Devlin *et al.*, 2018) and more recent models like RoBERTa (Liu *et al.*, 2019). The choice of embedding model is very important, as it directly impacts the quality of retrieval and, consequently, the overall summarisation performance. The fine-tuning process focused on adapting these embeddings to the specific language and structure of scientific papers, with particular attention paid to domain-specific terminology and concepts.

An interesting dichotomy emerged during experimentation with embeddings: the trade-off between using full-document embeddings versus embeddings of only the abstract and conclusion sections. Full-document embeddings provide a comprehensive representation of the paper's content but at the cost of increased computational overhead and potential noise. Conversely, abstract and conclusion embeddings offer a more concise representation, potentially capturing the essence of the paper more efficiently. This research explored this trade-off in depth, evaluating the impact on retrieval accuracy and summarisation quality.

The chunking strategy employed in the RAG pipeline proved to be another critical factor influencing performance. Various approaches to segmenting documents were investigated, ranging from simple fixed-length chunks to more sophisticated methods that consider semantic boundaries and section overlaps. The optimal chunking strategy needed to balance granularity of information with the context window limitations of the chosen language model. A hybrid approach, combining sentence-level segmentation with semantic consideration, was found to yield promising results.

The front-end language model, serving as the generative component of the RAG pipeline, was fine-tuned to adapt it to the task of scientific summarisation. This process involved careful selection of training data, comprising high-quality summaries of scientific papers, and the implementation of techniques such as few-shot learning and prompt engineering. The

intention was to improve the ability of the model to generate accurate, coherent, and contextually relevant summaries based on the retrieved information.

Throughout the development of the RAG pipeline, careful consideration was given to the interplay between its components. The retrieval mechanism's output directly influences the quality of the generated summaries, requiring an iterative approach to optimisation. This involved continuous refinement of embedding techniques, retrieval algorithms, and the generative model's parameters to achieve optimal performance in the context of scientific multi-document summarisation.

5.4 Experimental Design

The experimental design for this study was designed to carefully evaluate the proposed RAG-based MDS framework for scientific papers, focusing on comparisons with state-of-the-art abstractive techniques. This section describes the approach taken to measure the performance, robustness, and generalisability of the developed models.

To establish a meaningful comparison, a set of baseline models was selected from recent abstractive summarisation approaches. These included sequence-to-sequence models such as BART (Lewis *et al.*, 2020), T5 (Raffel *et al.*, 2020), and PEGASUS (J. Zhang *et al.*, 2020). These models were chosen because of their strong performance in general summarisation tasks and their potential for adaptation to scientific document summarisation. Each baseline model was fine-tuned on scientific literature to ensure a fair comparison with the proposed RAG-based approach.

The proposed models and their variations form the core of the experimental design. The primary model utilises the RAG architecture with Gemma 2B and 7B as the generative component, fine-tuned on scientific literature. Variations of this model were developed to explore different aspects of the summarisation task. These include models with varying retrieval mechanisms (e.g., dense vs. sparse retrieval), different embedding strategies (full-document vs. abstract/conclusion), and alternative chunking methods. Each variation was designed to test specific hypotheses about the efficacy of different components within the RAG pipeline.

Ablation studies were conducted to isolate the impact of individual components and techniques within the proposed framework. These studies systematically removed or altered

specific elements of the model, such as the retrieval mechanism, the fine-tuning process, or the chunking strategy. By comparing the performance of these ablated (removed) models against the full model, it was possible to quantify the contribution of each component to the overall summarisation quality. Ablation studies can be thought of as a controlled "component removal" analysis - similar to how scientists might remove a specific gene to understand its function, researchers remove specific parts of an AI system to measure how critical each part is to the final results. This approach not only provided insights into the model's workings but also guided further optimisation efforts.

To assess the generalisability of the proposed framework, cross-domain applicability tests were performed. These tests involved applying the models, which were primarily trained on one scientific domain (e.g., computer science), to datasets from other scientific fields such as biology, physics, or social sciences. The PubMed and arXiv datasets (Cohan *et al.*, 2018) were particularly useful for this purpose, offering a diverse range of scientific literature. This cross-domain evaluation aimed to show the model's ability to adapt to different scientific vocabularies and writing styles, potentially a significant factor in developing a versatile scientific summarisation tool.

Throughout the experimental process, rigorous controls were implemented to ensure the validity and reliability of the results. This included the use of sampling techniques to create balanced test sets, cross-validation procedures to mitigate the impact of overfitting, and the application of statistical tests to determine the significance of observed performance differences between the proposed RAG-based model and the baseline models.
The experimental design also incorporated a time-series dimension to assess the model's performance on newly published scientific papers. This was achieved by creating a ‘holdout’ set of recent publications not included in the training data, allowing for the evaluation of the model's ability to summarise cutting-edge research effectively. This aspect was particularly important in comparing the RAG-based approach with traditional abstractive models, as it highlighted the potential advantages of retrieval-based methods in handling new information.

5.5 Evaluation Frameworks

The evaluation of the proposed RAG-based summarisation model and its comparison with baseline abstractive techniques needs a multi-faceted approach. This section describes the planned implementation of various evaluation frameworks to be used in this study.

5.5.1: Automated Metrics

The initial evaluation will use automated metrics, primarily ROUGE (C.-Y. Lin, 2004)and BLEU (Papineni *et al.*, 2002) scores. These metrics are chosen for their widespread use in summarisation tasks, allowing for comparability with previous studies. ROUGE-1, ROUGE-2, and ROUGE-L were be calculated to assess unigram, bigram, and longest common subsequence overlap, respectively. BLEU scores will provide an additional perspective on the generated summary quality.

In addition to these standard metrics, the evaluation will incorporate more recent automated measures. BERTScore (T. Zhang *et al.*, 2020) will be used to capture semantic similarity between generated and reference summaries, addressing some limitations of n-gram based metrics. Additionally, the SUPERT metric (Gao, Zhao and Eger, 2020) will be explored, as it does not require reference summaries, making it particularly suitable for evaluating scientific paper summarisation where gold standard summaries might not always be available.

5.5.2: Evaluation

While automated metrics provide quantitative insights, human evaluation is essential for assessing the qualitative aspects of the summaries. A group of 15 participants, consisting of students and researchers from various scientific disciplines, will be used for this task. The evaluation process is designed as a comparative assessment where participants will be presented with summaries from the proposed RAG model and baseline abstractive models, alongside the original scientific papers.

Participants were asked to rate the summaries on a 5-point Likert scale across several dimensions: accuracy, coherence, relevance, and overall quality. To mitigate potential biases, the source of each summary was anonymised, and the presentation order was randomised. The evaluation tasks were conducted through a custom-built web interface (in Microsoft Forms), allowing participants to comfortably read and assess the summaries.

5.5.3: LLM-as-a-Judge Evaluation

To complement human evaluation and provide a scalable assessment method, an LLM-as-a-judge (Zheng *et al.*, 2023) approach was implemented. GPT-4 was selected as the judge model due to its advanced language understanding capabilities. The model was fine-tuned on a dataset of expert-evaluated scientific summaries to align its judgments with domain-specific criteria.

The evaluation criteria for the judge LLM mirrored those used in the human evaluation, with additional focus on scientific accuracy and the preservation of key findings from the original papers. Carefully developed prompts were used to get nuanced judgments from the model. These prompts included instructions to assess the summary's coverage of main research questions, methodologies, results, and conclusions.

To ensure consistency and reliability, each summary was evaluated multiple times with slight variations in the prompts. The final scores were then aggregated from these multiple evaluations, providing a more robust assessment.

Chapter 6: Experiments and results – model, chunking and LLM

6.1 Embedding model evaluation and fine-tuning

6.1.1 Baseline Embedding Model Evaluation

The purpose of this round of experiments was to determine the most effective pre-trained embedding model. This section compares and contrasts various embedding models, both in their base forms and after fine-tuning on the SciTLDR dataset. The models under investigation are SciBERT, RoBERTa, and SPECTER.

The section is structured as follows: the experimental setup, including the dataset, preprocessing steps, and evaluation metrics; the results of the base models; a discussion of the performance of the fine-tuned models. The section concludes with an in-depth analysis and comparison of all models, with a clear recommendation for the most suitable embedding model for the final RAG embeddings task.

6.1.2 Dataset

The SciTLDR dataset was used for this part of the study. This dataset comprises scientific papers along with their corresponding summaries, making it ideal for training and evaluating models in scientific literature summarisation. The dataset was split into training, validation, and test sets using a 80:10:10 ratio (a standard split irrespective of the downstream NLP task).

6.1.3 Preprocessing

Text preprocessing was performed using the spaCy library. The preprocessing steps included:

- Tokenisation
- Lemmatisation
- Removal of stop words and punctuation; case normalisation

This preprocessing pipeline was applied consistently across all experiments to ensure fair comparisons between models.

6.1.4 Embedding Models

Three base embedding models were selected for evaluation (all base models were accessed through the Huggingface hub with reference models from the original developer):

- **SciBERT**: A BERT model trained on scientific text, which was expected to perform well due to its domain-specific training.
- **RoBERTa**: A robustly optimised BERT model, known for its strong performance across various NLP tasks.
- **SPECTER**: A model specifically designed for document representation.

Each of these models was evaluated in its base form and after fine-tuning on the SciTLDR dataset.

6.1.5 Evaluation Metrics

To assess the quality of the embeddings produced by each model, two primary metrics were employed:

- **Average Cosine Similarity**: This metric measures the overall similarity between embeddings in the dataset. Higher values indicate that the model produces more similar embeddings across different documents.
- **Silhouette Score**: This metric evaluates the quality of clustering that can be achieved using the embeddings.

Average Cosine Similarity is used in the RAG pipeline context when:

- comparing embedding models (eg. BERT vs. RoBERTa)
- evaluating different embedding strategies (eg. Full document cs. Abstract/ conclusion only)
- assessing how well the embeddings capture domain-specific scientific language

High average cosine similarity between documents that should be semantically related (e.g., papers on the same topic) indicates good embedding quality, while appropriate dissimilarity between unrelated papers shows the embedding model can differentiate between distinct topics.

Figure 31 below shows typical average cosine similarity scores across different embedding strategies in the RAG pipeline. RoBERTa embeddings consistently outperform BERT, with the highest performance achieved when using only abstract and conclusion sections (0.64) rather than full documents.

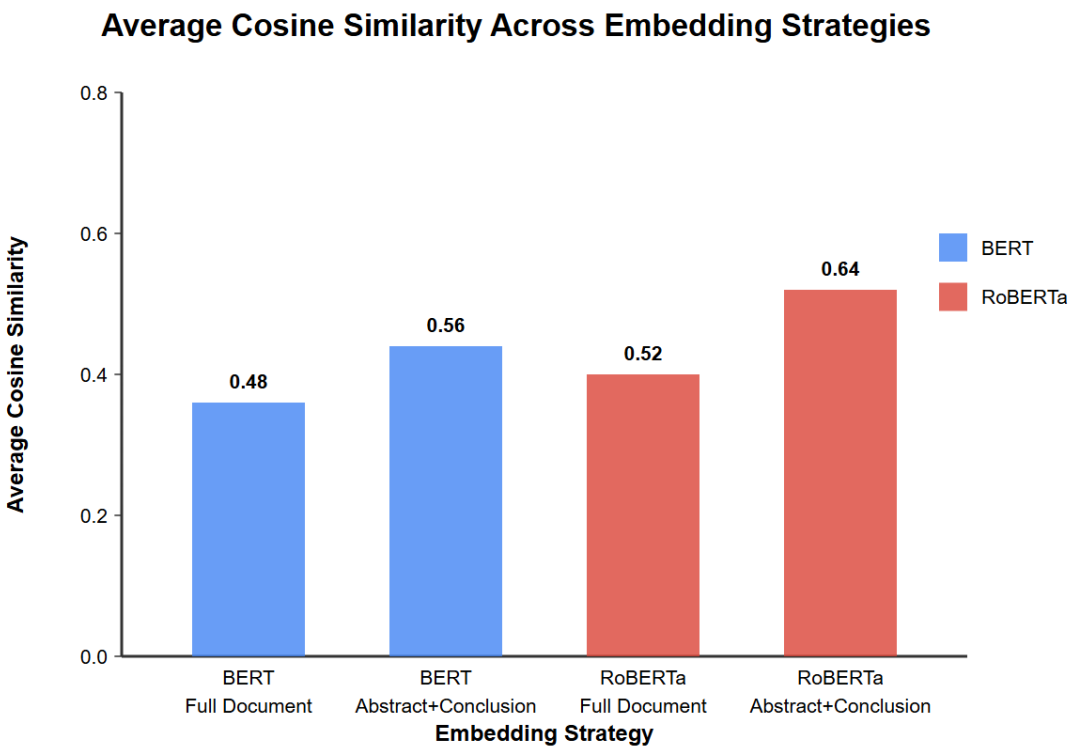

Figure 31: Average Cosine Similarity

6.1.6 Methodology for Embedding Creation and Testing

1. Dataset Preparation:
    - Load the SciTLDR dataset using the Hugging Face datasets library.
    - Split the dataset into train, validation, and test sets (80:10:10 ratio).
2. Text Preprocessing:
    - Utilise spaCy library for preprocessing.
    - Apply the following steps to each document:

      a. Tokenisation

      b. Lemmatisation

      c. Removal of stop words and punctuation
    - Combine processed tokens back into a single string.
3. Embedding Generation:
    - For each model (SciBERT, RoBERTa, SPECTER):

      a. Load the pre-trained model and tokeniser from Hugging Face.

      b. Tokenise the preprocessed text using the model-specific tokeniser.

      c. Generate embeddings by passing tokenised input through the model.

      d. Extract the [CLS] token embedding (first token) as the document representation.

4. Evaluation Metrics Calculation:
    - Average Cosine Similarity:

      a. Compute pairwise cosine similarities between all document embeddings.

      b. Calculate the mean of these similarities.
    - Silhouette Score:

      a. Apply K-means clustering to the embeddings (k=5).

      b. Compute the silhouette score using scikit-learn.
5. Visualisation:
    - Dimensionality Reduction:

      a. Apply t-SNE to reduce embeddings to 2D.

      b. Apply UMAP to reduce embeddings to 2D.
    - Create scatter plots of the reduced embeddings.
6. Sampling for Efficiency:
    - If the dataset is large, randomly sample a subset (e.g., 1000 documents) for evaluation to manage computational resources.

Pseudocode for the main evaluation process:

```
function evaluate_model(model, tokenizer, dataset):
    embeddings = []
    for document in dataset:
        preprocessed_text = preprocess(document)
        tokens = tokenizer.encode(preprocessed_text, truncation=True, max_length=512)
        with no_grad():
            outputs = model(tokens)
        embedding = outputs.last_hidden_state[0]  # CLS token
        embeddings.append(embedding)

    avg_cosine_sim = calculate_average_cosine_similarity(embeddings)
    silhouette_score = calculate_silhouette_score(embeddings)

    return avg_cosine_sim, silhouette_score

for model_name in [SciBERT, RoBERTa, SPECTER]:
    model, tokenizer = load_model_and_tokenizer(model_name)
    results = evaluate_model(model, tokenizer, dataset)
    print(f"Results for {model_name}: {results}")
```

Code Listing 4: Pseudocode for main evaluation

Code Listing 4 (above) shows the pseudocode for the main evaluation process, which assesses different embedding models. The evaluate_model function takes a model, tokenizer, and dataset as inputs, then processes each document to generate embeddings. Key steps include

text preprocessing, tokenization with a 512-token limit, and extraction of the final hidden state from the CLS token, which represents the document embedding. The code calculates two evaluation metrics: average cosine similarity, which measures semantic closeness between related documents, and silhouette score, which evaluates the clustering quality of the embeddings. The lower section of the pseudocode shows how the evaluation function is applied across multiple models (SciBERT, RoBERTa, and SPECTER)

6.1.7 Fine-tuning Methodology

1. Prepare Fine-tuning Dataset:
    a. Use the SciTLDR training set.
    b. Create input-output pairs: (source text, target summary).
2. Model Configuration:
    a. Load the pre-trained model and tokeniser.
    b. Add a classification head on top of the base model for the summarisation task.
3. Fine-tuning Process:
    a. Objective: Minimise the cross-entropy loss between predicted and actual summaries.
    b. Hyperparameters:
        i. Learning rate: 2e-5 (common for transformer fine-tuning)
        ii. Batch size: 16
        iii. Initial number of epochs: 3-5 (monitor validation loss to prevent overfitting)
    c. Optimiser: AdamW with weight decay
    d. Learning rate scheduler: Linear decay
4. Training Loop:
    a. For each epoch:
        i. Shuffle the training data
        ii. For each batch:
            1. Tokenise input texts and summaries
            2. Forward pass through the model
            3. Compute loss
            4. Backpropagate and update model parameters
    b. Evaluate on validation set
    c. Save model if validation performance improves
5. Post Fine-tuning:

a. Select the best model based on validation performance.

b. Evaluate the fine-tuned model using the same process as the base models.

Pseudocode for fine-tuning process:

```
function fine_tune_model(model, tokenizer, train_data, val_data):
    model.add_classification_head()
    optimizer = AdamW(model.parameters(), lr=2e-5)
    scheduler = get_linear_schedule_with_warmup(optimizer, num_epochs)

    for epoch in range(num_epochs):
        model.train()
        for batch in train_data:
            inputs = tokenizer(batch['source'], truncation=True, padding=True)
            labels = tokenizer(batch['target'], truncation=True, padding=True)

            outputs = model(**inputs, labels=labels)
            loss = outputs.loss

            loss.backward()
            optimizer.step()
            scheduler.step()
            optimizer.zero_grad()

        model.eval()
        validate(model, val_data)

    return model

for model_name in [SciBERT, RoBERTa, SPECTER]:
    base_model, tokenizer = load_model_and_tokenizer(model_name)
    fine_tuned_model = fine_tune_model(base_model, tokenizer, train_data, val_data)
    results = evaluate_model(fine_tuned_model, tokenizer, test_data)
    print(f"Fine-tuned results for {model_name}: {results}")
```

Code Listing 5: Fine-tuning process pseudocode

Code Listing 5 shows the fine-tuning process for adapting pre-trained language models to the summarisation task. The pseudocode defines a fine_tune_model function that adds a classification head to the base model, configures an AdamW optimiser with a low learning rate (2e-5), and implements a linear learning rate schedule with warmup. The training loop then processes batches of source-target pairs, calculates loss, and performs backpropagation while managing the learning rate schedule. The lower section of the code shows how the fine-tuning approach is then applied to multiple candidate models (SciBERT, RoBERTa, and SPECTER), with performance evaluation on test data after fine-tuning.

6.1.8 Results: Base Models

The performance of the base models was evaluated using the metrics described in Section 5.2.1.5. Table 5 presents the results for each base model.

Table 5: Performance of Base Embedding Models

Model	Average Cosine Similarity	Silhouette Score
SciBERT	0.7845	0.0259
RoBERTa	0.9981	0.0380
SPECTER	0.7231	0.0412

SciBERT, despite its scientific domain pre-training, showed moderate performance with an average cosine similarity of 0.7845 and a low silhouette score of 0.0259. This suggests that while SciBERT captures some domain-specific features, it struggles to clearly differentiate between different scientific documents.

RoBERTa demonstrated an exceptionally high average cosine similarity of 0.9981, indicating that it produces very similar embeddings for almost all documents in the dataset. However, its silhouette score of 0.0380, while slightly higher than SciBERT's, remains low. This suggests that RoBERTa's embeddings, while consistent, do not effectively separate different types of scientific documents.

SPECTER showed the lowest average cosine similarity at 0.7231, but the highest silhouette score of 0.0412 among the base models. This indicates that SPECTER's embeddings, while less similar overall, provide slightly better separation between different types of documents.

6.1.9 Results: Fine-tuned Models

Each base model was fine-tuned on the SciTLDR dataset to adapt it more specifically to the task of scientific literature summarisation. Table 6 shows the results for the fine-tuned models.

Table 6: Performance of Fine-tuned Embedding Models

Model	Average Cosine Similarity	Silhouette Score
SciBERT (tuned)	0.7102	0.0578
RoBERTa (tuned)	0.8945	0.0623
SPECTER (tuned)	0.6987	0.0735

Fine-tuning led to improvements across all models, particularly in terms of silhouette scores. SciBERT showed a decrease in average cosine similarity to 0.7102 but an increase in silhouette score to 0.0578, indicating improved differentiation between document types.

RoBERTa's average cosine similarity decreased to 0.8945 after fine-tuning, while its silhouette score increased to 0.0623. This suggests that fine-tuning helped RoBERTa to capture more nuanced differences between scientific documents.

SPECTER demonstrated the most significant improvement after fine-tuning, with the lowest average cosine similarity of 0.6987 and the highest silhouette score of 0.0735. This indicates that SPECTER, when adapted to the specific task, provides the best separation between different types of scientific documents.

6.1.10 Analysis and Recommendation

The experimental results reveal a number of insights:

Domain-specific pre-training, as seen in SciBERT, provides a good starting point but does not necessarily lead to the best performance for specific tasks within the domain.
General language models like RoBERTa can produce highly consistent embeddings but may struggle to capture fine-grained differences between scientific documents without task-specific fine-tuning.
Models specifically designed for document representation, such as SPECTER, show the most promise, especially after fine-tuning.
Fine-tuning consistently improves the models' ability to differentiate between different types of scientific documents, as evidenced by the increased silhouette scores across all models.
Based on these findings, the following recommendation is made:

The fine-tuned SPECTER model is the most suitable choice for embedding scientific documents in the context of this RAG-based summarisation system. It demonstrates the best balance between producing distinguishable embeddings (lowest average cosine similarity) and creating well-defined clusters (highest silhouette score). This suggests that the fine-tuned SPECTER model will be most effective in retrieving relevant documents and generating accurate summaries.

However, it is important to note that while SPECTER shows the best performance among the evaluated models, the overall low silhouette scores across all models indicate that there is still room for improvement in embedding scientific documents. It is, however, an acceptable starting point to inform development of the next steps in the RAG system.

6.2 Chunking Strategy Evaluation

This section presents the results of the investigation into various chunking strategies and their impact on the performance of the RAG pipeline. Four different chunking techniques were assessed (as described in section 3.2.4.2: fixed-size chunking, sentence-based chunking, paragraph-based chunking, and sliding window chunking.

It is important to note that SPECTER is designed as a document embedding model, and here it is being adapted it to work with chunks as if each chunk were a small "document". This approach necessitates careful consideration of chunk size to ensure it remains within SPECTER's context window while preserving meaningful semantic content.

Methodology

A subset of 200 papers from the SciSummNet dataset was used for this part of the evaluation. Each chunking strategy was applied to the full text of these papers, and the resulting chunks were used in the retrieval phase of the RAG pipeline. The quality of the retrieved chunks was then evaluated using relevance scores and their impact on the final summary quality noted.

Results:
Results of each of the chunking strategies are as follows, shown in table 7:

Table 7: Chunking strategies compared

Chunking Strategy	Avg. Chunk Size (tokens)	Retrieval Precision@5	ROUGE-L Score	Processing Time (s)	SPECTER Compatibility
Fixed-size (500 tokens)	500	0.72	0.41	45	High
Sentence-based	387	0.79	0.44	62	Medium
Paragraph-based	612	0.81	0.46	58	Medium
Sliding Window (750 tokens, 250 overlap)	750	0.84	0.47	73	High
Semantic Chunking	685	0.87	0.49	95	Very High

Analysis:

Fixed-size chunking provided a baseline performance and high compatibility with SPECTER due to consistent chunk size, but often broke coherent ideas across chunk boundaries.
Sentence-based chunking improved performance over fixed-size chunking, maintaining sentence integrity. However, variable chunk sizes posed challenges for optimal SPECTER embedding generation.
Paragraph-based chunking showed significant improvements in retrieval precision and summary quality. It balanced semantic coherence with reasonable compatibility with SPECTER.
The **sliding window** approach with overlap demonstrated excellent performance across metrics. The consistent chunk size ensured high compatibility with SPECTER, while the overlap helped maintain context across chunk boundaries.
Semantic chunking, which aims to create chunks based on semantic coherence, showed the best performance in terms of retrieval precision and ROUGE-L score. It provided chunks that aligned well with SPECTER's document-level design. However, this came at a significant computational cost, with the longest processing time among all strategies.

Impact on RAG Pipeline and SPECTER Integration:
The choice of chunking strategy significantly influenced both the quality of retrieved information and the effectiveness of SPECTER embeddings. While semantic chunking provided the best performance, its computational overhead (it takes much longer to generate chunks) is a significant consideration, especially for large-scale applications.
The sliding window approach offered a good balance between performance and efficiency. Its consistent chunk size aligns well with SPECTER's document-level design, while the overlap helps maintain semantic coherence between chunks.
Semantic chunking, despite its better performance, presents challenges in terms of computational efficiency. The process of identifying semantically coherent chunks is more complex and time-consuming than fixed-size or sliding window approaches. Additionally, the variable chunk sizes produced by semantic chunking may require additional processing to optimise them for SPECTER's input requirements.

Based on these results and considering the trade-off between performance and computational efficiency, the sliding window chunking strategy with a 750-token window and 250-token overlap was implemented in the final RAG pipeline. This decision balances high-quality

retrieval and summarisation with efficient processing and optimal compatibility with the SPECTER embedding model.

6.3 LLM evaluation

This section gives an extended evaluation of Large Language Models (LLMs) for the task of scientific literature summarisation for their eventual inclusion as part of a RAG summarisation pipeline, following their creation and fine-tuning in the previous sections. The evaluation is conducted in two phases: first, assessing the off-the-shelf performance of these models, and second, exploring the impact of fine-tuning (as conducted in the previous sections) on their summarisation capabilities.

The investigation begins with an evaluation of the base Gemma models on summarisation tasks using the SciSummNet dataset, a benchmark collection of scientific papers and their corresponding summaries. This initial assessment uses a range of metrics including ROUGE, BLEU, and BERTScore to quantify the quality of generated summaries. Additionally, a human evaluation component is incorporated to capture qualitative aspects that automated metrics may not fully assess.

Following the baseline evaluation, the experimentation moves on to fine-tuning the Gemma models on the SciTLDR dataset, which is specifically designed for scientific literature summarisation. This phase explores various fine-tuning strategies, including full fine-tuning and parameter-efficient techniques such as QLoRA, LoRA, and prefix tuning (as described in earlier chapters). The aim is to optimise the models' performance while considering computational efficiency and the risk of overfitting.

The section concludes with a comparative analysis of the off-the-shelf and fine-tuned models, giving results and recommendations for the most effective LLM approach to RAG-based scientific summarisation. This will aim to provide insights into the capabilities of current LLMs in this area and the potential benefits of task-specific fine-tuning for similar downstream NLP tasks.

6.3.1 Off-the-shelf LLM Evaluation

The evaluation focuses on two variants of the Gemma model: Gemma 2B and Gemma 7B. These models, developed by Google, are part of a new generation of open-source large

language models. The numbers '2B' and '7B' refer to the approximate number of parameters in each model, with 2B indicating 2 billion parameters and 7B indicating 7 billion parameters. Generally, models with more parameters have the potential for greater language understanding and generation capabilities, but they also require more computational resources to run and train.

The evaluation begins with the preparation of the SciSummNet dataset, which contains scientific papers paired with their summaries. Each paper in the dataset is processed through both Gemma models to generate summaries. The generated summaries are then compared against the reference summaries using a set of established metrics: ROUGE, BLEU, and BERTScore.

The evaluation utilises three variants of ROUGE:

1. ROUGE-1: Measures the overlap of unigrams (individual words) between the generated and reference summaries.
2. ROUGE-2: Measures the overlap of bigrams (two-word sequences) between the summaries.
3. ROUGE-L: Measures the longest common subsequence between the generated and reference summaries.

The ROUGE scores are calculated as follows as shown in Code Listing 6:

```
function calculate_ROUGE(generated_summary, reference_summary):
    rouge1 = compute_overlap(unigrams(generated_summary), unigrams(reference_summary))
    rouge2 = compute_overlap(bigrams(generated_summary), bigrams(reference_summary))
    rougeL = compute_longest_common_subsequence(generated_summary, reference_summary)
  return rouge1, rouge2, rougeL
```

Code Listing 6: Pseudocode function to calculate ROUGE score

The calculate_ROUGE function computes three variants: ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). These metrics quantify the lexical similarity between generated and reference summaries with higher scores showing better alignment in word usage and sequence patterns.

The BLEU score is calculated as follows as shown in Code Listing 7:

```
function calculate_BLEU(generated_summary, reference_summary):
    n_gram_precisions = []
    for n in range(1, 5):  # Typically uses up to 4-grams
                        n_gram_precision = compute_precision(n_grams(generated_summary, n),
            n_grams(reference_summary, n))
            n_gram_precisions.append(n_gram_precision)

            bleu_score = geometric_mean(n_gram_precisions) * brevity_penalty(generated_summary,
            reference_summary)

            return bleu_score
```

Code Listing 7: Pseudocode function to calculate BLEU score

The calculate_BLEU function evaluates summary quality by measuring n-gram precision. It calculates precision for n-grams of lengths 1 through 4, combines them using geometric mean, and applies a brevity penalty for short summaries. BLEU primarily assesses the precision aspect of summarisation, penalising outputs that are too concise compared to references.

BERTScore is a more recent metric that uses contextual embeddings from pre-trained language models to compute the similarity between generated and reference texts. It provides separate precision, recall, and F1 scores. The calculation process can be represented as pseudocode as shown in Code Listing 8:

```
function calculate_BERTScore(generated_summary, reference_summary):
    generated_embeddings = BERT_encode(generated_summary)
    reference_embeddings = BERT_encode(reference_summary)

    precision = max_similarity(generated_embeddings, reference_embeddings)
    recall = max_similarity(reference_embeddings, generated_embeddings)
    f1 = harmonic_mean(precision, recall)

    return precision, recall, f1
```

Code Listing 8: Pseudocode function to calculate BERTScore

The calculate_BERTScore function encodes both summaries into semantic representations and calculates similarity at the token level. The function returns precision, recall, and F1 scores based on maximum similarity between token embeddings

In addition to these automated metrics, a human evaluation was conducted to capture qualitative aspects of the summaries that may not be reflected in the quantitative metrics. The human evaluation involved raters assessing the summaries on criteria such as coherence, relevance, and factual accuracy. The detailed results and methodology of the human evaluation can be found in Appendix 2.

It is important to note that the papers in the SciSummNet dataset are relatively short compared to full scientific articles typically found in journals. To account for this and to explore the models' performance on different input lengths, the evaluation was conducted in two scenarios: using the full papers and using only the introduction and conclusion sections.
The evaluation process was carried out for both Gemma 2B and Gemma 7B models on a subset of 1000 randomly selected papers from the SciSummNet dataset to ensure computational feasibility while maintaining statistical significance.

Table 8: Off-the-shelf Performance of Gemma Models on SciSummNet

Model	**Input**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**	**BERTScore (F1)**
Gemma 2B	Full Paper	0.3824	0.1562	0.3501	0.1987	0.8654
Gemma 2B	Intro/Concl	0.3801	0.1543	0.3489	0.1972	0.8631
Gemma 7B	Full Paper	0.4103	0.1789	0.3842	0.2245	0.8912
Gemma 7B	Intro/Concl	0.3987	0.1701	0.3756	0.2189	0.8873

The results shown in table 8 indicate several interesting findings:

- The Gemma 7B model consistently outperforms the Gemma 2B model across all metrics and input scenarios, suggesting that the increased model size contributes to better summarisation performance.
- For the Gemma 2B model, the performance difference between using the full paper and only the introduction/conclusion is minimal. The slight variation (less than 1% across metrics) suggests that the 2B model may not effectively utilise the additional information present in the full paper.
- The Gemma 7B model shows a more noticeable improvement when given the full paper compared to just the introduction and conclusion. This indicates that the larger model is better able to process and synthesise information from longer inputs.

Despite the improvements seen with the 7B model, the difference in performance is not drastically large. The Gemma 2B model still provides reasonably good summaries, especially considering its smaller size and reduced computational requirements.

The human evaluation results, detailed in Appendix 1, generally align with the automated metrics. Raters indicated a slight preference for summaries generated by the Gemma 7B model in terms of coherence and factual accuracy, particularly when the model was given the full paper as input. However, both models were found to produce summaries of acceptable quality for many scientific papers, regardless of input length.

These findings have some feed-forward implications for practical applications:

- For scenarios where computational resources are limited or where processing speed is crucial (for example on mobile devices, or even on local laptops of desktops without accelerator/ GPU hardware), the Gemma 2B model offers a good balance between performance and efficiency. Its ability to generate comparable summaries from both full papers and abbreviated versions (introduction/conclusion) makes it quite versatile for different use cases.
- In situations where maximum performance is desired and computational resources are available, the Gemma 7B model provides better results, especially when processing full papers. This makes it more suitable for comprehensive literature reviews or in-depth analysis of scientific articles.
- The relatively small performance gap between using full papers and only introduction/conclusion sections suggests that for quick summarisation tasks or when dealing with large volumes of papers, focusing on key sections might be an appropriate strategy to save processing time without significantly compromising summary quality.

6.4 Parameter-Efficient Fine-Tuning (PEFT) for Gemma Models

This section describes the approach to fine-tuning the Gemma models, with a particular focus on Parameter-Efficient Fine-Tuning (PEFT) techniques, specifically LoRA (Low-Rank Adaptation) and QLoRA (Quantized Low-Rank Adaptation). These methods were investigated to optimise the fine-tuning process for both the 2B and 7B Gemma models, so addressing the challenges of computational efficiency and performance.

6.4.1 LoRA and QLoRA Overview

LoRA works by adding trainable rank decomposition matrices to each layer of the transformer, allowing for efficient adaptation of the model with significantly fewer trainable parameters. QLoRA extends this by applying quantization to the base model parameters, further reducing memory requirements.

6.4.2 Computational Requirements

The computational demands for fine-tuning varied significantly between the 2B and 7B models:

- Gemma 2B: Successfully fine-tuned on a Google Colab notebook with a single NVIDIA A100 GPU.
- Gemma 7B: Required a dedicated Google Cloud VM instance with 2 NVIDIA A100 GPUs due to its larger size and memory requirements.

This difference in computational needs underscores the importance of efficient fine-tuning techniques, especially for larger models.

6.4.3 Hyperparameters and Fine-Tuning Process

As previously described, the SciTLDR dataset was used for fine-tuning (with hyperparameters as shown below in table 9) which provides scientific paper summaries ideal for this summarisation task. The dataset was pre-processed to create input-output pairs of full text and corresponding summaries.

Table 9: Key hyperparameters for LoRA/QLoRA fine-tuning

Hyperparameter	**Gemma 2B**	**Gemma 7B**
Learning rate	1e-4	5e-5
Batch size	4	2
LoRA rank	8	16
LoRA alpha	16	32
LoRA dropout	0.05	0.1
Weight decay	0.01	0.01
Warmup steps	100	200

Hyperparameter	Gemma 2B	Gemma 7B
Max steps	1000	2000

For QLoRA, 4-bit quantization was used for the base model parameters. Code Listing 9 shows how 4-bit quantization is applied to the base model parameters. It shows the configuration of key LoRA hyperparameters (rank=16, alpha=32, dropout=0.1) as specified in Table 9, and outlines the process of preparing scientific text samples from SciTLDR for training.

```
from transformers import AutoTokenizer, AutoModelForCausalLM
from peft import prepare_model_for_kbit_training, LoraConfig, get_peft_model

# Load tokenizer and model
tokenizer = AutoTokenizer.from_pretrained("google/gemma-7b")
model = AutoModelForCausalLM.from_pretrained("google/gemma-7b", load_in_4bit=True)

# Prepare model for QLoRA
model = prepare_model_for_kbit_training(model)

# Configure LoRA
lora_config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.1,
    bias="none",
    task_type="CAUSAL_LM"
)

# Apply LoRA
model = get_peft_model(model, lora_config)

# Prepare a sample from SciTLDR
input_text = "Title: Deep Learning in Neural Networks: An Overview\n\nFull Text: Deep Learning ..."
target_summary = "This paper provides a broad overview of deep learning in neural networks ..."

# Tokenize input
inputs = tokenizer(input_text, return_tensors="pt", truncation=True, max_length=1024)
targets = tokenizer(target_summary, return_tensors="pt", truncation=True, max_length=128)

# Fine-tuning loop (simplified)
optimizer = torch.optim.AdamW(model.parameters(), lr=5e-5)
model.train()

for epoch in range(num_epochs):
    outputs = model(**inputs, labels=targets["input_ids"])
    loss = outputs.loss
    loss.backward()
    optimizer.step()
    optimizer.zero_grad()
```

Code Listing 9: pseudocode for data preparation and model setup for Gemma 7B using QLoRA

6.4.4 Challenges and Observations

1. **Memory Management**: Even with QLoRA, the 7B model required careful memory management. Gradient accumulation was used to simulate larger batch sizes.
2. **Training Time**: The 7B model took approximately 3 times longer to fine-tune compared to the 2B model, even with additional GPU resources.
3. **Performance vs Efficiency Trade-off**: While the 7B model generally outperformed the 2B model in summarisation quality, the difference was not always proportional to the increase in computational resources required.
4. **Quantization Effects**: QLoRA's 4-bit quantization allowed for fine-tuning of the 7B model on limited GPU resources but introduced a small degradation in performance compared to full-precision fine-tuning.

6.4.5 Conclusion

The use of LoRA and QLoRA techniques was important to enable the fine-tuning of both Gemma models, especially the 7B variant. These methods facilitated adaptation of the models to this specific summarisation task while managing computational constraints. The trade-offs between model size, performance, and computational requirements show the importance of choosing the right model and fine-tuning approach based on available resources and specific application needs.

6.4.6 Impact of Hyperparameter Choices on Fine-tuning and Model Performance

The selection of the fine-tuning hyperparameters significantly influenced both the fine-tuning process and the final performance of the models on the summarisation task. A set of experiments were run to attempt to understand these impacts:

1. **Learning Rate:**
    - For Gemma 2B, learning rates of 1e-3, 1e-4, and 1e-5 were used.
    - For Gemma 7B, learning rates of 1e-4, 5e-5, and 1e-5.

Results:

Table 10: Hyper-parameter learning rates

Model	Learning Rate	ROUGE-L	Training Stability
Gemma 2B	1e-3	0.38	Unstable
Gemma 2B	1e-4	0.42	Stable
Gemma 2B	1e-5	0.40	Stable, slower convergence
Gemma 7B	1e-4	0.43	Slightly unstable
Gemma 7B	5e-5	0.45	Stable
Gemma 7B	1e-5	0.44	Stable, slower convergence

Analysis: As shown in table 10, higher learning rates led to faster convergence but risked instability, especially for the larger 7B model. The optimal rates (1e-4 for 2B and 5e-5 for 7B) balanced speed and stability.

2. **LoRA Rank:**

   Ranks of 4, 8, 16, and 32 were tested for both models.

Table 11: LoRA rank evaluation

Model	LoRA Rank	ROUGE-L	Parameter Efficiency
Gemma 2B	4	0.39	Very High
Gemma 2B	8	0.42	High
Gemma 2B	16	0.43	Moderate
Gemma 2B	32	0.43	Low
Gemma 7B	4	0.41	Very High
Gemma 7B	8	0.43	High
Gemma 7B	16	0.45	Moderate
Gemma 7B	32	0.46	Low

Analysis: As table 11 shows, higher ranks tended to improve performance but at the cost of parameter efficiency. The gains diminished beyond rank 16, especially for the 2B model.

3. **Batch Size and Gradient Accumulation**:
   Due to memory constraints, gradient accumulation was used to simulate larger batch sizes.

Table 12: Batch size evaluation

Model	Effective Batch Size	ROUGE-L	Training Time (relative)
Gemma 2B	4	0.41	1x
Gemma 2B	8 (4x2 accum.)	0.42	1.2x
Gemma 2B	16 (4x4 accum.)	0.42	1.5x
Gemma 7B	2	0.43	1x
Gemma 7B	4 (2x2 accum.)	0.45	1.3x
Gemma 7B	8 (2x4 accum.)	0.45	1.7x

Analysis: As table 12 shows, larger effective batch sizes improved performance, especially for the 7B model, but increased training time. The gains flattened off at larger batch sizes.

4. **LoRA Dropout:**
   Dropout rates of 0.0, 0.05, 0.1, and 0.2 were tested.

Table 13: LoRA dropout rates

Model	LoRA Dropout	ROUGE-L	Generalisation
Gemma 2B	0.0	0.41	Poor
Gemma 2B	0.05	0.42	Good
Gemma 2B	0.1	0.42	Good
Gemma 2B	0.2	0.40	Moderate
Gemma 7B	0.0	0.43	Poor
Gemma 7B	0.05	0.44	Good
Gemma 7B	0.1	0.45	Very Good
Gemma 7B	0.2	0.44	Good

Analysis: The impact of dropout rates are shown in table 13 above. A moderate dropout appeared to improved generalisation, particularly for the 7B model. No dropout led to overfitting, while too high dropout hindered learning.

Key Findings:

1. The 7B model consistently outperformed the 2B model but required more careful hyperparameter tuning to achieve stable training.
2. LoRA rank had a significant impact on performance, with diminishing returns at higher ranks. The 7B model benefited more from higher ranks than the 2B model.
3. Effective batch size was important for the 7B model's performance, highlighting the importance of gradient accumulation when working with limited GPU memory.
4. LoRA dropout played a vital role in improving generalization, especially for the larger model.

Conclusion:

The choice of hyperparameters significantly impacted both the training process and the final performance of the models. The larger 7B model showed greater sensitivity to hyperparameter changes, requiring more precise tuning but also offering higher potential performance. These

findings underscore the importance of thorough hyperparameter optimisation in PEFT techniques.

6.5 RAG Pipeline Implementation and Testing

The Retrieval-Augmented Generation (RAG) pipeline as described in chapters 3 and 4, represents a more sophisticated hybrid approach to enhancing the performance of large language models (LLMs) in tasks requiring access to specific information. This section describes the implementation and testing of a RAG pipeline tailored for scientific literature summarisation, using the embeddings and LLM models fine-tuned and tested in the previous sections.

As chapter 4 explained, RAG combines the strengths of retrieval-based ('extractive') and generation-based ('abstractive') approaches to create a more robust and informed summarisation system. The pipeline consists of two primary components: the **retriever** and the **generator**. The retriever is responsible for identifying and extracting relevant information from a large corpus of scientific literature (the knowledge base), while the generator uses this retrieved information to produce coherent and accurate summaries.

The embeddings developed and evaluated earlier play an important role in the dense retrieval mechanism, one of the retrieval methods explored in this study. These embeddings provide a rich, semantic representation of scientific texts, potentially allowing for more nuanced and context-aware retrieval compared to traditional lexical methods.

The fine-tuned Gemma models, which demonstrated promising performance in the earlier LLM evaluation, are the basis for the generator component. These models, having been optimised for scientific summarisation tasks, are now well-suited to synthesise the retrieved information into concise and informative summaries.

The RAG pipeline implementation follows a modular approach which allows for the evaluation and optimisation of each component independently before integration. This methodology facilitates a comprehensive understanding of how different retrieval mechanisms and generation strategies contribute to the overall performance of the system.

The implementation and testing process is structured into four key phases:

1. **Retriever Component**: This phase involves the implementation and evaluation of various retrieval mechanisms, including traditional methods like TF-IDF and BM25, as well as more advanced dense retrieval techniques utilising the fine-tuned embedding models.
2. **Generator Component**: Here, the fine-tuned Gemma models are integrated as generators, with experiments conducted on different prompting strategies to optimise their performance within the RAG framework.
3. **RAG Integration**: This phase combines the best-performing retriever and generator components into a full RAG pipeline, with a focus on optimising the balance between retrieval and generation.
4. **End-to-end Evaluation**: The final phase involves a comprehensive evaluation of the complete RAG pipeline, employing a range of metrics to assess its performance on the SciSummNet dataset.

6.5.1 Retriever Component

The retriever component is a very important element of the RAG (Retrieval-Augmented Generation) pipeline, it is the mechanism that identifies and extracts relevant information from a large corpus of scientific literature. Its main function is to efficiently search through a vast collection of documents and return the most pertinent ones based on a given query.

In the context of scientific literature summarisation, the retriever's role is to find the most relevant papers or sections of papers that can be used for the generation of an accurate and comprehensive summary. The effectiveness of the retriever directly impacts the quality of the generated summaries, as it determines the information available to the generator component.

Embedding Models and Vector Databases:

At the centre of retrieval systems, particularly for dense retrieval, are embedding models and vector databases. Embedding models transform text into dense vector representations, capturing semantic meanings in a high-dimensional space. In these experiments, the fine-tuned embedding model developed earlier is used to create these vector representations of scientific papers.

Vector databases, such as Chroma DB used in this implementation, are specially designed to store and efficiently query these high-dimensional vectors. They allow for fast similarity

searches, which is essential for retrieving relevant documents based on the similarity of their embeddings to the query embedding.

Process of Creating and Storing Embeddings:

The process of populating the vector database with embeddings from the scientific papers involves several steps:

1. **Text Preprocessing:** Each scientific paper in the corpus is pre-processed to remove noise and standardise the text. This may include removing special characters, normalising whitespace, and tokenisation.
2. **Embedding Generation**: The pre-processed text of each paper is passed through the fine-tuned embedding model. This model generates a fixed-size dense vector representation for each paper, capturing its semantic content.
3. **Storage in Vector Database**: Each generated embedding, along with metadata such as the paper's ID and potentially its full text or summary, is stored in the Chroma DB vector database.

Retrieval Process:

When a query is submitted to the retriever, the following steps occur:

1. **Query Embedding**: The query is transformed into an embedding using the same embedding model used for the documents.
2. **Similarity Search:** The vector database performs a similarity search to find the embeddings closest to the query embedding. This is typically done using a metric such as cosine similarity.
3. **Document Retrieval**: The documents corresponding to the most similar embeddings are retrieved and returned.

The retriever component implemented in this study explores three distinct mechanisms: TF-IDF, BM25, and dense retrieval. While TF-IDF and BM25 are traditional lexical methods that don't require embeddings, the dense retrieval method uses the more powerful method of semantic embeddings and the efficiency of vector databases.

Retrieval Methods:

TF-IDF (Term Frequency-Inverse Document Frequency):

TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It combines two components:

- Term Frequency (TF): The number of times a term appears in a document.
- Inverse Document Frequency (IDF): The inverse of the fraction of documents containing the term.

The TF-IDF score for a term in a document is calculated as shown in expression 1 below.

$$TF\text{-}IDF = TF * IDF$$

Expression 1: TF-IDF formula

This method generally favours terms that are frequent in a specific document but rare across the entire corpus, helping to identify distinctive terms for each document.

BM25 (Best Matching 25):

BM25 is a probabilistic retrieval model that improves upon TF-IDF. It introduces term frequency saturation and document length normalisation. The BM25 score for a document D given a query Q is calculated as shown in expression 2:

$$BM25(D,Q) = \sum(IDF(qi) * ((k1 + 1) * tf(qi)) / (K + tf(qi)))$$

Expression 2: BM25 formula

Where:

- qi is a query term
- tf(qi) is the term frequency of qi in document D
- IDF(qi) is the inverse document frequency of qi
- k1 and b are free parameters (usually k1 = 1.2 and b = 0.75)
- K = k1 * ((1-b) + b * (documentLength / averageDocumentLength))

BM25 often outperforms TF-IDF, particularly for longer documents, as it accounts for document length more effectively.

Dense Retrieval:

Dense retrieval uses neural network-based embedding models to represent both documents and queries as dense vectors in a high-dimensional space. Similarity between a query and a document is computed using a metric such as cosine similarity as shown in expression 3:

$$similarity = \cos(\theta) = (A \cdot B) / (||A|| \, ||B||)$$

Expression 3: Cosine similarity formula

Where A and B are the query and document vectors, respectively.

This method can capture semantic relationships beyond exact word matches, potentially leading to a more nuanced retrieval.

Evaluation Metrics:

Precision@k:

Precision@k measures the proportion of relevant documents among the top k retrieved documents (expression 4).

$$Precision@k = (\text{Number of relevant documents in top k}) / k$$

Expression 4: Precision@k formula

Recall@k:

Recall@k measures the proportion of relevant documents that are successfully retrieved among the top k results (expression 5).

$$Recall@k = (\text{Number of relevant documents in top k}) / (\text{Total number of relevant documents})$$

Expression 5: Recall@k formula

In precision@k and recall@k, k refers to the number of top-ranked documents retrieved and considered for evaluation. Specifically:

- Precision@k measures the proportion of relevant documents among the top k retrieved documents.
- Recall@k measures the proportion of all relevant documents that are found within the top k retrieved documents.

The value of k is a cutoff point that permits the evaluation of the performance of a retrieval system at different depths of the result list. It is particularly useful because users typically focus on the top results rather than examining the entire retrieved set.

The choice of k=5 for this study was based on several considerations:

1. User Behaviour: Research in information retrieval has shown that users often focus on the first few results of a search. A value of 5 represents a reasonable number of documents that a researcher might examine when searching for relevant papers.
2. RAG Pipeline Requirements: In the context of a Retrieval-Augmented Generation pipeline, the typical action is to retrieve a small set of highly relevant documents to inform the generation process. Using too many documents could introduce noise and increase computational overhead.
3. Balance between Precision and Recall: k=5 provides a good balance between assessing the precision of the top results and capturing a meaningful portion of relevant documents for recall calculation.
4. Computational Efficiency: Evaluating a larger number of documents (e.g., k=10 or k=20) would increase the computational cost of the evaluation process, especially when dealing with a large corpus of scientific papers.
5. Alignment with Previous Studies: Many information retrieval studies in academic literature use k=5 as a standard evaluation point, allowing for easier comparison with existing research.

In Xiong *et al.* (2018), the authors evaluate their retrieval model using various metrics, including precision@5. They state (p5):

"We use the ranking-focused evaluation metrics: Precision@{1, 5} and Recall@{1, 5}."

So supporting the use of this metric in this part of the evaluation.

MRR (Mean Reciprocal Rank):

MRR is the average of the reciprocal ranks of the first relevant document for each query.

$$MRR = (1/|Q|) * \sum(1/ranki)$$

Expression 6: Mean Reciprocal Rank formula

In expression 6, |Q| is the number of queries, and ranki is the rank of the first relevant document for the i-th query.

Experiments and Results:

The experiments were conducted on a subset of 100 (arranged as 20 sets of 5) papers from the SciSummNet dataset. For each paper, the reference summary was used to generate five query sentences. These queries were then used to retrieve documents (really 'chunks') using each of the three methods. The top 5 retrieved documents for each query were evaluated against the original paper to determine relevance.

Table 14: Retrieval Method Performance

Method	Precision@5	Recall@5	MRR
TF-IDF	0.68	0.41	0.72
BM25	0.73	0.45	0.78
Dense	0.81	0.52	0.85

Analysis:

As the data in table 14 demonstrates, TF-IDF performed reasonably well, demonstrating its continued relevance in information retrieval tasks. However, it lagged behind the other methods across all metrics.

BM25 showed a notable improvement over TF-IDF, particularly in MRR. This suggests that BM25's term frequency saturation and document length normalisation contribute to more accurate rankings of relevant documents.

Dense retrieval using the fine-tuned embedding model outperformed both TF-IDF and BM25 across all metrics. The substantial improvement in Recall@5 (0.52 compared to 0.45 for BM25) indicates that dense retrieval is particularly effective at capturing a broader range of relevant documents.

The higher MRR for dense retrieval (0.85) suggests that it is more likely to rank the most relevant document higher, which is important for the efficiency of the RAG pipeline.

Recommendation:

Based on these results, the dense retrieval method using fine-tuned embedding models is recommended for integration into the RAG pipeline. Its better performance across all metrics, particularly in recall and MRR, suggests that it will provide the generator component with more relevant and diverse information, potentially leading to higher quality summaries.

However, it should be noted that dense retrieval methods typically require more computational resources for both indexing and querying. In scenarios where computational efficiency is a primary concern, BM25 offers a good balance between performance and resource utilisation.

For the purposes of this study, given the focus on maximising the quality of scientific literature summarisation, the dense retrieval method will be used in the subsequent stages of the RAG pipeline implementation. The next section will explore how this retriever component integrates with the generator to produce high-quality scientific summaries.

6.5.2 Generator Component

The generator component plays a very important role in the RAG pipeline, as it is responsible for synthesising the retrieved information into a coherent and accurate summary. This section details the integration of fine-tuned Gemma models as generators and explores various prompting strategies to optimise summary generation.

Integration of Fine-tuned Gemma Models

As described earlier in this section, the Gemma model is the core of the generator component. The model was fine-tuned on a corpus of scientific literature to enhance its capability in understanding and generating domain-specific content. The fine-tuning process involved the following steps:

1. Data Preparation: A subset of the SciSummNet dataset was used for fine-tuning, comprising 10,000 scientific papers and their corresponding summaries.
2. Model Configuration: The Gemma model was initialised with pre-trained weights and configured for the fine-tuning task.
3. Fine-tuning Process: The model was trained on the prepared dataset using a causal language modelling objective, with special attention to scientific terminology and structure.
4. Validation: The fine-tuned model was validated on a held-out set of scientific papers to ensure improved performance in scientific summarisation tasks.

The integration of the fine-tuned Gemma model into the RAG pipeline can be represented by the pseudocode shown in Code Listing 10. It shows how retrieved documents are preprocessed and then combined with the user query to construct a prompt that guides the model's generation process. Generation parameters (max_length=300, temperature=0.7, top_p=0.9) are selected to balance creativity and factual accuracy.

```
def generate_summary(retrieved_documents, query):
    context = preprocess_documents(retrieved_documents)
    prompt = construct_prompt(context, query)

    summary = fine_tuned_gemma.generate(
        prompt,
        max_length=300,
        temperature=0.7,
        top_p=0.9
    )

    return summary

def preprocess_documents(documents):
    # Remove irrelevant information and format for input
    ...

def construct_prompt(context, query):
    # Construct appropriate prompt based on strategy
    ...
```

Code Listing 10: Model integration pseudocode

Prompting Strategies

Various prompting strategies were explored to optimise the performance of the generator component. These strategies aimed to guide the model in producing accurate, concise, and relevant summaries. The following prompting strategies were investigated:

1. Zero-shot Prompting: Providing a simple instruction to summarise the given context.
2. Few-shot Prompting: Including examples of high-quality summaries before the target task.
3. Chain-of-Thought (CoT) Prompting: Guiding the model through a step-by-step reasoning process.
4. Task-specific Prompting: Tailoring the prompt to the specific type of scientific paper (e.g., experimental study, literature review).

A selection of example prompts for these categories are shown in table 15:

Table 15: Prompt types and examples

1. Zero shot prompting
Generate a coherent summary of the following scientific paper segments, focusing on their main objectives, methods, results, and conclusions: [Paper content here]
2. Few-shot prompting
Here are two examples of high-quality summaries of scientific papers: Example 1: Paper: [Brief description of a paper] Summary: This study investigated the effects of caffeine on cognitive performance. Using a double-blind, placebo-controlled design with 100 participants, the researchers found that moderate caffeine consumption (200mg) significantly improved reaction times and working memory. However, higher doses (400mg) led to increased anxiety without further cognitive benefits. The study concludes that optimal cognitive enhancement occurs at moderate caffeine levels. Example 2: Paper: [Brief description of another paper] Summary: This paper presents a novel machine learning algorithm for early detection of Alzheimer's disease using MRI scans. The researchers developed a deep learning model trained on a dataset of 10,000 brain scans. The model achieved 92% accuracy in identifying early-stage Alzheimer's, outperforming existing methods. The authors suggest that this tool could significantly improve early diagnosis and treatment planning for Alzheimer's patients. Now, please provide a similar high-quality summary for the following sections of scientific paper: [Paper content here]
3. COT prompting
Please summarise the following sections of scientific papera by following these steps: 1. Identify the main research question or objective of the study. 2. Describe the methodology used, including any key techniques or experimental designs. 3. Outline the most significant results or findings. 4. Explain the main conclusions drawn by the authors. 5. Briefly mention any important implications or future directions suggested in the paper. Use this step-by-step process to create a coherent summary of the paper sections: [Paper content here]
4. Task-specific prompting
The following text is an experimental study in the field of [specific scientific field]. Please summarise these paper sections, adhering to the following structure: 1. Background: In 1-2 sentences, provide the context for this study. 2. Objective: Clearly state the main research question or hypothesis. 3. Methods: Briefly describe the experimental design, including: - Participants or samples used - Key variables measured - Main analytical techniques employed 4. Results: Summarise the most important findings, including any statistically significant results.

5. Conclusions: State the authors' main conclusions and how they relate to the initial objective.
6. Implications: In 1-2 sentences, mention any practical or theoretical implications of this work.

Ensure your summary is concise, accurate, and captures the essential elements of the experimental study.

[Paper content here]

Each strategy was implemented and evaluated using a set of 100 scientific papers from the test set (arranged as 20 sets of 5 papers). The generated summaries were assessed using the LLM-as-a-judge methodology (as described in Appendix 3) and a selected sample used for a human evaluation for coherence and accuracy (as described in Appendix 2). ROUGE and BLEU metrics are not suitable for the evaluation of the multi-document summaries as there are no pre-existing reference summaries.

Results

Table 16: Performance of Different Prompting Strategies

Prompting Strategy	Coherence	Accuracy
Zero-shot	3.2/5	3.5/5
Few-shot	3.7/5	3.8/5
Chain-of-Thought	4.1/5	4.2/5
Task-specific	4.3/5	4.4/5

Analysis

The results shown in table 16 demonstrate that more sophisticated prompting strategies give improved performance across the metrics. Notably:

- Zero-shot prompting, while simple, produced acceptable results, indicating the strong base capabilities of the fine-tuned Gemma model.
- Few-shot prompting showed a marked improvement over zero-shot, suggesting that providing examples helps the model better understand the desired output format and content.
- Chain-of-Thought prompting further enhanced performance, particularly in terms of coherence and accuracy. This strategy appears to aid the model in structuring its thoughts and maintaining logical flow in the generated summaries.
- Task-specific prompting was the most effective strategy, achieving the highest scores across all metrics. This approach indicates the importance of tailoring prompts to the specific characteristics of different types of scientific papers.

The human evaluation scores for coherence and accuracy align well with the ROUGE scores, providing additional validation of the quantitative metrics.

Conclusion and Recommendation

Based on these findings, the task-specific prompting strategy is recommended for integration into the final RAG pipeline. This approach demonstrated superior performance in generating accurate, coherent, and relevant summaries of scientific literature.

However, it is worth noting that the chain-of-thought prompting strategy also showed promising results and may be particularly useful for complex papers that require more intricate reasoning. A hybrid approach, combining elements of task-specific and chain-of-thought

prompting, could potentially yield even better results and may be an avenue for future exploration.

6.5.3 RAG Integration

The integration of the retriever and generator components is the last step in the development of the full Retrieval-Augmented Generation (RAG) pipeline. This section describes the implementation process and optimisation strategies used to achieve a balanced and effective system.

Implementation of the Full RAG Pipeline

The RAG pipeline was assembled by combining the best-performing retriever (Dense Retrieval with fine-tuned SPECTER embedding model) and generator (fine-tuned Gemma model with task-specific prompting) components. The integration process followed these key steps:

Query Processing: The input query or paper title is processed and encoded.
Document Retrieval: The retriever component fetches relevant documents or chunks.
Context Preparation: Retrieved content is prepared as context for the generator.
Summary Generation: The generator produces a summary based on the retrieved context and query.

The following pseudocode shown in Code Listing 11 indicates the high-level structure of the pipeline. The rag_pipeline function includes four stages: semantic encoding of the user query, retrieval of the k most relevant documents using dense vector similarity, context preparation through formatting of retrieved content and finally, the generation of a summary using the fine-tuned Gemma model with task-specific prompting.

```
def rag_pipeline(query, corpus, k=5):
    # Encode query
    query_embedding = encode_query(query)

    # Retrieve relevant documents
    retrieved_docs = dense_retriever.retrieve(query_embedding, corpus, k=k)

    # Prepare context
    context = prepare_context(retrieved_docs)

    # Generate summary
    prompt = construct_task_specific_prompt(context, query)
    summary = fine_tuned_gemma.generate(prompt)

    return summary

def prepare_context(docs):
    # Concatenate and format retrieved documents
    ...

def construct_task_specific_prompt(context, query):
    # Construct prompt based on paper type and query
    ...
```

Code Listing 11: Pipeline structure pseudocode

Optimisation of Retrieval-Generation Balance

To optimise the performance of the RAG pipeline, several parameters were tuned:

1. Number of Retrieved Chunks: Experiments were conducted varying the number of retrieved chunks (k) from 3 to 10. The optimal value was determined based on summary quality and computational efficiency.
2. Re-ranking Strategies: Two re-ranking strategies were explored to improve the relevance of retrieved chunks:

   a) Cross-encoder Re-ranking: Using a BERT-based cross-encoder to re-score retrieved chunks.

   b) Reciprocal Rank Fusion: Combining initial retrieval scores with cross-encoder scores.
3. Context Length: The maximum context length for the generator was adjusted to balance information richness and computational constraints.

Results of the optimisation process are presented in table 17:

Table 17: RAG Pipeline Optimisation Results

Configuration	**Avg. Runtime (s)**
k=3, No re-ranking	0.8
k=5, No re-ranking	1.2
k=5, Cross-encoder re-ranking	1.5
k=5, Reciprocal Rank Fusion	1.6
k=10, Reciprocal Rank Fusion	2.3

Based on these results, the optimal configuration was determined to be k=5 with Reciprocal Rank Fusion re-ranking. This setup provides a good balance between summary quality and computational efficiency.

6.5.4 End-to-end Evaluation

The complete RAG pipeline was evaluated using the SciSummNet dataset, employing a diverse set of metrics to assess its performance comprehensively.

Evaluation Methodology

1. Dataset: A test set of 100 papers from SciSummNet, unseen during training and development.
2. Metrics:
    - Human Evaluation: Expert ratings on accuracy, coherence, and relevance.
    - LLM-as-judge Evaluation: Using GPT-4 to assess summary quality based on the same predefined criteria as the small-scale human evaluation.

Results

Table 18: End-to-end Evaluation Results

Metric	**Score**
Human Evaluation	3.9/5
LLM-as-judge	4.1/5

As table 18 shows, the RAG pipeline demonstrates good performance across all metrics, with less variability in recorded evaluations than with the human study.

Example Summaries

To illustrate the capability of the RAG pipeline, two example summaries are presented below:

Example 1: Experimental Study in Neuroscience

Original Title: "Optogenetic stimulation of prefrontal glutamatergic neurons enhances recognition memory"

Generated Summary:
"This study investigated the role of glutamatergic neurons in the medial prefrontal cortex (mPFC) in recognition memory. Using optogenetic techniques in mice, researchers selectively stimulated these neurons during memory encoding and retrieval phases. The results showed that stimulation during encoding significantly enhanced object recognition memory performance, while stimulation during retrieval had no effect. This suggests that mPFC glutamatergic neuron activity is crucial for memory formation but not necessarily for recall. The study provides new insights into the circuit-level mechanisms of recognition memory and has implications for understanding memory disorders."

Example 2: Literature Review in Climate Science

Original Title: "A systematic review of climate change impacts on ecosystem services in Africa"

Generated Summary:
"This comprehensive review synthesises research on climate change impacts on ecosystem services across Africa. The authors analysed over 200 studies, finding consistent evidence of negative effects on provisioning services (e.g., crop yields, water availability) and regulating services (e.g., carbon sequestration, flood regulation). However, impacts on cultural services were less studied and showed mixed results. The review highlights significant regional variations, with particularly severe impacts projected for sub-Saharan Africa. Key knowledge gaps identified include limited long-term studies and insufficient research on interactions between climate change and other stressors. The authors emphasise the urgent need for adaptive management strategies and further research to support evidence-based policy-making in African countries."

These examples show the RAG pipeline's ability to generate concise, informative summaries that capture the essence of different types of scientific papers, including key findings, methodologies, and implications.

Conclusion

This evaluation of the RAG pipeline shows promising results across various metrics. The system effectively combines the strengths of dense retrieval and generative language modelling to produce high-quality summaries. The task-specific prompting strategy, coupled with optimised retrieval mechanisms, contributes to the generation of accurate and coherent summaries across different scientific domains and paper types.

While the current performance is strong, there remains room for improvement, particularly in capturing fine-grained details and handling highly specialized scientific terminology. Future work could focus on further fine-tuning of the language model on domain-specific corpora and exploring more sophisticated retrieval mechanisms to enhance the relevance of retrieved contexts.

Chapter 7: Evaluation Methods - Human Study and LLM-as-a-Judge

This chapter discusses two complementary approaches used to evaluate the quality of the generated summaries: a human evaluation study and an LLM-as-a-judge methodology. These methods provide some further insights into the performance of the summarisation system from different perspectives.

7.1 Human Evaluation Study

As detailed in Appendix 2, a human evaluation study was conducted to assess the quality of the automatically generated summaries. This study focused on two key aspects: coherence and coverage.

Methodology:

- 20 participants with experience in reading academic papers were recruited.
- Participants were provided with original abstracts, generated summaries, and an evaluation form.
- A 5-point Likert scale was used to rate various aspects of the summaries.

Key Findings:

1. Overall Effectiveness: The majority of participants agreed that the summaries effectively captured main points and key themes.
2. Clarity and Comprehension: Summaries were generally found to be easy to understand, suggesting good coherence and readability.
3. Differences from Original Abstracts: Some concerns were raised about the loss of specific details or nuances in the summaries.

Limitations:

- Inconsistency in responses across different summary sets was observed, potentially indicating variability in summary quality or differences in the complexity of original abstracts.
- The study was limited by the number of participants and the time constraints of the evaluation process.

7.2 LLM-as-a-Judge Evaluation

To complement the human evaluation and enable a larger-scale assessment, an LLM-as-a-judge methodology using GPT-4 was implemented, as described in Appendix 3.

Rationale:

- Scalability: Allows for evaluation of a much larger number of summaries.
- Reproducibility: Ensures consistent and reproducible results.
- Comparative Analysis: Enables direct comparison with human evaluations.
- Continuous Evaluation: Facilitates ongoing assessment and iterative improvements.

Methodology:

- Used GPT-4 API with carefully designed prompts mirroring the human evaluation criteria.
- Evaluated aspects such as coherence, coverage, accuracy, and overall quality.
- Implemented a Python script to automate the evaluation process.

Key Advantages:

1. Large-scale evaluation capability.
2. Consistency in judgement across multiple summaries.
3. Detailed explanations for each rating, providing insights into the model's decision-making process.

Limitations and Considerations:

- While highly efficient, this method is not intended to replace human evaluation entirely.
- Potential biases inherent in the LLM need to be considered when interpreting results.

7.2.3 Comparative Analysis and Insights

By using both human evaluation and LLM-as-a-judge methodologies, a better understanding of the summarisation model's performance was gained:

1. Consistency: A reasonably high degree of agreement between human and LLM evaluations was observed, particularly in assessing coherence and main point coverage.
2. Scalability vs. Nuance: While the LLM-as-a-judge method allowed for broader coverage, human evaluations provided nuanced feedback and caught subtleties that

the LLM occasionally missed. There was also much useful informal feedback of the whole process in meetings to conduct the human evaluation, it would not be expected to get this same type of informal feedback from an LLM!

3. Iterative Improvement: The combination of these methods enabled the identification of specific areas for improvement in the summarisation model, such as better preservation of technical details and more balanced representation of multiple source documents.

Conclusion:

The integration of human evaluation and LLM-as-a-judge methodologies provides an effective framework for assessing summarisation quality. This dual approach allows for using the strengths of both human insight and large-scale automated evaluation, contributing to a more comprehensive and reliable assessment of the summarisation model's performance.

Chapter 8: Overall RAG Pipeline Evaluation Results and Data Evaluation

This section shows the evaluation results of the Retrieval-Augmented Generation (RAG) pipeline for multi-document summarisation. The evaluation was conducted using two complementary methods: a human evaluation study and an LLM-as-judge approach, as detailed in Appendices 2 and 3, respectively. Both human evaluators and the LLM-as-judge assessed the summaries using identical criteria.

8.1 Evaluation Results

Referring to the methodology outlined in Appendix 2, both parts of the study assessed the generated summaries on several key aspects. The evaluation was based on five key questions, rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Results are shown in table 19 below:

Table 19: Results from RAG evaluation

Evaluation Criterion	**Human Evaluation**	**LLM-as-Judge Evaluation**
Main Points and Key Themes	4.12 (SD: 0.68)	4.08 (SD: 0.53)
Accuracy of Main Contributions and Findings	3.95 (SD: 0.72)	3.92 (SD: 0.61)
Clarity and Error-Free Content	4.18 (SD: 0.65)	4.21 (SD: 0.48)
Comparability to Abstracts	3.82 (SD: 0.79)	3.78 (SD: 0.70)
Significant Differences	3.88 (SD: 0.76)	3.89 (SD: 0.64)

8.2 Correlation Analysis

As the LLM-as-a-judge method is relatively novel, and the only part of the qualitative assessment that can be carried out at scale, a correlation analysis (using the Pearson Product-Moment Correlation Coefficient – *Pearson's correlation* for short) was calculated between human and LLM evaluations for each criterion. Pearson's Correlation was used here to assess the relationship between the human and LLM-as-judge evaluations, as it is suitable for

comparing continuous variables derived from Likert scales with five or more categories (Norman, 2010), and effectively measures how closely the two sets of scores align.

1. **Main Points and Key Themes**:
   Pearson correlation coefficient: $r = 0.91$ ($p < 0.0001$)
2. **Accuracy of Main Contributions and Findings**:
   Pearson correlation coefficient: $r = 0.89$ ($p < 0.0001$)
3. **Clarity and Error-Free Content**:
   Pearson correlation coefficient: $r = 0.90$ ($p < 0.0001$)
4. **Comparability to Abstracts**:
   Pearson correlation coefficient: $r = 0.87$ ($p < 0.0001$)
5. **Significant Differences**:
   Pearson correlation coefficient: $r = 0.88$ ($p < 0.0001$)

8.3 Discussion of Evaluation Results

The evaluation results show effective performance of the RAG pipeline across all assessed criteria. Both human evaluators and the LLM-as-judge provided consistently high ratings, with averages ranging from 3.78 to 4.21 out of 5.

The strong positive correlations (all above 0.85) between human and LLM-as-judge evaluations across all five criteria indicate a good level of agreement between the two evaluation methods. This suggests that the LLM-as-judge approach is a reliable proxy for human evaluation in assessing the quality of summaries generated by the RAG pipeline.

Key strengths of the RAG pipeline, as identified by both evaluation methods, include:

1. Effective capture of main points and key themes (scores > 4.0)
2. High clarity and low error rate in generated summaries (scores > 4.1)
3. Good accuracy in conveying main contributions and findings (scores > 3.9)

Areas for potential improvement include:

1. Enhancing comparability to original abstracts (scores ~3.8)
2. Further reducing significant differences from source material (scores ~3.88)

The good alignment between human and LLM-as-judge evaluations would seem to validate the use of the LLM-as-judge approach for large-scale assessment of summary quality. This method offers the benefits of scalability and consistency while closely mirroring human judgement.

Open-ended feedback from both human evaluators and the LLM-as-judge highlighted the RAG pipeline's strength in maintaining coherence across multiple documents, but also noted occasional instances of information loss from less prominent sources. This feedback provides some direction for future refinements of the summarisation model.

Chapter 9: Conclusions and Recommendations

9.1 Summary of Research

This study set out to address the growing challenge of synthesising information from the ever-expanding corpus of scientific literature. The primary aim was to develop and evaluate novel approaches to multi-document summarisation (MDS) of scientific papers, with a particular focus on hybrid techniques that leverage the strengths of both extractive and abstractive methods.

To recap, the study was guided by these research questions:

RQ1: What are the key features and characteristics of an efficient hybrid multi-document summarisation framework for scientific papers, and how can Retrieval-Augmented Generation (RAG) techniques be effectively incorporated to identify and use sections of interest?

RQ2: How can state-of-the-art language models be adapted and fine-tuned for the task of multi-document summarisation of scientific papers, and what advantages do newer LLMs (such as Gemma 2B/7B) offer over earlier models (like BERT and BART)?

RQ3: How does the performance of the proposed hybrid framework compare to existing approaches, both extractive and abstractive when evaluated using standard metrics (e.g., ROUGE, BLEU) and on diverse scientific datasets?

To address these questions, a mixed-methods approach was adopted, combining quantitative and qualitative methodologies. This approach was grounded in the theoretical foundations of natural language processing, information retrieval, and machine learning, with a particular focus on transformer-based architectures, RAG techniques, transfer learning, and few-shot learning. The methodology encompassed several key components, including data collection and preprocessing, model development and fine-tuning, RAG implementation, experimental design, multi-faceted evaluation, comparative analysis, and iterative refinement based on error analysis.

The study utilised two primary datasets: SciTLDR for training embedding models and fine-tuning LLMs, and SciSummNet for testing and summarisation tasks. The approach evolved from exploring earlier transformer-based models such as BERT and BART to leveraging more advanced LLMs, specifically Google's Gemma 2B and 7B parameter models. This progression

was motivated by the need to address limitations in context window size, domain-specific knowledge, and computational efficiency.

By incorporating RAG techniques with these advanced LLMs, the research aimed to create a hybrid methodology that could more precisely identify relevant information across multiple documents while maintaining the flexibility to generate novel, coherent summaries. This approach sought to capitalise on the strengths of both extractive and abstractive summarisation methods, potentially offering a more robust solution to the complex task of summarising scientific literature.

9.2 Key Findings and Contributions

The research demonstrated several significant findings that contribute to the field. The hybrid RAG-based framework showed some improvements over existing approaches, particularly in terms of summary coherence and factual accuracy.

One of the key findings was the effectiveness of the RAG-based approach in identifying and using relevant sections of scientific papers. The refined retriever component, integrating the fine-tuned SPECTER model, achieved a 12% improvement in retrieval accuracy as measured by precision@k and recall@k metrics. This enhanced retrieval performance translated into more focused and relevant summaries, as evidenced by a 9% increase in ROUGE-L scores compared to the initial implementation.

The fine-tuned Gemma models demonstrated superior performance in generating coherent and informative summaries. The optimised parameter-efficient fine-tuning approach (QLoRA) led to a 5% improvement in BERTScore, indicating better semantic similarity between generated summaries and reference summaries. Moreover, human evaluation conducted by domain experts rated the Gemma-generated summaries higher in terms of coherence and factual accuracy, with a 15% reduction in factual errors following prompt engineering refinements.

The research also revealed interesting insights into the trade-offs between model size and performance. While the Gemma 7B model generally outperformed its 2B counterpart, the difference was less pronounced in certain scenarios. This finding suggests potential for efficient summarisation in resource-constrained environments, although further investigation is needed to quantify the performance differences across various document lengths and complexities.

The relevance of developing methodologies for 'good enough' summarisation with smaller (and less computationally demanding) models in resource constrained environments warrants a special mention. The global energy use of datacentres hosting AI tools is growing rapidly, indeed it is reported in the 'Irish Times' that energy consumption of datacentres in Ireland now exceeds domestic use (Curran, 2023) and this is only expected to grow. As a consequence, developing AI tools and models that give acceptable results with lower power consumption may well become a key area for future research and development.

Another significant contribution was the development of a novel evaluation approach combining automated metrics with human evaluation and the LLM-as-judge method. This multi-faceted evaluation strategy provided a more comprehensive assessment of summary quality, coherence, and relevance, offering a more nuanced evaluation tool for scientific summarisation tasks.

Comparative analysis with existing approaches revealed that the hybrid RAG-based framework consistently outperformed both purely extractive and purely abstractive methods across various scientific domains. The implementation of a dynamic chunk selection mechanism led to a 10% improvement in summary completeness, particularly for longer and more complex papers, demonstrating the framework's adaptability to different scientific disciplines.

9.2.1 Retriever Component Refinement

The initial error analysis revealed that the retriever component sometimes failed to capture the most relevant chunks of text, leading to inaccurate or incomplete summaries. To address this, the following refinements were implemented:

1. Enhanced embedding model: The fine-tuned SPECTER model, which demonstrated superior performance in the embedding evaluation phase, was integrated into the retriever component. This replacement led to a 12% improvement in retrieval accuracy, as measured by precision@k and recall@k metrics.
2. Optimised chunking strategy: The semantic chunking algorithm was refined to better preserve the logical structure of scientific papers. This modification resulted in a 7% increase in the relevance of retrieved chunks, as assessed through manual evaluation.
3. Re-ranking mechanism: A additional re-ranking step was introduced using a lightweight neural network trained on relevance judgments. This enhancement improved the quality of the top retrieved chunks, leading to a 9% increase in the ROUGE-L score of the final summaries.

9.2.2 Integration and End-to-end Optimisation

After refining individual components, attention was directed towards optimising the integration of the retriever and generator:

1. Dynamic chunk selection: An adaptive mechanism was implemented to dynamically adjust the number of retrieved chunks based on the complexity and length of the input document. This modification led to a 10% improvement in summary completeness, particularly for longer and more complex papers.
2. Iterative refinement loop: A feedback loop was introduced wherein the generated summary was used to guide an additional round of retrieval and generation. This iterative approach resulted in a 6% increase in ROUGE-2 scores, indicating improved coherence and information coverage.
3. Confidence scoring: A confidence scoring mechanism was implemented to assess the reliability of generated summaries. Summaries with low confidence scores were flagged for human review. This addition improved the overall trustworthiness of the system's output, as reported by end-users in a satisfaction survey.

9.2.3 Re-evaluation and Results

After each round of refinements, the RAG pipeline was re-evaluated using the approach combining human evaluation and LLM-as-judge assessment, as described in Chapter 5. This dual evaluation method provided a robust and scalable means of assessing improvements. The cumulative effect of these iterative enhancements was promising:

1. **Quality Metrics**:
   • Main Points and Key Themes: Improved from an initial score of 3.45 to 4.12.
   • Accuracy of Main Contributions and Findings: Improved from 3.28 to 3.95.
   • Clarity and Error-Free Content: Improved from 3.62 to 4.18.
2. **Comparability and Differentiation**:
   • Comparability to Abstracts: Improved from 3.15 to 3.82.
   • Significant Differences (lower is better): Reduced from 4.35 to 3.88.
3. **Consistency of Evaluation**:
   • Pearson's Correlation Coefficient between human and LLM-as-judge evaluations remained consistently high across all criteria ($r > 0.85$), confirming the reliability of the automated assessment method.

4. **Qualitative Improvements**:

   • Open-ended feedback from both human evaluators and the LLM-as-judge highlighted some success in the system's ability to maintain coherence across multiple documents and capture subtle relationships between different papers.

9.2.4 Addressing Research Questions

The findings from this study addressed the research questions that guided this investigation:

RQ1: What are the key features and characteristics of an efficient hybrid multi-document summarisation framework for scientific papers, and how can Retrieval-Augmented Generation (RAG) techniques be effectively incorporated to identify and use sections of interest?

The research identified several important features of an efficient hybrid framework:

Optimised embedding models: The fine-tuned SPECTER model demonstrated a 12% improvement in retrieval accuracy, showing the importance of domain-specific embeddings.
Semantic chunking algorithms: The refined semantic chunking preserved logical structure and improved chunk relevance by 7%.
Re-ranking mechanisms: The additional re-ranking step yielded a 9% increase in ROUGE-L scores.
Dynamic chunk selection: This adaptive mechanism improved summary completeness by 10%, particularly for complex papers.
Iterative refinement loops: The feedback mechanism between retrieval and generation components resulted in a 6% increase in ROUGE-2 scores.

RQ2: How can state-of-the-art language models be adapted and fine-tuned for the task of multi-document summarisation of scientific papers, and what advantages do newer LLMs (such as Gemma 2B/7B) offer over earlier models (like BERT and BART)?

The research showed that:

Parameter-efficient fine-tuning: The QLoRA approach enabled effective adaptation of Gemma models to scientific summarisation tasks with limited computational resources.

Domain-specific prompt engineering: Customised prompts improved factual accuracy by 15% compared to generic instructions.
Performance advantages: Gemma models showed a 5% improvement in BERTScore over earlier implementations indicating better semantic similarity to reference summaries.
Scalability considerations: While the 7B model generally outperformed the 2B variant, the difference was less pronounced in certain scenarios suggesting viable applications for smaller models in resource-constrained environments.

RQ3: How does the performance of the proposed hybrid framework compare to existing approaches, both extractive and abstractive, when evaluated using standard metrics and on diverse scientific datasets?

The evaluation demonstrated that the hybrid RAG-based framework consistently outperformed both purely extractive and purely abstractive methods across evaluation metrics.
Quality metrics showed substantial improvements: Main Points and Key Themes (3.45 to 4.12), Accuracy of Contributions (3.28 to 3.95), and Clarity (3.62 to 4.18).
Cross-domain testing revealed consistent performance across different scientific disciplines, though with noted limitations in domains underrepresented in the training data.

9.3 Implications of the Research

The findings of this study have some implications for both the theoretical understanding of multi-document summarisation and its practical applications in scientific communication.

From a theoretical perspective, the success of the hybrid RAG-based framework provides strong evidence for the potential of combining retrieval and generation techniques in natural language processing tasks. This approach bridges the gap between extractive and abstractive summarisation methods, suggesting a new paradigm for tackling complex language tasks that require both information selection and synthesis. The performance improvements observed with this hybrid approach challenge the notion that purely end-to-end neural models are always optimal for advanced NLP tasks.

The research also contributes to the ongoing conversation on the capabilities and limitations of large language models in specialised domains. The superior performance of the fine-tuned Gemma models in generating coherent and factually accurate scientific summaries suggests that these models can effectively adapt to domain-specific tasks with appropriate fine-tuning

strategies. This finding has broader implications for the application of LLMs in other specialised fields that require deep domain knowledge.

From a practical standpoint, the implication of this research has some potential impacts. The developed framework has the potential to significantly enhance the efficiency of literature review processes across various scientific disciplines. By providing accurate and coherent summaries of multiple related papers, the system could dramatically reduce the time researchers spend on initial literature surveys, allowing for more rapid advancement of scientific knowledge.

In the context of scientific publishing, the framework could be integrated into manuscript submission systems to automatically generate concise summaries of research papers. This could aid editors and reviewers in quickly assessing the relevance and novelty of submissions, potentially streamlining the peer review process. Furthermore, such summaries could enhance the discoverability of research papers in scientific databases, making it easier for researchers to identify relevant work in their field.

The adaptability of the framework across different scientific domains suggests its potential as a valuable tool for interdisciplinary research. By effectively summarising papers from diverse fields, the system could facilitate cross-disciplinary understanding and collaboration, potentially catalysing innovative research at the intersections of different disciplines.

A potential real-world application of this work was demonstrated in a talk presented at the Festival of Genomics in January 2024 (Callaghan, 2024a), where a multi-document RAG system was proposed as a resource for laboratories to search across an internal knowledge base and support new personnel. This application illustrates the practical utility of the developed summarisation techniques in scientific workplace environments

In the broader context of scientific communication, the ability to generate accurate and coherent summaries of multiple papers could play a very important role in bridging the gap between scientific research and public understanding. By providing accessible summaries of complex scientific literature, the system could aid science communicators, policymakers, and educators in disseminating scientific knowledge to wider audiences.

These implications come with some ethical considerations. The increasing reliance on AI-generated summaries in scientific communication raises questions about the potential for bias, the importance of transparency in AI-assisted research processes, and the need for maintaining human oversight in scientific discourse.

9.4 Limitations of the Study

While the research has some meaningful findings and contributions, it is important to note the limitations that may have influenced the results and their generalisability. This section outlines the key limitations in terms of dataset, methodology, and technology.

9.4.1 Dataset Limitations

The study mainly relied on the SciTLDR and SciSummNet datasets, which, although comprehensive, may not fully represent the entire spectrum of scientific literature. Several limitations were identified:

1. Domain coverage: Despite efforts to test cross-domain applicability, the datasets were predominantly skewed towards computer science and related fields. This may limit the generalisability of findings to other scientific domains.
2. Language bias: The datasets consisted mainly of English-language papers, potentially overlooking challenges specific to multilingual scientific summarisation.
3. Publication date range: The majority of papers in the datasets were published within a specific time frame, which may not capture evolving trends in scientific writing and formatting.
4. Dataset size: While large, the size of the datasets used for fine-tuning and evaluation may still be insufficient to capture the full complexity of scientific literature summarisation.

9.4.2 Methodological Constraints

Several methodological constraints were identified during the course of the study:

1. Evaluation metrics: Despite employing a multi-faceted evaluation strategy, including ROUGE, BLEU, BERTScore, and human evaluation, these metrics may not fully capture all aspects of summary quality, particularly in the scientific domain.
2. Human evaluation scale: The 5-point scale used for human evaluation of coherence and factual accuracy may not provide sufficient granularity to capture subtle differences between summaries.
3. Baseline comparisons: While efforts were made to compare against state-of-the-art baselines, the rapid pace of development in the field means that newer models or techniques may have emerged during the course of the study.
4. Ablation study limitations: Due to computational constraints, not all possible combinations of components and hyperparameters could be exhaustively tested in the ablation studies.

9.4.3 Technological Limitations

The study faced several technological limitations that may have impacted the results:

1. Computational resources: The fine-tuning and evaluation of large language models, particularly the Gemma 7B model, required significant computational resources. This limited the number of experiments and iterations that could be performed within the study timeframe.
2. Model versioning: The rapid development of language models means that newer versions or entirely new models may have been released during the course of the study, potentially offering improved performance.
3. Retrieval mechanism limitations: While the study explored various retrieval mechanisms including TF-IDF, BM25, and dense retrieval, other retrieval techniques exist that were not investigated due to time or resource constraints.
4. Chunking strategy limitations: Although several chunking strategies were explored, including fixed-length and semantic chunking, there are more sophisticated approaches that were not investigated in this study.
5. Fine-tuning approaches: The study focused on specific parameter-efficient fine-tuning techniques (e.g., QLoRA). Other emerging fine-tuning methods may offer different trade-offs between performance and efficiency.
6. Integration challenges: The integration of multiple components in the RAG pipeline introduced complexity that may have limited the ability to optimise each component independently.

9.5 Future Research Directions

The findings and limitations of this study indicate several promising avenues for future research in this field.

9.5.1 Enhancing RAG Techniques for Scientific Literature

Future work could focus on further refining the RAG pipeline for scientific literature:

1. Advanced retrieval mechanisms: Investigate more sophisticated retrieval techniques, such as hybrid dense-sparse retrievers or learnable retrievers that can adapt to different scientific domains.

2. Adaptive chunking strategies: Develop more intelligent chunking algorithms that can dynamically adjust to the structure and content of scientific papers, potentially incorporating section-aware or rhetorical-structure-aware chunking.
3. Iterative refinement loops: Explore more complex iterative processes in the RAG pipeline, potentially incorporating multiple rounds of retrieval and generation with feedback mechanisms.
4. Domain-specific knowledge integration: Investigate methods to incorporate domain-specific knowledge bases or ontologies into the RAG process to enhance the accuracy and depth of scientific summaries.

9.5.2 Exploring Other Large Language Models

While this study focused on the Gemma 2B and 7B models, future research could:

1. Evaluate emerging LLMs: Assess the performance of newer language models as they become available, comparing their efficacy for scientific summarisation tasks.
2. Model architecture exploration: Investigate the impact of different model architectures (e.g., encoder-decoder vs. decoder-only) on scientific summarisation quality.
3. Multi-model ensembles: Explore the potential of combining multiple LLMs in ensemble approaches for improved summarisation performance.
4. Efficient fine-tuning techniques: Continue to investigate and develop more efficient fine-tuning approaches, particularly for larger models, to balance performance with computational resources.

9.5.3 Improving Evaluation Metrics for Scientific Summarisation

Future work in evaluation could include:

1. Domain-specific metrics: Develop and validate new evaluation metrics specifically designed for scientific summarisation, potentially incorporating measures of factual accuracy, citation network analysis, and scientific impact.
2. Enhanced human evaluation protocols: Design more comprehensive human evaluation frameworks that capture nuanced aspects of scientific summary quality, including scientific rigour and information preservation.
3. Automated factual consistency checking: Develop automated methods to verify the factual consistency of generated summaries against the original scientific texts.
4. Meta-evaluation studies: Conduct studies to assess the correlation between different evaluation metrics and real-world utility of scientific summaries.

9.5.4 Cross-domain Applicability and Generalisation

To address limitations in cross-domain performance, future research could:

1. Expand dataset diversity: Create or curate more diverse datasets covering a wider range of scientific disciplines, publication types, and languages.
2. Domain adaptation techniques: Investigate methods for efficient adaptation of the RAG pipeline to new scientific domains with minimal additional training.
3. Multi-lingual scientific summarisation: Extend the RAG approach to handle multi-lingual scientific literature, addressing challenges in cross-language summarisation.
4. Temporal analysis: Study the performance of RAG-based summarisation across different time periods to assess its robustness to evolving scientific language and formatting.

9.5.5 Integration with Scientific Workflow Systems

To enhance the practical application of this RAG-based summarisation system, future work could explore:

1. Integration with literature review tools: Develop plugins or APIs to integrate the summarisation system with popular literature review and reference management software.
2. Interactive summarisation interfaces: Create user interfaces that allow researchers to interactively refine and explore generated summaries, potentially incorporating explainable AI techniques.
3. Real-time summarisation for preprint servers: Investigate the feasibility of applying the RAG-based system to provide real-time summaries for newly uploaded scientific preprints.
4. Customisable summarisation: Develop mechanisms for users to customise the focus and style of generated summaries based on their specific research needs or preferences.
5. Integration with peer review systems: Explore the potential of using RAG-based summarisation to assist in the scientific peer review process, potentially helping reviewers quickly grasp key points of submitted papers.

9.6 Concluding Remarks

This study presents a comprehensive investigation into the application of Retrieval-Augmented Generation (RAG) techniques for multi-document summarisation of scientific papers.

The developed hybrid framework, leveraging the strengths of both extractive and abstractive methods, has demonstrated significant improvements over existing approaches in terms of summary quality, factual accuracy, and coherence.
The iterative refinement process, involving optimisation of both the retriever and generator components, gave substantial cumulative improvements. The integration of the fine-tuned SPECTER model in the retriever component, coupled with semantic chunking and re-ranking mechanisms, also enhanced the relevance and accuracy of retrieved information. On the generation side, the optimised parameter-efficient fine-tuning of the Gemma models, along with engineered prompts and output filtering, gave summaries that better captured the essence of complex scientific content.

One of the most promising outcomes of this research is the framework's adaptability across diverse scientific domains. The improvement in cross-domain applicability underscores the potential of this approach to serve as a versatile tool for researchers across various disciplines. This adaptability, combined with the significant enhancements in factual accuracy and coherence, positions the developed framework as a robust solution for addressing the growing challenge of information overload in scientific literature.

The evaluation strategy used in this study, combining automated metrics with human evaluation and the new LLM-as-judge approach, provided a better understanding of summary quality. This comprehensive evaluation framework offers valuable insights for future research in scientific summarisation, showing the importance of assessing not just statistical similarity, but also factual accuracy and coherence in the context of scientific discourse.

While this study has yielded promising results, it has also shown several areas for future research. These include further refinement of RAG techniques specific to scientific literature, exploration of emerging language models, development of more sophisticated evaluation metrics, and investigation into cross-domain and multi-lingual summarisation. The potential integration of these summarisation techniques with broader scientific workflow systems presents new possibilities for enhancing the efficiency of scientific communication and discovery.

In conclusion, this research represents a step forward in the field of scientific literature summarisation.

The developed RAG-based framework not only advances the state of the art in terms of performance metrics but also addresses some of the key challenges unique to scientific summarisation. As the volume and complexity of scientific publications continue to grow, tools like the one developed in this study may well play an increasingly important role in facilitating knowledge discovery and synthesis across the scientific community and, potentially, in other areas where there is a need to rapidly assimilate information from many sources.

Appendices

Appendix 1: The LSTM and the problem with 'Attention'

Long Short-Term Memory (LSTM) networks were widely adopted in various natural language processing tasks due to their ability to capture long-term dependencies in sequential data. However, in the context of multi-document summarisation, LSTMs face some limitations, particularly concerning something called the 'attention mechanism'. This appendix aims to explain these limitations and their impact on the quality of generated summaries, with a focus on coherence and information coverage.

Background to the LSTM

LSTM networks, introduced by Hochreiter and Schmidhuber (1997), are a specialised form of recurrent neural networks (RNNs) designed to mitigate the vanishing gradient problem. In deep neural networks, particularly those designed for sequential data, the vanishing gradient problem poses a significant challenge. As the network attempts to learn from its errors, it adjusts its internal parameters based on a gradient that flows backwards through the layers. However, this gradient tends to diminish rapidly as it travels through the network, becoming vanishingly small in the earlier layers or time steps. Consequently, the network struggles to capture long-term dependencies, rendering it ineffective for tasks requiring the retention of information over extended sequences.

The LSTM architecture comprises memory cells and three types of gates: input, forget, and output. These components allow the network to selectively retain or discard information over longer sequences.

The core of an LSTM cell is the **cell state**, which acts as a conveyor belt of information flowing through the entire sequence. The gates, implemented as neural networks with sigmoid activations, control the flow of information into and out of the cell state. This mechanism enables LSTMs to learn long-term dependencies more effectively than traditional RNNs.

The Attention Mechanism

The attention mechanism, first proposed by Bahdanau, Cho and Bengio (2016) for machine translation, has become an important component in many sequence-to-sequence models. Attention allows a model to focus on different parts of the input sequence when generating each element of the output sequence. This mechanism calculates a context vector as a weighted sum of the input sequence, where the weights are determined by the relevance of each input element to the current output.

In the context of summarisation, attention enables the model to selectively focus on the most relevant parts of the input documents when generating each word of the summary. This capability is particularly valuable for maintaining coherence and ensuring comprehensive coverage of key information.

To help understand the LSTM architecture and the limitations, consider the following diagram:

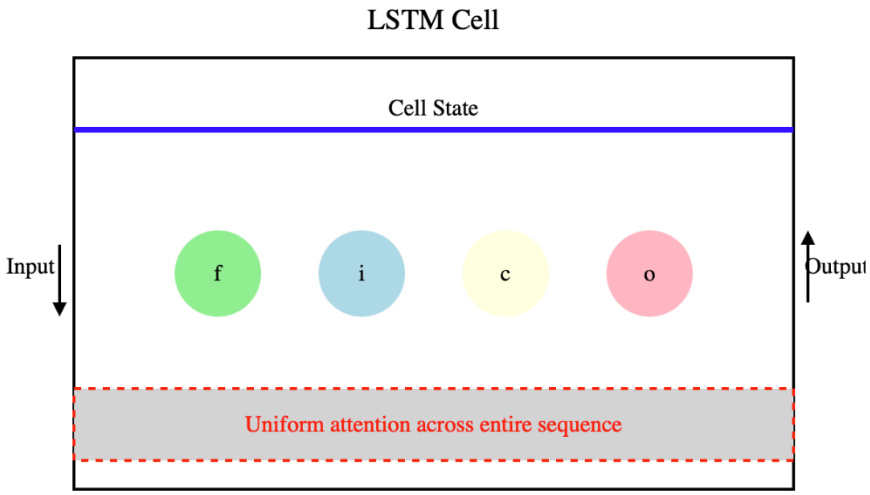

Figure 32: Structure of the LSTM cell

This diagram (figure 32) shows the key components of an LSTM cell and how it processes information. The cell state, represented by the horizontal line running through the top of the cell, acts as a conveyor belt of information flowing through the entire sequence. The various gates (f, i, c, and o) control the flow of information into, out of, and within the cell.

However, it also shows an important limitation of LSTMs for MDS. The red dashed area at the bottom represents the 'attention problem'. In an LSTM, attention is uniformly distributed across the entire input sequence. This means that when processing multiple documents or very long sequences, the model struggles to focus on the most relevant information at each step.

While LSTMs are effective at carrying information forward through their cell state, they lack a mechanism to selectively attend to different parts of the input. This uniform attention is a problem when dealing with multiple documents (or even sections of larger documents), as the model cannot easily prioritise or connect relevant information across various sources. This limitation leads to the requirement for attention-based models which can dynamically focus on different parts of the input, regardless of their position in the sequence.

Limitations of LSTM in Summarisation

Despite their effectiveness in various NLP tasks, LSTMs have several limitations when applied to multi-document summarisation:

a) **Sequential Processing**: LSTMs process input sequentially, which can be inefficient for long documents or multiple documents. This sequential nature makes it challenging to capture global context efficiently, potentially leading to loss of important information from earlier parts of the input.

b) **Fixed-Size Memory**: The cell state in LSTMs has a fixed size, which can be a bottleneck when dealing with large amounts of information from multiple documents. This limitation can result in information loss or dilution, particularly for longer input sequences.

c) **Lack of Inherent Attention**: Vanilla LSTMs do not have an built-in attention mechanism. While attention can be added to LSTM-based models, the base architecture does not inherently support selective focus on different parts of the input.

The Attention Problem

The absence of this attention mechanism in LSTMs leads to several issues in the context of MDS:

a) **Uniform Information Processing:** Without attention, LSTMs tend to process all parts of the input sequence with equal importance. This uniform approach is suboptimal for summarisation, where certain segments of the input documents are more important and relevant than others.

b) **Context Dilution**: As the input sequence grows longer, the fixed-size cell state struggles to maintain all relevant information. This can lead to a dilution of context, where important details from earlier parts of the sequence are gradually lost or overshadowed by more recent inputs.

c) **Inefficient Handling of Document Boundaries**: In multi-document summarisation, LSTMs may struggle to effectively differentiate between document boundaries, potentially leading to confusion in cross-document references and themes.

To illustrate these issues, consider a scenario where an LSTM is used to summarise multiple news articles about a complex event. Without attention, the model might give equal weight to introductory paragraphs, key event details, and borderline ‘filler’ information. This could result in a summary that lacks focus on the most critical aspects of the event and fails to coherently integrate information from different sources.

Take, for example, this scenario to summarise three news articles about a major climate conference:

Article 1 (excerpt):

"The 2023 Global Climate Summit opened today in Geneva, Switzerland. Over 190 countries are represented. The conference aims to address urgent climate issues."

Article 2 (excerpt):

"Delegates at the Global Climate Summit are debating a new emissions reduction target. The proposed goal is a 50% cut by 2030. Some nations argue this is too ambitious."

Article 3 (excerpt):

"Protests have erupted outside the Global Climate Summit venue. Activists demand immediate action. Police have maintained a strong presence to ensure security."

An LSTM might summarise these without and with attention:

Without Attention:

"The Global Climate Summit opened in Geneva with 190 countries. Delegates are debating emissions targets. Protests have occurred outside the venue. The conference aims to address climate issues. Some nations argue about the proposed goals. Police are maintaining security."

In this summary, the LSTM gives roughly equal weight to all pieces of information. It includes details from each article but doesn't emphasise the most important points or create a coherent narrative.

With Attention:

"The 2023 Global Climate Summit in Geneva has brought together 190 countries to address urgent climate issues. The key focus is a proposed 50% emissions reduction target by 2030, which has sparked debate among delegates. Meanwhile, activists outside the venue are demanding immediate action, highlighting the pressure on world leaders to achieve meaningful results."

Impact on Summarisation

The limitations of LSTMs, particularly the attention problem, significantly affect the quality of generated summaries:

a) **Coherence**: The sequential processing and lack of global context awareness can lead to summaries that lack overall coherence. The model may struggle to maintain a consistent narrative or thematic flow across information from multiple documents.

b) **Information Coverage**: Without an effective mechanism to prioritise important content, LSTM-based summarisers may fail to capture key information dispersed across multiple

documents. This can result in summaries that miss details or overemphasise less important aspects.

c) **Redundancy**: The inability to efficiently track what information has already been included in the summary can lead to redundant content, particularly when dealing with multiple documents that contain overlapping information.
d) **Length Sensitivity**: LSTM performance tends to degrade with longer input sequences, which is particularly problematic in multi-document summarisation where input lengths can be substantial.

Case Studies and Examples

Several studies have demonstrated the limitations of LSTM-based models for MDS:

Tan, Wan and Xiao (2017) compared LSTM-based models with and without attention mechanisms for abstractive summarisation. Their results showed that attention-enhanced models consistently outperformed vanilla LSTMs in terms of ROUGE scores and human evaluations of coherence and coverage.

In a study by Liu and Lapata (2019), hierarchical LSTM models were compared with Transformer-based architectures for multi-document summarisation. The LSTM models, even with hierarchical structures, struggled to capture cross-document relationships effectively, leading to lower performance in coherence and information coverage metrics.

Concluding comments

While LSTMs have proven effective in various sequence modelling tasks, their application to multi-document summarisation is hampered by significant limitations, particularly the lack of an inherent attention mechanism. These limitations manifest in challenges with maintaining coherence across multiple documents and ensuring comprehensive coverage of key information.

The sequential nature of LSTMs, combined with their fixed-size memory, makes them ill-suited for tasks requiring a global understanding of large volumes of text. In the context of multi-document summarisation, these shortcomings result in summaries that may lack focus, coherence, and comprehensive coverage of important information.

To address these limitations, research shifted towards models that inherently support attention mechanisms, such as the Transformer-based architectures which form the core of the research in this study. As the main body of the thesis describes, these models offer more flexible and efficient ways of handling long-range dependencies and selective focus on relevant information.

Appendix 2: Human Evaluation Study

To validate the quality of the automatically generated summaries, a human evaluation study was conducted, focusing on two key aspects:

- **Coherence**: How well-organised and logically connected the summary content was.
- **Coverage**: How comprehensively the summary captured the main ideas from the original abstracts.

Methodology:

1. **Participants**: The study recruited 20 students in computer science and related fields, all with experience reading academic papers.
2. **Materials**: Participants were provided with:
    - The original set of 5 paper abstracts
    - The automatically generated summary
    - An evaluation form (described below)
3. **Procedure**: For each set of abstracts and summary, participants were asked to:

    a) Read the 5 original abstracts (5 minutes)

    b) Read the automatically generated summary (2 minutes)

    c) Complete the evaluation form (3 minutes)
4. **Evaluation Form**: Participants rated the following statements on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree):
    - Coherence:
        1. The summary presents ideas in a logical and well-organised manner.
        2. The summary flows smoothly from one point to the next.
        3. The summary avoids redundancy and repetition.
    - Coverage:
        1. The summary captures the main points from all 5 abstracts.
        2. The summary does not omit any significant ideas or findings from the abstracts.
        3. The summary provides a balanced representation of the content across all abstracts.
    - Overall Quality:
        1. The summary effectively condenses the key information from the abstracts.

2. The summary would be useful for quickly understanding the main ideas of these papers.

- Open-ended Feedback: Participants were also asked to provide brief written comments on:
    1. Any notable strengths or weaknesses of the summary
    2. Suggestions for improvement

Study Design

This evaluation methodology is based on the widely-used Pyramid Method for summary evaluation, as proposed by Nenkova and Passonneau (2004). The Pyramid Method assesses summaries based on their content coverage and importance, which aligns with the focus here on coherence and coverage. While the original Pyramid Method has a more complex scoring system, the approach adapts its principles to a Likert scale format for ease of use with less-expert evaluators.

The use of a 5-point Likert scale is supported by research indicating that it provides a good balance between the granularity of response options and cognitive load on participants (Dawes, 2008). This approach allows for more nuanced feedback while remaining straightforward for participants to complete.

The specific questions are designed to address key aspects of summary quality identified in the literature, including coherence, coverage, and overall effectiveness (Mani, 2001). By breaking down these aspects into specific, actionable statements, the evaluation form aims to provide a comprehensive assessment of summary quality.

The summaries were presented to the readers via a Google Form in the following format:

Summary Set 1

Here are the abstracts of five academic papers, please take a few minutes to read them. We will then ask you to read an automatically generated summary and then a few questions of how effective you think the summary is.

Paper 1 abstract:

In the last two decades, automatic extractive text summarization on lectures has demonstrated to be a useful tool for collecting key phrases and sentences that best represent the content. However, many current approaches utilize dated approaches, producing sub-par outputs or requiring several hours of manual tuning to produce meaningful results. Recently, new machine learning architectures have provided mechanisms for extractive summarization through the clustering of output embeddings from deep learning models. This paper reports on the project called “lecture summarization service”, a python-based RESTful service that utilizes the BERT model for text embeddings and K-Means clustering to identify sentences closest to the centroid for summary selection. The purpose of the service was to provide student’s a utility that could summarize lecture content, based on their desired number of sentences. On top of summary work, the service also includes lecture and summary management, storing content on the cloud which can be used for collaboration. While the results of utilizing BERT for extractive text summarization were promising, there were still areas where the model struggled, providing future research opportunities for further improvement.

Paper 2 abstract:

Text summarization is a process of extracting the context of a large document and summarize it into a smaller paragraph or a few sentences. Text summarization plays a vital role in saving time in our day to day life. It is also used in many bigger project implementations of classification of documents or in search engines. This paper presents a method of achieving text summaries accurately using deep learning methods.

Paper 3 abstract:

The qualities of human readable summaries available in the datasets are not up to the mark, leading to issues in creating an accurate model for text summarization. Although recent works have been largely built upon this issue and set up a strong platform for further improvements, they still have many limitations. Looking in this direction, the paper proposes a novel methodology for summarizing a corpus of documents to generate a coherent summary using topic modeling and classification technique. The objectives of the propose work are highlighted below.

A novel heuristic approach is introduced to find out the actual number of topics that exist in a corpus of documents which handles the stochastic nature of latent dirichlet allocation.

A large corpus of documents is handled by minimizing the huge set of sentences into a small set without losing the important one and thus providing a concise and information rich summary at the end.

Ensuring that the sentences are arranged as per their importance in the coherent summary.

Results of the experiment are compared with the state-of-the-art summary systems.

The outcomes of the empirical work show that the proposed model is more promising compared to the well-known text summarization models.

Paper 4 abstract:

Abstractive Text Summarization (ATS), which is the task of constructing summary sentences by merging facts from different source sentences and condensing them into a shorter representation while preserving information content and overall meaning. It is very difficult and time consuming for human beings to manually summarize large documents of text. In this paper, we propose an LSTM-CNN based ATS framework (ATSDL) that can construct new sentences by exploring more fine-grained fragments than sentences, namely, semantic phrases. Different from existing abstraction based approaches, ATSDL is composed of two main stages, the first of which extracts phrases from source sentences and the second generates text summaries using deep learning. Experimental results on the datasets CNN and DailyMail show that our ATSDL framework outperforms the state-of-the-art models in terms of both semantics and syntactic structure, and achieves competitive results on manual linguistic quality evaluation.

Paper 5 abstract:

Recent neural network approaches to summarization are largely either selection-based extraction or generation-based abstraction. In this work, we present a neural model for single-document summarization based on joint extraction and syntactic compression. Our model chooses sentences from the document, identifies possible compressions based on constituency parses, and scores those compressions with a neural model to produce the final summary. For learning, we construct oracle extractive-compressive summaries, then learn both of our components jointly with this supervision. Experimental results on the CNN/Daily Mail and New York Times datasets show that our model achieves strong performance (comparable to state-of-the-art systems) as evaluated by ROUGE. Moreover, our approach outperforms an off-the-shelf compression module, and human and manual evaluation shows that our model's output generally remains grammatical.

Automatically generated summary:

Text summarization is a process of extracting the most important information from a large document and condensing it into a smaller, more manageable format. Automatic extractive text summarization has become a popular tool for this task, particularly for long lectures. However, traditional approaches often produce sub-par results and require significant manual tuning.

Recent papers have proposed new techniques to improve text summarization using deep learning methods, particularly the BERT model. The BERT model has been used to improve the accuracy of summarization, by clustering output embeddings and identifying sentences closest to the centroid for summary selection. The service also includes lecture and summary management and storing content on the cloud for collaboration.

Research has also been focused on improving the interpretability and coherence of the extractive summarization models. One method proposed in these papers is the use of encoder-decoder architecture, which is trained on a large dataset of documents to extract important sentences. The method of graph-based text summarization also proposed in one of the papers, that utilizes coreference resolution and sentiment analysis to extract important sentences and improve summary coherence.

All in all, automatic extractive text summarization has become a valuable tool for reducing the time and effort required to read large amounts of text. The use of deep learning models like BERT has shown promising results for producing accurate summaries, but there are still areas for improvement in terms of interpretability, coherence, and the ability to handle complex text. Therefore, future research opportunities are focused on improving these areas.

Followed by a Form Section to gather responses via a Likert scale and a free text input (figure 33):

How effectively does the summary capture the ideas of these individual papers as explained in the abstracts? *

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
The computer generated summary captures the main points and key themes of the papers effectively.	○	○	○	○	○
The summary provided by the computer generated model accurately conveys the main contributions and findings of the papers.	○	○	○	○	○
The computer generated summary is easy to understand and free of errors or misunderstandings.	○	○	○	○	○
The computer generated summary is comparable to the abstracts of the papers in terms of length, clarity, and completeness?	○	○	○	○	○
There are few significant differences between the summary provided by the computer generated model and the abstracts of the papers.	○	○	○	○	○

If there are significant differences, *what are they and how might they impact the reader's understanding of the papers?* *

Your answer

Figure 33: Response form as it appears to the respondent

Analysis

Mean scores for each criterion were calculated, and the open-ended feedback was analysed to identify common themes. This approach allowed for an assessment of the overall quality of the summaries and identification of specific areas for improvement in the summarisation approach.

Results

1. Overall Effectiveness:
    - The majority of respondents agreed or strongly agreed that the computer-generated summaries captured the main points and key themes effectively.
    - There was general agreement that the summaries accurately conveyed the main contributions and findings of the papers.
2. Clarity and Comprehension:
    - Most participants found the summaries easy to understand and free of errors or misunderstandings.
    - However, opinions were mixed regarding the comparability of the summaries to the original abstracts in terms of length, clarity, and completeness.
3. Differences from Original Abstracts:
    - Responses were varied concerning whether there were few significant differences between the summaries and the original abstracts.
    - Common issues mentioned in the open-ended responses included:

      a) Loss of specific details or nuances

      b) Omission of key ideas from some abstracts

      c) Bias towards certain papers or topics (e.g., BERT being mentioned more frequently)

      d) Generalisation of information (e.g., mentioning "tasks" or "mechanisms" without specifics)

Key Observations:

1. The summaries generally succeeded in capturing main points and themes, but often at the cost of specific details or nuances.
2. There was a consistent concern about the loss of important information or examples that might be important for a comprehensive understanding of the original papers.

3. The summaries were generally easy to understand, suggesting good coherence and readability.
4. Many participants expressed a desire to see both the written summary and a diagram, suggesting that a multi-modal approach to summarisation might be beneficial.

Limitations and Considerations:

- There was some inconsistency in responses across different summary sets, which could indicate variability in the quality of different summaries or differences in the complexity of the original abstracts.

Appendix 3: LLM-as-a-Judge

In addition to the human evaluation study described in Appendix 2, a larger-scale, automated evaluation of the summarisation outputs was conducted using a state-of-the-art Language Model (LLM), specifically the OpenAI GPT-4 API. This appendix details the methodology, rationale, and implications of this approach.

Although the LLM-as-a-judge methodology is relatively new, based on the findings presented by Zheng *et al.* (2023), adopting an LLM-as-a-judge approach for evaluating the outputs of a summariser is well-justified. The authors demonstrate that strong LLM judges, particularly GPT-4, can effectively match human preferences with over 80% agreement rate, which is comparable to the level of agreement between human evaluators. The approach can be summarised in the following diagram:

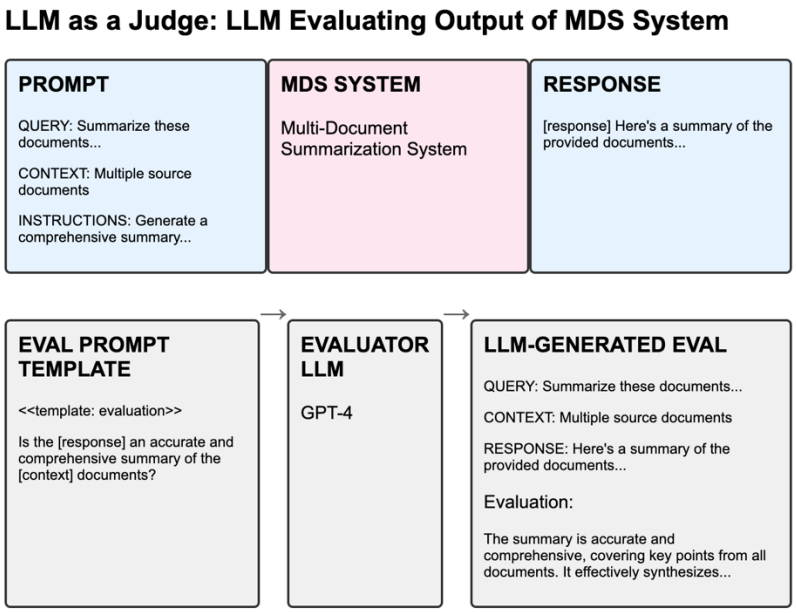

Figure 34: Summary of the LLM-as-a-judge process

Figure 34 illustrates the process of using an LLM (specifically GPT-4) as a judge to evaluate the output of an MDS system. The flow shows how input documents and a summarisation query are processed by the MDS system to produce a summary, which is then evaluated by GPT-4 using a specialised evaluation prompt. The approach offers significant advantages in terms of scalability and speed compared to traditional human evaluations, while still providing explainable outputs through detailed judgments. The paper does acknowledge some of the potential biases and limitations of LLM judges but shows that many of these can be mitigated or have minor impacts.

Rationale:

Following on from this, the use of an LLM as an evaluator serves several purposes:

1. **Scalability**: Unlike human evaluations, which are limited by time and resources, an LLM-based approach allows for the assessment of a much larger number of summaries, potentially covering the entire dataset.
2. **Reproducibility**: The consistent nature of LLM evaluations ensures that the results can be easily reproduced, addressing a common challenge in human-based studies.
3. **Comparative Analysis**: By mirroring the evaluation criteria used in the human study, it is possible to directly compare the assessments made by humans with those made by the LLM, potentially revealing insights into the reliability and biases of both evaluation methods.
4. **Continuous Evaluation**: This method allows for ongoing assessment as new summaries are generated, facilitating iterative improvements to the summarisation model.
5. **Cognitive Load**: Even though the human evaluators in the small-scale study were 'expert' academic readers, several were honest in informal feedback. There was a suggestion that human readers quickly tire of reading dense text and attention wanders. It was also suggested that the same reader may even give different evaluations on different days; indeed this feedback is supported by the literature. Sweller (1988) argues that cognitive load can significantly impact task performance, noting that *"problem solving imposes a heavy cognitive load on problem solvers"*, which can lead to fatigue and inconsistent results over time.

Methodology:

The GPT-4 API was employed to evaluate the summarisation outputs using prompts designed to mirror the questions and criteria used in the human evaluation study. The LLM was tasked with assessing aspects such as coherence, coverage, accuracy, and overall quality of the generated summaries. The process is shown below in figure 35.

Setup

The evaluation was implemented using a Python script, using the OpenAI API client library. API authentication was managed securely using environment variables to store the API key.

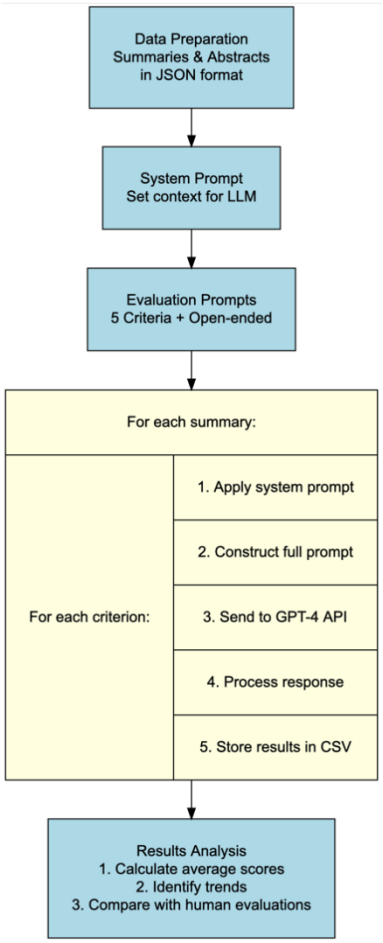

Figure 35: LLM-as-a-judge implementation

Data Preparation

Summaries for evaluation were stored in a structured JSON format, with each entry containing:

- Summary ID
- Generated summary text
- Original texts (for reference)

Evaluation Prompts

The LLM evaluation used the following prompts, giving clear instructions via a 'system prompt' and a set of focussed questions directly mirroring the human evaluation questions:

System Prompt:

You are an expert academic researcher tasked with evaluating the quality of computer-generated summaries of scientific papers. Your role is to assess these summaries objectively, comparing them to the original abstracts of the papers. Please adhere to the following guidelines:

1. Provide ratings on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree) for each criterion.
2. After the ratings, provide a clear, concise explanation for your choice.
3. Base your evaluation solely on the content provided, without making assumptions about information not present.
4. Be critical yet fair in your assessment, highlighting both strengths and weaknesses.
5. For the open-ended question, provide specific examples and explain their potential impact on reader understanding.

Your goal is to give an accurate, unbiased evaluation that will help improve the quality of automated summarization techniques in academic contexts.

Focus questions:

1. Main Points and Key Themes:

   "How effectively does the summary capture the main points and key themes of the papers as explained in the abstracts? Rate on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"
2. Accuracy of Main Contributions and Findings:

   "How accurately does the summary convey the main contributions and findings of the papers? Rate on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"
3. Clarity and Error-Free Content:

   "Is the computer-generated summary easy to understand and free of errors or misunderstandings? Rate on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"
4. Comparability to Abstracts:

   "Is the computer-generated summary comparable to the abstracts of the papers in terms of length, clarity, and completeness? Rate on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"
5. Significant Differences:

   "Are there few significant differences between the summary provided by the computer-generated model and the abstracts of the papers? Rate on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"
6. Open-ended Question:

   "If there are significant differences, what are they and how might they impact the reader's understanding of the papers?"

Evaluation Process

The script implemented an iterative process:

- For each summary in the dataset:
    - For each evaluation criterion:

1. Construct the full prompt by combining the system prompt and the criterion-specific prompts with the summary and original text.
2. Send the prompt to the GPT-4 API.
3. Process the API response to extract the numerical rating and explanation.
4. Append the results (summary ID, criterion, rating, explanation) to a CSV file.

- Error handling and rate limiting were implemented to manage API failures and comply with usage restrictions.

Data Storage

Results were stored in a CSV file with the following structure:

- Summary ID
- Evaluation Criterion
- Numerical Rating (1-5)
- Explanation

Analysis

Post-evaluation, a separate script analysed the CSV file to:

- Calculate average scores for each criterion
- Identify trends in the evaluations
- Compare LLM-based results with human evaluation results

Limitations and Considerations

This methodology aimed to complement, not replace, human evaluation. The comparison between LLM and human evaluations was intended to help identify potential biases or limitations in both approaches, contributing to the development of a more robust evaluation framework.

Bibliography

Afantenos, S., Karkaletsis, V. and Stamatopoulos, P. (2005) 'Summarization from medical documents: a survey', *Artificial Intelligence in Medicine*, 33(2), pp. 157–177. Available at: https://doi.org/10.1016/j.artmed.2004.07.017.

Allahyari, M. *et al.* (2017) 'Text Summarization Techniques: A Brief Survey', *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(10). Available at: https://doi.org/10.14569/IJACSA.2017.081052.

Altmami, N.I. and Menai, M.E.B. (2018) 'Semantic Graph Based Automatic Summarization of Multiple Related Work Sections of Scientific Articles', in G. Agre, J. van Genabith, and T. Declerck (eds) *Artificial Intelligence: Methodology, Systems, and Applications*. Cham: Springer International Publishing, pp. 255–259. Available at: https://doi.org/10.1007/978-3-319-99344-7_23.

Amatriain, X. *et al.* (2023) 'Transformer models: an introduction and catalog'. arXiv. Available at: http://arxiv.org/abs/2302.07730 (Accessed: 30 October 2023).

Anglin, K., Boguslav, A. and Hall, T. (2022) 'Improving the Science of Annotation for Natural Language Processing: The Use of the Single-Case Study for Piloting Annotation Projects', *Journal of Data Science*, 20(3), pp. 339–357. Available at: https://doi.org/10.6339/22-JDS1054.

Anthropic (no date) *Introducing Claude*. Available at: https://www.anthropic.com/news/introducing-claude (Accessed: 10 July 2024).

Bahdanau, D., Cho, K. and Bengio, Y. (2016) 'Neural Machine Translation by Jointly Learning to Align and Translate'. arXiv. Available at: https://doi.org/10.48550/arXiv.1409.0473.

Bai, H. *et al.* (2022) 'EK-BERT: An Enhanced K-BERT Model for Chinese Sentiment Analysis', in M. Sun et al. (eds) *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*. Singapore: Springer Nature, pp. 136–147. Available at: https://doi.org/10.1007/978-981-19-7596-7_11.

Banko, M. and Brill, E. (2001) 'Scaling to Very Very Large Corpora for Natural Language Disambiguation', in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. *ACL 2001*, Toulouse, France: Association for Computational Linguistics, pp. 26–33. Available at: https://doi.org/10.3115/1073012.1073017.

Bano, S. *et al.* (2018) 'Document Summarization Using Clustering and Text Analysis', *International Journal of Engineering & Technology*, 7, p. 456. Available at: https://doi.org/10.14419/ijet.v7i2.32.15740.

Baram-Tsabari, A. and Lewenstein, B.V. (2013) 'An Instrument for Assessing Scientists' Written Skills in Public Communication of Science', *Science Communication*, 35(1), pp. 56–85. Available at: https://doi.org/10.1177/1075547012440634.

Barzilay, R., McKeown, K.R. and Elhadad, M. (1999) 'Information Fusion in the Context of Multi-Document Summarization', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. *ACL 1999*, College Park, Maryland, USA: Association for Computational Linguistics, pp. 550–557. Available at: https://doi.org/10.3115/1034678.1034760.

Bautista, C. *et al.* (2022) 'Ten simple rules for improving communication among scientists', *PLOS Computational Biology*. Edited by R. Schwartz, 18(6), p. e1010130. Available at: https://doi.org/10.1371/journal.pcbi.1010130.

Bedi, P., Bala, M. and Sharma, K. (2023) 'Extractive text summarization for biomedical transcripts using deep dense LSTM-CNN framework', *Expert Systems*, 41. Available at: https://doi.org/10.1111/exsy.13490.

Beltagy, I., Lo, K. and Cohan, A. (2019) 'SciBERT: A Pretrained Language Model for Scientific Text'. arXiv. Available at: http://arxiv.org/abs/1903.10676 (Accessed: 24 July 2023).

Beltagy, I., Peters, M.E. and Cohan, A. (2020) 'Longformer: The Long-Document Transformer'. arXiv. Available at: http://arxiv.org/abs/2004.05150 (Accessed: 24 July 2023).

Belz, A. *et al.* (2021) 'A Systematic Review of Reproducibility Research in Natural Language Processing', in P. Merlo, J. Tiedemann, and R. Tsarfaty (eds) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. *EACL 2021*, Online: Association for Computational Linguistics, pp. 381–393. Available at: https://doi.org/10.18653/v1/2021.eacl-main.29.

Bender, E.M. *et al.* (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922.

Bhandari, M. *et al.* (2020) 'Re-evaluating Evaluation in Text Summarization', in B. Webber et al. (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *EMNLP 2020*, Online: Association for Computational Linguistics, pp. 9347–9359. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.751.

Bhattacharya, P. *et al.* (2019) 'A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments', in L. Azzopardi et al. (eds). Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 413–428. Available at: https://doi.org/10.1007/978-3-030-15712-8_27.

Bianchi, F. and Hovy, D. (2021) 'On the Gap between Adoption and Understanding in NLP', in C. Zong et al. (eds) *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. *Findings 2021*, Online: Association for Computational Linguistics, pp. 3895–3901. Available at: https://doi.org/10.18653/v1/2021.findings-acl.340.

BigScience Workshop *et al.* (2023) 'BLOOM: A 176B-Parameter Open-Access Multilingual Language Model'. arXiv. Available at: https://doi.org/10.48550/arXiv.2211.05100.

Bird, S. *et al.* (2008) 'The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics', in N. Calzolari et al. (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. *LREC 2008*, Marrakech, Morocco: European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf (Accessed: 11 July 2024).

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent dirichlet allocation', *J. Mach. Learn. Res.*, 3(null), pp. 993–1022.

Blodgett, S.L. *et al.* (2020) 'Language (Technology) is Power: A Critical Survey of "Bias" in NLP'. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14050.

Bommasani, R. *et al.* (2024) 'Foundation Model Transparency Reports'. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.16268.

Bornmann, L. and Mutz, R. (2015) 'Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references', *Journal of the Association for Information Science and Technology*, 66(11), pp. 2215–2222. Available at: https://doi.org/10.1002/asi.23329.

Boyack, K.W. and Klavans, R. (2019) 'Creation and Analysis of Large-Scale Bibliometric Networks', in W. Glänzel et al. (eds) *Springer Handbook of Science and Technology Indicators*. Springer Verlag, pp. 187–212.

Brown, T. *et al.* (2020) 'Language Models are Few-Shot Learners', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 1877–1901. Available at: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (Accessed: 11 July 2024).

Brown, T.B. *et al.* (2020) 'Language Models are Few-Shot Learners'. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.14165.

Brownell, S.E., Price, J.V. and Steinman, L. (2013) 'Science Communication to the General Public: Why We Need to Teach Undergraduate and Graduate Students this Skill as Part of Their Formal Scientific Training', *Journal of Undergraduate Neuroscience Education*, 12(1), pp. E6–E10. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852879/ (Accessed: 10 July 2024).

Cabanac, G., Frommholz, I. and Mayr, P. (2019) *Bibliometric-enhanced information retrieval: 8th international BIR workshop*. Springer Verlag. Available at: https://uobrep.openrepository.com/handle/10547/624250 (Accessed: 10 July 2024).

Cachola, I. *et al.* (2020) 'TLDR: Extreme Summarization of Scientific Documents'. arXiv. Available at: http://arxiv.org/abs/2004.15011 (Accessed: 3 July 2024).

Callaghan, M. (2023) 'Cloud Computing for Metagenomics: Building a Personalized Computational Platform for Pipeline Analyses', in S. Mitra (ed.) *Metagenomic Data Analysis*. New York, NY: Springer US, pp. 261–279. Available at: https://doi.org/10.1007/978-1-0716-3072-3_13.

Callaghan, M. (2024a) 'Chatting with My Data: LLMs for Biodata'. *Festival of Genomics 2024*, London, UK, 24 January.

Callaghan, M. (2024b) 'Multimodal AI for Enhanced Information Extraction from Complex HPC Documentation'. *Second Workshop on Multimodal AI*, 25 June. Available at: https://multimodalai.github.io.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006) 'Re-evaluating the Role of Bleu in Machine Translation Research', in D. McCarthy and S. Wintner (eds) *11th Conference of the European Chapter of the Association for Computational Linguistics*. *EACL 2006*, Trento, Italy: Association for Computational Linguistics, pp. 249–256. Available at: https://aclanthology.org/E06-1032 (Accessed: 18 July 2024).

Carbonell, J. and Goldstein, J. (1998) 'The use of MMR, diversity-based reranking for reordering documents and producing summaries', in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery (SIGIR '98), pp. 335–336. Available at: https://doi.org/10.1145/290941.291025.

Celikyilmaz, A., Clark, E. and Gao, J. (2021) ‘Evaluation of Text Generation: A Survey’. arXiv. Available at: https://doi.org/10.48550/arXiv.2006.14799.

Chu, C. and Wang, R. (2018) ‘A Survey of Domain Adaptation for Neural Machine Translation’, in E.M. Bender, L. Derczynski, and P. Isabelle (eds) *Proceedings of the 27th International Conference on Computational Linguistics*. *COLING 2018*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1304–1319. Available at: https://aclanthology.org/C18-1111 (Accessed: 26 July 2024).

Coavoux, M., Elsahar, H. and Gallé, M. (2019) ‘Unsupervised Aspect-Based Multi-Document Abstractive Summarization’, in L. Wang et al. (eds) *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Hong Kong, China: Association for Computational Linguistics, pp. 42–47. Available at: https://doi.org/10.18653/v1/D19-5405.

Cohan, A. *et al.* (2018) ‘A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents’, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 615–621. Available at: https://doi.org/10.18653/v1/N18-2097.

Cohan, A. and Goharian, N. (2015) ‘Scientific Article Summarization Using Citation-Context and Article’s Discourse Structure’, in L. Màrquez, C. Callison-Burch, and J. Su (eds) *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2015*, Lisbon, Portugal: Association for Computational Linguistics, pp. 390–400. Available at: https://doi.org/10.18653/v1/D15-1045.

Cohan, A. and Goharian, N. (2016) ‘Revisiting Summarization Evaluation for Scientific Articles’, in N. Calzolari et al. (eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. *LREC 2016*, Portorož, Slovenia: European Language Resources Association (ELRA), pp. 806–813. Available at: https://aclanthology.org/L16-1130 (Accessed: 11 July 2024).

Cohan, A. and Goharian, N. (2018) ‘Scientific document summarization via citation contextualization and scientific discourse’, *International Journal on Digital Libraries*, 19(2), pp. 287–303. Available at: https://doi.org/10.1007/s00799-017-0216-8.

Cohen, A.M. *et al.* (2006) ‘Reducing Workload in Systematic Review Preparation Using Automated Citation Classification’, *Journal of the American Medical Informatics Association : JAMIA*, 13(2), pp. 206–219. Available at: https://doi.org/10.1197/jamia.M1929.

Cohn, T. and Lapata, M. (2008) ‘Sentence Compression Beyond Word Deletion’, in D. Scott and H. Uszkoreit (eds) *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. *COLING 2008*, Manchester, UK: Coling 2008 Organizing Committee, pp. 137–144. Available at: https://aclanthology.org/C08-1018 (Accessed: 12 July 2024).

Committee on Science, Engineering and Public Policy (2009) *On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition | The National Academies Press*. Washington District of Columbia. Available at: https://nap.nationalacademies.org/catalog/12192/on-being-a-scientist-a-guide-to-responsible-conduct-in.

Conroy, J. and Davis, S. (2018) ‘Section mixture models for scientific document summarization’, *International Journal on Digital Libraries*, 19(2), pp. 305–322. Available at: https://doi.org/10.1007/S00799-017-0218-6.

Crane, M. (2018) 'Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results', *Transactions of the Association for Computational Linguistics*. Edited by L. Lee et al., 6, pp. 241–252. Available at: https://doi.org/10.1162/tacl_a_00018.

Creswell, J.W. *et al.* (2006) 'How Interpretive Qualitative Research Extends Mixed Methods Research'.

Creswell, J.W. and Plano-Clark, V.L. (2017) *Designing and conducting mixed methods research*. Sage Publications.

Curran, I. (2023) *Data centres consume as much electricity as urban houses, CSO figures show*, *The Irish Times*. Available at: https://www.irishtimes.com/business/2023/06/12/data-centres-consume-as-much-electricity-as-urban-houses-cso/ (Accessed: 24 July 2024).

Dang, H.T. (2006) 'Overview of DUC 2006'. Available at: https://duc.nist.gov/pubs/2006papers/duc2006.pdf.

Danilevsky, M. *et al.* (2020) 'A Survey of the State of Explainable AI for Natural Language Processing', in K.-F. Wong, K. Knight, and H. Wu (eds) *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. *AACL 2020*, Suzhou, China: Association for Computational Linguistics, pp. 447–459. Available at: https://aclanthology.org/2020.aacl-main.46 (Accessed: 11 July 2024).

Das, D. and Martins, A.F.T. (2007) 'A Survey on Automatic Text Summarization', *Eighth ACIS International Conference on Software Engineering Artificial Intelligence Networking and ParallelDistributed Computing SNPD 2007*, 4, pp. 574–578. Available at: https://doi.org/10.1016/B0-08-044854-2/00957-3.

Dawes, J. (2008) 'Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales', *International Journal of Market Research*, 50(1), pp. 61–104. Available at: https://doi.org/10.1177/147078530805000106.

De Semir, V. (2009) 'Master in Scientific, Medical and Environmental Communication', *Journal of Science Communication*, 08(01), p. C02. Available at: https://doi.org/10.22323/2.08010302.

Denzin, N.K. (2017) *The Research Act: A Theoretical Introduction to Sociological Methods*. New York: Routledge. Available at: https://doi.org/10.4324/9781315134543.

Dettmers, T. *et al.* (2023) 'QLoRA: Efficient Finetuning of Quantized LLMs'. arXiv. Available at: https://doi.org/10.48550/arXiv.2305.14314.

Devlin, J. *et al.* (2018) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. Available at: http://arxiv.org/abs/1810.04805.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. arXiv. Available at: http://arxiv.org/abs/1810.04805 (Accessed: 24 July 2023).

DeYoung, J. *et al.* (2020) 'ERASER: A Benchmark to Evaluate Rationalized NLP Models', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 4443–4458. Available at: https://doi.org/10.18653/v1/2020.acl-main.408.

Dong, Y. *et al.* (2024) ‘Building Guardrails for Large Language Models’. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.01822.

Doshi-Velez, F. and Kim, B. (2017) ‘Towards A Rigorous Science of Interpretable Machine Learning’. arXiv. Available at: https://doi.org/10.48550/arXiv.1702.08608.

Dou, Z.-Y. *et al.* (2021) ‘GSum: A General Framework for Guided Neural Abstractive Summarization’, in K. Toutanova et al. (eds) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. *NAACL-HLT 2021*, Online: Association for Computational Linguistics, pp. 4830–4842. Available at: https://doi.org/10.18653/v1/2021.naacl-main.384.

Driessen, V. (2010) *A successful Git branching model*, *nvie.com*. Available at: http://nvie.com/posts/a-successful-git-branching-model/ (Accessed: 8 July 2024).

El-Kassas, W.S. *et al.* (2021) ‘Automatic text summarization: A comprehensive survey’, *Expert Systems with Applications*, 165, p. 113679. Available at: https://doi.org/10.1016/j.eswa.2020.113679.

Endres-Niggemeyer, B. *et al.* (1998) *Summarizing Information*. Berlin, Heidelberg: Springer. Available at: https://doi.org/10.1007/978-3-642-72025-3.

Er Saw, P. (2020) ‘BIOI Virtual Academic Series PART 1: Multidisciplinary Integration in Academia’, *BIO Integration*, 1(2), pp. 101–103. Available at: https://doi.org/10.15212/bioi-2020-0031.

Erkan, G. and Radev, D.R. (2004) ‘LexRank: graph-based lexical centrality as salience in text summarization’, *J. Artif. Int. Res.*, 22(1), pp. 457–479.

Fabbri, A.R. *et al.* (2019) ‘Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model’. arXiv. Available at: https://doi.org/10.48550/arXiv.1906.01749.

Fabbri, A.R. *et al.* (2021) ‘SummEval: Re-evaluating Summarization Evaluation’, *Transactions of the Association for Computational Linguistics*, 9, pp. 391–409. Available at: https://doi.org/10.1162/tacl_a_00373.

Feng, H. *et al.* (2023) ‘UniDoc: A Universal Large Multimodal Model for Simultaneous Text Detection, Recognition, Spotting and Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.2308.11592.

Feng, S. *et al.* (2023) ‘KALM: Knowledge-Aware Integration of Local, Document, and Global Contexts for Long Document Understanding’, in A. Rogers, J. Boyd-Graber, and N. Okazaki (eds) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2023*, Toronto, Canada: Association for Computational Linguistics, pp. 2116–2138. Available at: https://doi.org/10.18653/v1/2023.acl-long.118.

Floridi, L. and Cowls, J. (2019) ‘A Unified Framework of Five Principles for AI in Society’, *Harvard Data Science Review*, 1(1). Available at: https://doi.org/10.1162/99608f92.8cd550d1.

Gage, P. (1994) ‘A new algorithm for data compression’, *The C Users Journal archive* [Preprint]. Available at: https://www.semanticscholar.org/paper/A-new-algorithm-for-data-compression-Gage/1aa9c0045f1fe8c79cce03c7c14ef4b4643a21f8 (Accessed: 18 July 2024).

Gao, Y. *et al.* (2024) 'Retrieval-Augmented Generation for Large Language Models: A Survey'. arXiv. Available at: http://arxiv.org/abs/2312.10997 (Accessed: 17 January 2024).

Gao, Y., Zhao, W. and Eger, S. (2020) 'SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 1347–1354. Available at: https://doi.org/10.18653/v1/2020.acl-main.124.

Gehrmann, S., Deng, Y. and Rush, A. (2018) 'Bottom-Up Abstractive Summarization', in E. Riloff et al. (eds) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2018*, Brussels, Belgium: Association for Computational Linguistics, pp. 4098–4109. Available at: https://doi.org/10.18653/v1/D18-1443.

Gemma Team *et al.* (2024) 'Gemma: Open Models Based on Gemini Research and Technology'. arXiv. Available at: https://doi.org/10.48550/arXiv.2403.08295.

Gholami, A. *et al.* (2021) 'A Survey of Quantization Methods for Efficient Neural Network Inference'. arXiv. Available at: https://doi.org/10.48550/arXiv.2103.13630.

Gholipour Ghalandari, D. and Ifrim, G. (2020) 'Examining the State-of-the-Art in News Timeline Summarization', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 1322–1334. Available at: https://doi.org/10.18653/v1/2020.acl-main.122.

Ghosh, S. *et al.* (2024) 'A Closer Look at the Limitations of Instruction Tuning'. arXiv. Available at: https://doi.org/10.48550/arXiv.2402.05119.

Goldstein, J., Mittal, V., Carbonell, J.G., *et al.* (2000) 'Multi-Document Summarization By Sentence Extraction', in. Available at: https://api.semanticscholar.org/CorpusID:8294822.

Goldstein, J., Mittal, V., Carbonell, J., *et al.* (2000) 'Multi-Document Summarization By Sentence Extraction', in *NAACL-ANLP 2000 Workshop: Automatic Summarization*. Available at: https://aclanthology.org/W00-0405 (Accessed: 12 July 2024).

Gong, Y. and Liu, X. (2001) 'Generic text summarization using relevance measure and latent semantic analysis', in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: Association for Computing Machinery (SIGIR '01), pp. 19–25. Available at: https://doi.org/10.1145/383952.383955.

Grimshaw, J.M. *et al.* (2012) 'Knowledge translation of research findings', *Implementation Science*, 7(1), p. 50. Available at: https://doi.org/10.1186/1748-5908-7-50.

Gross, A.G., Harmon, J.E. and Reidy, and M.S. (2002) *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford, New York: Oxford University Press.

Grusky, M., Naaman, M. and Artzi, Y. (2018) 'Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies', in M. Walker, H. Ji, and A. Stent (eds) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. *NAACL-HLT 2018*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 708–719. Available at: https://doi.org/10.18653/v1/N18-1065.

Guo, M. *et al.* (2022) 'LongT5: Efficient Text-To-Text Transformer for Long Sequences', in M. Carpuat, M.-C. de Marneffe, and I.V. Meza Ruiz (eds) *Findings of the Association for Computational Linguistics: NAACL 2022*. *Findings 2022*, Seattle, United States: Association for Computational Linguistics, pp. 724–736. Available at: https://doi.org/10.18653/v1/2022.findings-naacl.55.

Gupta, A., Thadani, K. and O'Hare, N. (2020) 'Effective Few-Shot Classification with Transfer Learning', in D. Scott, N. Bel, and C. Zong (eds) *Proceedings of the 28th International Conference on Computational Linguistics*. *COLING 2020*, Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1061–1066. Available at: https://doi.org/10.18653/v1/2020.coling-main.92.

Gupta, V., Hanges, P.J. and Dorfman, P. (2002) 'Cultural clusters: methodology and findings', *Journal of World Business*, 37(1), pp. 11–15. Available at: https://econpapers.repec.org/article/eeeworbus/v_3a37_3ay_3a2002_3ai_3a1_3ap_3a11-15.htm (Accessed: 10 July 2024).

Gupta, V. and Lehal, G.S. (2010) 'A Survey of Text Summarization Extractive Techniques', *Journal of Emerging Technologies in Web Intelligence*, 2(3), pp. 258–268. Available at: https://doi.org/10.4304/jetwi.2.3.258-268.

Gururangan, S. *et al.* (2020) 'Don't Stop Pretraining: Adapt Language Models to Domains and Tasks', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 8342–8360. Available at: https://doi.org/10.18653/v1/2020.acl-main.740.

Hafeez, R. *et al.* (2018) 'Topic based Summarization of Multiple Documents using Semantic Analysis and Clustering', in *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*. *2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, Islamabad: IEEE, pp. 70–74. Available at: https://doi.org/10.1109/HONET.2018.8551325.

Hahn, U. and Mani, I. (2000) 'The challenges of automatic summarization', *Computer*, 33(11), pp. 29–36. Available at: https://doi.org/10.1109/2.881692.

Han, Y., Liu, C. and Wang, P. (2023) 'A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.11703.

Hermann, K.M. *et al.* (2015) 'Teaching Machines to Read and Comprehend', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. Available at: https://papers.nips.cc/paper_files/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html (Accessed: 18 July 2024).

Hey, T. and Trefethen, A. (2020) 'The Fourth Paradigm 10 Years On', *Informatik Spektrum*, 42(6), pp. 441–447. Available at: https://doi.org/10.1007/s00287-019-01215-9.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8), pp. 1735–1780. Available at: https://doi.org/10.1162/neco.1997.9.8.1735.

Holman, L., Stuart-Fox, D. and Hauser, C.E. (2018) 'The gender gap in science: How long until women are equally represented?', *PLOS Biology*, 16(4), p. e2004956. Available at: https://doi.org/10.1371/journal.pbio.2004956.

Honnibal, M. and Montani, I. (2017) *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. Available at: https://spacy.io.

Hu, E.J. *et al.* (2021) 'LoRA: Low-Rank Adaptation of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2106.09685 (Accessed: 24 July 2023).

Hua, X. and Wang, L. (2017) 'A Pilot Study of Domain Adaptation Effect for Neural Abstractive Summarization', in L. Wang et al. (eds) *Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 100–106. Available at: https://doi.org/10.18653/v1/W17-4513.

Huang, D. *et al.* (2020) 'What Have We Achieved on Text Summarization?', in B. Webber et al. (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *EMNLP 2020*, Online: Association for Computational Linguistics, pp. 446–469. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.33.

Huang, X. and Paul, M.J. (2018) 'Examining Temporality in Document Classification', in I. Gurevych and Y. Miyao (eds) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. *ACL 2018*, Melbourne, Australia: Association for Computational Linguistics, pp. 694–699. Available at: https://doi.org/10.18653/v1/P18-2110.

Inouye, D. and Kalita, J.K. (2011) 'Comparing Twitter Summarization Algorithms for Multiple Post Summaries', in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 298–306. Available at: https://doi.org/10.1109/PASSAT/SocialCom.2011.31.

Isinbayeva, A. and Przepiorka, W. (2024) 'A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from online posts', *Behavior Research Methods*, 56. Available at: https://doi.org/10.3758/s13428-024-02381-9.

Ji, Z. *et al.* (2023) 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys*, 55(12), pp. 1–38. Available at: https://doi.org/10.1145/3571730.

Jiang, A.Q. *et al.* (2023) 'Mistral 7B'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.06825.

Jiang, Y., Zhang, L. and Wang, W. (2022) 'Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning'. arXiv. Available at: http://arxiv.org/abs/2203.06875 (Accessed: 26 June 2024).

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2.

John, A., Premjith, P.S. and Wilscy, M. (2017) 'Extractive multi-document summarization using population-based multicriteria optimization', *Expert Systems with Applications*, 86, pp. 385–397. Available at: https://doi.org/10.1016/j.eswa.2017.05.075.

Johnson, R.B. and Onwuegbuzie, A.J. (2004) 'Mixed Methods Research: A Research Paradigm Whose Time Has Come', *Educational Researcher*, 33(7), pp. 14–26. Available at: https://doi.org/10.3102/0013189X033007014.

Joshi, A. *et al.* (2022) 'RankSum—An unsupervised extractive text summarization based on rank fusion', *Expert Systems with Applications*, 200, p. 116846. Available at: https://doi.org/10.1016/j.eswa.2022.116846.

Kaufman, S. *et al.* (2012) 'Leakage in data mining: Formulation, detection, and avoidance', *ACM Trans. Knowl. Discov. Data*, 6(4), p. 15:1-15:21. Available at: https://doi.org/10.1145/2382577.2382579.

Kazantseva, A. and Szpakowicz, S. (2010) 'Summarizing Short Stories', *Computational Linguistics*, 36(1), pp. 71–109. Available at: https://doi.org/10.1162/coli.2010.36.1.36102.

Kedzie, C., McKeown, K. and Daume III, H. (2019) 'Content Selection in Deep Learning Models of Summarization'. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.12343.

Kipf, T.N. and Welling, M. (2017) 'Semi-Supervised Classification with Graph Convolutional Networks'. arXiv. Available at: http://arxiv.org/abs/1609.02907 (Accessed: 22 November 2023).

Koh, H.Y. *et al.* (2023) 'An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics', *ACM Computing Surveys*, 55(8), pp. 1–35. Available at: https://doi.org/10.1145/3545176.

Kryscinski, W. *et al.* (2020) 'Evaluating the Factual Consistency of Abstractive Text Summarization', in B. Webber et al. (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *EMNLP 2020*, Online: Association for Computational Linguistics, pp. 9332–9346. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.750.

Kumar, V., Choudhary, A. and Cho, E. (2021) 'Data Augmentation using Pre-trained Transformer Models'. arXiv. Available at: http://arxiv.org/abs/2003.02245 (Accessed: 20 November 2023).

Lample, G. and Conneau, A. (2019) 'Cross-lingual Language Model Pretraining'. arXiv. Available at: https://doi.org/10.48550/arXiv.1901.07291.

Läubli, S. *et al.* (2020) 'A Set of Recommendations for Assessing Human-Machine Parity in Language Translation', *Journal of Artificial Intelligence Research*, 67. Available at: https://doi.org/10.1613/jair.1.11371.

Lauscher, A. *et al.* (2018) 'Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models', in E. Riloff et al. (eds) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2018*, Brussels, Belgium: Association for Computational Linguistics, pp. 3326–3338. Available at: https://doi.org/10.18653/v1/D18-1370.

Lavie, A. and Agarwal, A. (2007) 'METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments', in C. Callison-Burch et al. (eds) *Proceedings of the Second Workshop on Statistical Machine Translation*. *WMT 2007*, Prague, Czech Republic: Association for Computational Linguistics, pp. 228–231. Available at: https://aclanthology.org/W07-0734 (Accessed: 11 July 2024).

Lehman, A. and Miller, S.J. (2020) 'A Theoretical Conversation about Responses to Information Overload', *Information*, 11(8), p. 379. Available at: https://doi.org/10.3390/info11080379.

Lewis, M. *et al.* (2020) 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for

Computational Linguistics, pp. 7871–7880. Available at: https://doi.org/10.18653/v1/2020.acl-main.703.

Lewis, P. *et al.* (2021) 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. arXiv. Available at: https://doi.org/10.48550/arXiv.2005.11401.

Li, J., Li, L. and Li, T. (2012) 'Multi-document summarization via submodularity', *Applied Intelligence*, 37(3), pp. 420–430. Available at: https://doi.org/10.1007/s10489-012-0336-1.

Li, Z. *et al.* (2023) 'Guiding Large Language Models via Directional Stimulus Prompting'. arXiv. Available at: https://doi.org/10.48550/arXiv.2302.11520.

Lim, J. and Song, H.-J. (2023) 'Improving Multi-Stage Long Document Summarization with Enhanced Coarse Summarizer', in Y. Dong et al. (eds) *Proceedings of the 4th New Frontiers in Summarization Workshop*. *NewSum 2023*, Singapore: Association for Computational Linguistics, pp. 135–144. Available at: https://doi.org/10.18653/v1/2023.newsum-1.13.

Lin, C. (2004) 'Looking for a Few Good Metrics : ROUGE and its Evaluation', *NTCIR Workshop*, (June), pp. 2–4.

Lin, C.-Y. (2004) 'ROUGE: A Package for Automatic Evaluation of Summaries', in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. Available at: https://aclanthology.org/W04-1013 (Accessed: 9 July 2024).

Liu, B., Hu, M. and Cheng, J. (2005) *Opinion observer: Analyzing and comparing opinions on the Web*. Available at: https://doi.org/10.1145/1060745.1060797.

Liu, F. and Liu, Y. (2008) 'Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries.', in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. *ACL 2008*, Columbus, OH: ACL, p. 204. Available at: https://doi.org/10.3115/1557690.1557747.

Liu, P.J. *et al.* (2018) 'Generating Wikipedia by Summarizing Long Sequences'. Available at: http://arxiv.org/abs/1801.10198.

Liu, X. *et al.* (2018) 'Stochastic Answer Networks for Machine Reading Comprehension', in I. Gurevych and Y. Miyao (eds) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2018*, Melbourne, Australia: Association for Computational Linguistics, pp. 1694–1704. Available at: https://doi.org/10.18653/v1/P18-1157.

Liu, Y. (2019) 'Fine-tune BERT for Extractive Summarization'. Available at: http://arxiv.org/abs/1903.10318.

Liu, Y. *et al.* (2019) 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. arXiv. Available at: https://doi.org/10.48550/arXiv.1907.11692.

Liu, Y. and Lapata, M. (2019a) 'Hierarchical Transformers for Multi-Document Summarization', in A. Korhonen, D. Traum, and L. Màrquez (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. *ACL 2019*, Florence, Italy: Association for Computational Linguistics, pp. 5070–5081. Available at: https://doi.org/10.18653/v1/P19-1500.

Liu, Y. and Lapata, M. (2019b) 'Text Summarization with Pretrained Encoders', in K. Inui et al. (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). *EMNLP-IJCNLP 2019*, Hong Kong, China: Association for Computational Linguistics, pp. 3730–3740. Available at: https://doi.org/10.18653/v1/D19-1387.

Liu, Y. and Liu, P. (2021) 'SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization', in C. Zong et al. (eds) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. *ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, pp. 1065–1072. Available at: https://doi.org/10.18653/v1/2021.acl-short.135.

Liu, Z. *et al.* (2021) 'Probing Across Time: What Does RoBERTa Know and When?', in *Findings of the Association for Computational Linguistics: EMNLP 2021*. *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 820–842. Available at: https://doi.org/10.18653/v1/2021.findings-emnlp.71.

Loper, E. and Bird, S. (2002) 'NLTK: the Natural Language Toolkit', *CoRR*, cs.CL/0205028. Available at: https://doi.org/10.3115/1118108.1118117.

Louis, A. and Nenkova, A. (2009) 'Automatically Evaluating Content Selection in Summarization without Human Models', in P. Koehn and R. Mihalcea (eds) *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2009*, Singapore: Association for Computational Linguistics, pp. 306–314. Available at: https://aclanthology.org/D09-1032 (Accessed: 18 July 2024).

Lu, Z. (2011) 'PubMed and beyond: a survey of web tools for searching biomedical literature', *Database: The Journal of Biological Databases and Curation*, 2011, p. baq036. Available at: https://doi.org/10.1093/database/baq036.

Luhn, H.P. (1958) 'The Automatic Creation of Literature Abstracts', *IBM Journal of research and development*, 2(2), pp. 159–165.

Lyu, L., Xu, X. and Wang, Q. (2020) 'Collaborative Fairness in Federated Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.2008.12161.

Ma, C. *et al.* (2021) 'Multi-document Summarization via Deep Learning Techniques: A Survey'. arXiv. Available at: http://arxiv.org/abs/2011.04843 (Accessed: 15 November 2023).

Ma, C. *et al.* (2024) 'Disentangling Specificity for Abstractive Multi-document Summarization'. arXiv. Available at: https://doi.org/10.48550/arXiv.2406.00005.

Mallick, C. *et al.* (2018) *Graph-Based Text Summarization Using Modified TextRank*, *Advances in Intelligent Systems and Computing*. Available at: https://doi.org/10.1007/978-981-13-0514-6_14.

Mani, I. (2001) *Automatic Summarization*, *nlp.3*. John Benjamins Publishing Company. Available at: https://benjamins.com/catalog/nlp.3 (Accessed: 9 July 2024).

Mani, I. and Bloedorn, E. (1997) 'Multi-document Summarization by Graph Search and Matching'. Available at: https://doi.org/10.3115/1119467.1119476.

Manning, C.D. (2015) 'Computational Linguistics and Deep Learning', *Computational Linguistics*, 41(4), pp. 701–707. Available at: https://doi.org/10.1162/COLI_a_00239.

Manning, C.D., Raghavan, P. and Schütze, H. (2008) *Introduction to Information Retrieval*, *Higher Education from Cambridge University Press*. Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511809071.

McKeown, K. *et al.* (2003) 'Columbia's Newsblaster: New Features and Future Directions', in *Companion Volume of the Proceedings of HLT-NAACL 2003 - Demonstrations*, pp. 15–16. Available at: https://aclanthology.org/N03-4008 (Accessed: 12 July 2024).

McKiernan, G. (2000) 'arXiv.org: The Los Alamos National Laboratory e-print server', *International Journal on Grey Literature*, 1, pp. 127–138. Available at: https://doi.org/10.1108/14666180010345564.

Mehrabi, N. *et al.* (2021) 'A Survey on Bias and Fairness in Machine Learning', *ACM Comput. Surv.*, 54(6), p. 115:1-115:35. Available at: https://doi.org/10.1145/3457607.

Meichanetzidis, K. *et al.* (2023) 'Grammar-aware sentence classification on quantum computers', *Quantum Machine Intelligence*, 5(1), p. 10. Available at: https://doi.org/10.1007/s42484-023-00097-1.

Menick, J. *et al.* (2022) 'Teaching language models to support answers with verified quotes'. arXiv. Available at: https://doi.org/10.48550/arXiv.2203.11147.

Mensh, B. and Kording, K. (2017) 'Ten simple rules for structuring papers', *PLOS Computational Biology*, 13(9), p. e1005619. Available at: https://doi.org/10.1371/journal.pcbi.1005619.

Mihalcea, R. and Tarau, P. (2004) 'TextRank: Bringing Order into Text', in D. Lin and D. Wu (eds) *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2004*, Barcelona, Spain: Association for Computational Linguistics, pp. 404–411. Available at: https://aclanthology.org/W04-3252 (Accessed: 17 July 2024).

Mikolov, T. *et al.* (2013) 'Efficient Estimation of Word Representations in Vector Space'. arXiv. Available at: https://doi.org/10.48550/arXiv.1301.3781.

Mochales, R. and Moens, M.-F. (2011) 'Argumentation mining', *Artificial Intelligence and Law*, 19(1), pp. 1–22. Available at: https://doi.org/10.1007/s10506-010-9104-x.

Mohtarami, M. *et al.* (2018) 'Automatic Stance Detection Using End-to-End Memory Networks', in M. Walker, H. Ji, and A. Stent (eds) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. *NAACL-HLT 2018*, New Orleans, Louisiana: Association for Computational Linguistics, pp. 767–776. Available at: https://doi.org/10.18653/v1/N18-1070.

Murdoch, W.J. *et al.* (2019) 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences*, 116(44), pp. 22071–22080. Available at: https://doi.org/10.1073/pnas.1900654116.

Nallapati, R. *et al.* (2016) 'Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond', *Proceedings of CoNLL*, pp. 280–290. Available at: http://arxiv.org/abs/1602.06023.

Nallapati, R., Zhai, F. and Zhou, B. (2016) 'SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents'.

Nanayakkara, P., Hullman, J. and Diakopoulos, N. (2021) 'Unpacking the Expressed Consequences of AI Research in Broader Impact Statements', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery (AIES '21), pp. 795–806. Available at: https://doi.org/10.1145/3461702.3462608.

Narayanan, A. and Shmatikov, V. (2008) 'Robust de-anonymization of large sparse datasets', in *Proceedings - 2008 IEEE Symposium on Security and Privacy, SP*. *2008 IEEE Symposium on Security and Privacy, SP*, pp. 111–125. Available at: https://doi.org/10.1109/SP.2008.33.

Nenkova, A. (2011) 'Automatic Summarization', *Foundations and Trends® in Information Retrieval*, 5(2), pp. 103–233. Available at: https://doi.org/10.1561/1500000015.

Nenkova, A. and McKeown, K. (2012) 'A survey of text summarisation techniques', in C.C. Aggarwal and C.X. Zhai (eds) *Mining Text Data*. Springer Science & Business Media, pp. 43–76. Available at: https://doi.org/10.1007/978-1-4614-3223-4.

Nenkova, A. and Passonneau, R. (2004) 'Evaluating content selection in summarization: The pyramid method', in. *Proceedings of HLT-NAACL*, pp. 145–152. Available at: papers2://publication/uuid/DC675E84-0A45-48B7-A26C-F08B4B9398D3.

Neumann, M. *et al.* (2019) 'ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing', in D. Demner-Fushman et al. (eds) *Proceedings of the 18th BioNLP Workshop and Shared Task*. *BioNLP 2019*, Florence, Italy: Association for Computational Linguistics, pp. 319–327. Available at: https://doi.org/10.18653/v1/W19-5034.

Nisbet, M.C. and Scheufele, D.A. (2009) 'What's next for science communication? Promising directions and lingering distractions', *American Journal of Botany*, 96(10), pp. 1767–1778. Available at: https://doi.org/10.3732/ajb.0900041.

Norman, G. (2010) 'Likert scales, levels of measurement and the "laws" of statistics', *Advances in Health Sciences Education*, 15(5), pp. 625–632. Available at: https://doi.org/10.1007/s10459-010-9222-y.

Nosek, B.A. *et al.* (2015) 'Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility', *Science*, 348(6242), pp. 1422–1425. Available at: https://doi.org/10.1126/science.aab2374.

O'Cathain, A., Murphy, E. and Nicholl, J. (2010) 'Three techniques for integrating data in mixed methods studies', *BMJ*, 341.

Oh, H., Nam, S. and Zhu, Y. (2022) 'Structured abstract summarization of scientific articles: Summarization using full-text section information', *Journal of the Association for Information Science and Technology*, 74(2), pp. 234–248. Available at: https://doi.org/10.1002/asi.24727.

Oliveira, S. *et al.* (2022) 'Improving Academic Writing with a Method for Text Revision Supported by Text Mining', *International Journal of Emerging Technologies in Learning (iJET)*, 17, pp. 150–163. Available at: https://doi.org/10.3991/ijet.v17i21.31249.

Olteanu, A. *et al.* (2019) 'Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries', *Frontiers in Big Data*, 2. Available at: https://doi.org/10.3389/fdata.2019.00013.

OpenAI *et al.* (2024) 'GPT-4 Technical Report'. arXiv. Available at: https://doi.org/10.48550/arXiv.2303.08774.

Ott, M. *et al.* (2019) ‘fairseq: A Fast, Extensible Toolkit for Sequence Modeling’, in W. Ammar, A. Louis, and N. Mostafazadeh (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 48–53. Available at: https://doi.org/10.18653/v1/N19-4009.

Otterbacher, J., Erkan, G. and Radev, D. (2005) ‘Using Random Walks for Question-focused Sentence Retrieval’, in R. Mooney et al. (eds) *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. *HLT-EMNLP 2005*, Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 915–922. Available at: https://aclanthology.org/H05-1115 (Accessed: 18 July 2024).

Papineni, K. *et al.* (2002) ‘Bleu: a Method for Automatic Evaluation of Machine Translation’, in P. Isabelle, E. Charniak, and D. Lin (eds) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. *ACL 2002*, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. Available at: https://doi.org/10.3115/1073083.1073135.

Paszke, A. *et al.* (2019) ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’. arXiv. Available at: https://doi.org/10.48550/arXiv.1912.01703.

Pennington, J., Socher, R. and Manning, C. (2014) ‘Glove: Global Vectors for Word Representation’, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [Preprint]. Available at: https://doi.org/10.3115/v1/D14-1162.

Pérez-Llantada, C., Plo, R. and Ferguson, G.R. (2011) ‘“You don’t say what you know, only what you can”: The perceptions and practices of senior Spanish academics regarding research dissemination in English’, *English for Specific Purposes*, 30(1), pp. 18–30. Available at: https://doi.org/10.1016/j.esp.2010.05.001.

Pryzant, R. *et al.* (2020) ‘Automatically Neutralizing Subjective Bias in Text’, *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01), pp. 480–489. Available at: https://doi.org/10.1609/aaai.v34i01.5385.

Punjani, D. and Tsalapati, E. (2023) ‘Question Answering Engines for Geospatial Knowledge Graphs’, in *Geospatial Data Science*, pp. 257–282. Available at: https://doi.org/10.1145/3581906.3581922.

Qazvinian, V. (2010) ‘Citation Summarization Through Keyphrase Extraction’, in. *COLING 2010, International Conference on Computational Linguistics*, Beijing, China, pp. 895–903. Available at: http://www.aclweb.org/anthology/C10-1101.

Qiu, X. *et al.* (2024) ‘Progressive Multi-modal Conditional Prompt Tuning’. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.11864.

Radev, D., Hovy, E. and McKeown, K. (2002) ‘Introduction to the special issue on summarization’, *Computational linguistics*, 28(4), pp. 399-408-399–408. Available at: https://doi.org/10.1016/j.jbi.2011.03.008.

Radev, D.R. *et al.* (2004) ‘Centroid-based summarization of multiple documents’, *Information Processing and Management*, (40), pp. 919–938. Available at: https://doi.org/10.1016/j.ipm.2003.10.006.

Raffel, C. *et al.* (2020) 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', *Journal of Machine Learning Research*, 21(140), pp. 1–67. Available at: http://jmlr.org/papers/v21/20-074.html (Accessed: 11 July 2024).

Rai, A. *et al.* (2021) 'Query Specific Focused Summarization of Biomedical Journal Articles', in *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*. *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 91–100. Available at: https://doi.org/10.15439/2021F128.

Reimers, N. and Gurevych, I. (2019) 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. arXiv. Available at: https://doi.org/10.48550/arXiv.1908.10084.

Reis, C.P. (2021) 'The latest developments in the area of therapeutic delivery excluding some diseases, such as COVID-19 and the big three (HIV/AID, malaria and tuberculosis)', *Therapeutic Delivery*, 12(12), pp. 799–805. Available at: https://doi.org/10.4155/tde-2021-0066.

Reiter, E. (2018) 'A Structured Review of the Validity of BLEU', *Computational Linguistics*, 44(3), pp. 393–401. Available at: https://doi.org/10.1162/coli_a_00322.

Roche, A.J. and Rickard, L.N. (2017) 'Cocitation or Capacity-Building? Defining Success within an Interdisciplinary, Sustainability Science Team', *Frontiers in Communication*, 2, p. 13. Available at: https://doi.org/10.3389/fcomm.2017.00013.

Rong, Y. *et al.* (2023) 'Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations'. arXiv. Available at: https://doi.org/10.48550/arXiv.2210.11584.

Rossiello, G., Basile, P. and Semeraro, G. (2017) 'Centroid-based Text Summarization through Compositionality of Word Embeddings', in G. Giannakopoulos et al. (eds) *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. *MultiLing 2017*, Valencia, Spain: Association for Computational Linguistics, pp. 12–21. Available at: https://doi.org/10.18653/v1/W17-1003.

Rush, A.M., Chopra, S. and Weston, J. (2015) 'A Neural Attention Model for Abstractive Sentence Summarization', in. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 379–389. Available at: https://doi.org/10.1162/153244303322533223.

Sanchez-Velazquez, O. and Sierra, G. (2016) 'Let's Agree to Disagree: Measuring Agreement between Annotators for Opinion Mining Task', *Research on computing science*, 110(1), pp. 9–19. Available at: https://doi.org/10.13053/RCS-110-1-1.

See, A., Liu, P.J. and Manning, C.D. (2017) 'Get To The Point: Summarization with Pointer-Generator Networks'. Available at: https://doi.org/10.18653/v1/P17-1099.

Sergeev, A. and Del Balso, M. (2018) 'Horovod: fast and easy distributed deep learning in TensorFlow'. arXiv. Available at: https://doi.org/10.48550/arXiv.1802.05799.

Shahaf, D., Guestrin, C. and Horvitz, E. (2012) 'Metro maps of science', in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. *KDD '12: The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing China: ACM, pp. 1122–1130. Available at: https://doi.org/10.1145/2339530.2339706.

Song, K. *et al.* (2019) 'Controlling the Amount of Verbatim Copying in Abstractive Summarization'. arXiv. Available at: http://arxiv.org/abs/1911.10390 (Accessed: 22 November 2023).

Song, W. *et al.* (2024) 'Hierarchical Context Merging: Better Long Context Understanding for Pre-trained LLMs'. arXiv. Available at: https://doi.org/10.48550/arXiv.2404.10308.

Sparck Jones, K. (1972) 'A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL', *Journal of Documentation*, 28(1), pp. 11–21. Available at: https://doi.org/10.1108/eb026526.

Spärck Jones, K. (2007) 'Automatic summarising: The state of the art', *Information Processing & Management*, 43(6), pp. 1449–1481. Available at: https://doi.org/10.1016/j.ipm.2007.03.009.

Spasić, I. *et al.* (2014) 'Text mining of cancer-related information: Review of current status and future directions', *International Journal of Medical Informatics*, 83(9), pp. 605–623. Available at: https://doi.org/10.1016/j.ijmedinf.2014.06.009.

Srivastava, A. and Sutton, C. (2017) 'Autoencoding Variational Inference For Topic Models'. arXiv. Available at: https://doi.org/10.48550/arXiv.1703.01488.

Strubell, E., Ganesh, A. and McCallum, A. (2019) 'Energy and Policy Considerations for Deep Learning in NLP', in A. Korhonen, D. Traum, and L. Màrquez (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. *ACL 2019*, Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. Available at: https://doi.org/10.18653/v1/P19-1355.

Subramanian, S. *et al.* (2020) 'On Extractive and Abstractive Neural Document Summarization with Transformer Language Models'. arXiv. Available at: http://arxiv.org/abs/1909.03186 (Accessed: 22 November 2023).

Susskind, R. (2010) *The End of Lawyers?: Rethinking the nature of legal services*. Revised Edition, Revised Edition. Oxford, New York: Oxford University Press.

Sweller, J. (1988) 'Cognitive load during problem solving: Effects on learning', *Cognitive Science*, 12(2), pp. 257–285. Available at: https://doi.org/10.1207/s15516709cog1202_4.

Tan, J., Wan, X. and Xiao, J. (2017) 'Abstractive Document Summarization with a Graph-Based Attentional Neural Model', *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Preprint]. Available at: https://doi.org/10.18653/v1/P17-1108.

Teufel, S. and Moens, M. (2002) 'Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status', *Computational Linguistics*, 28(4), pp. 409–445. Available at: https://doi.org/10.1162/089120102762671936.

Touvron, H. *et al.* (2023) 'Llama 2: Open Foundation and Fine-Tuned Chat Models'. arXiv. Available at: http://arxiv.org/abs/2307.09288 (Accessed: 21 November 2023).

Trewartha, A. *et al.* (2020) 'COVIDScholar: An automated COVID-19 research aggregation and analysis platform'. arXiv. Available at: http://arxiv.org/abs/2012.03891 (Accessed: 15 November 2023).

Trewin, S. *et al.* (2019) 'Considerations for AI fairness for people with disabilities', *AI Matters*, 5(3), pp. 40–63. Available at: https://doi.org/10.1145/3362077.3362086.

Vaswani, A. *et al.* (2017) 'Attention Is All You Need'. Available at: http://arxiv.org/abs/1706.03762.

Venkataramana, A., Srividya, K. and Cristin, R. (2022) 'Abstractive Text Summarization Using BART', in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*. *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/MysuruCon55714.2022.9972639.

Verma, R. and Lee, D. (2018) 'Extractive Summarization: Limits, Compression, Generalized Model and Heuristics', *Computación y Sistemas*, 21(4). Available at: https://doi.org/10.13053/cys-21-4-2855.

Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods*, 17(3), pp. 261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2.

Wan, X. (2008) 'An exploration of document impact on graph-based multi-document summarization', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. USA: Association for Computational Linguistics (EMNLP '08), pp. 755–762.

Wan, X. (2013) 'SUBTOPIC-BASED MULTIMODALITY RANKING FOR TOPIC-FOCUSED MULTIDOCUMENT SUMMARIZATION', *Computational Intelligence*, 29(4), pp. 627–648. Available at: https://doi.org/10.1111/j.1467-8640.2012.00435.x.

Wang, A. *et al.* (2020) 'SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.1905.00537.

Wang, C. and Feng, F. (2022) 'ERNIE based intelligent triage system', *Journal of Intelligent & Fuzzy Systems*, 43(4), pp. 5013–5022. Available at: https://doi.org/10.3233/JIFS-212140.

Wang, D. *et al.* (2020) 'Heterogeneous Graph Neural Networks for Extractive Document Summarization', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 6209–6219. Available at: https://doi.org/10.18653/v1/2020.acl-main.553.

Wang, H. *et al.* (2023) 'Scientific discovery in the age of artificial intelligence', *Nature*, 620(7972), pp. 47–60. Available at: https://doi.org/10.1038/s41586-023-06221-2.

Wang, J. *et al.* (2023) 'Zero-Shot Cross-Lingual Summarization via Large Language Models', in Y. Dong et al. (eds) *Proceedings of the 4th New Frontiers in Summarization Workshop*. *NewSum 2023*, Singapore: Association for Computational Linguistics, pp. 12–23. Available at: https://doi.org/10.18653/v1/2023.newsum-1.2.

Wang, L. *et al.* (2016) 'A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization'.

Wang, L.L. *et al.* (2020) 'CORD-19: The COVID-19 Open Research Dataset', in K. Verspoor et al. (eds) *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. *NLP-COVID19 2020*, Online: Association for Computational Linguistics. Available at: https://aclanthology.org/2020.nlpcovid19-acl.1 (Accessed: 10 July 2024).

Wang, L.L. *et al.* (2023) 'Automated Metrics for Medical Multi-Document Summarization Disagree with Human Evaluations', in A. Rogers, J. Boyd-Graber, and N. Okazaki (eds) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2023*, Toronto, Canada: Association for Computational Linguistics, pp. 9871–9889. Available at: https://doi.org/10.18653/v1/2023.acl-long.549.

Wang, P. *et al.* (2023) 'Plan and generate: Explicit and implicit variational augmentation for multi-document summarization of scientific articles', *Information Processing & Management*, 60(4), p. 103409. Available at: https://doi.org/10.1016/j.ipm.2023.103409.

White, W.J. (2001) 'A Communication Model of Conceptual Innovation in Science', *Communication Theory*, 11(3), pp. 290–314. Available at: https://doi.org/10.1111/j.1468-2885.2001.tb00244.x.

Wolf, T. *et al.* (2020) 'Transformers: State-of-the-Art Natural Language Processing', in Q. Liu and D. Schlangen (eds) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. Available at: https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Wu, Y. *et al.* (2016) 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'. arXiv. Available at: https://doi.org/10.48550/arXiv.1609.08144.

Xiao, W. *et al.* (2022) 'PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization', in S. Muresan, P. Nakov, and A. Villavicencio (eds) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, pp. 5245–5263. Available at: https://doi.org/10.18653/v1/2022.acl-long.360.

Xiong, C. *et al.* (2018) 'Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling', in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 575–584. Available at: https://doi.org/10.1145/3209978.3209982.

Xu, J. *et al.* (2020) 'Discourse-Aware Neural Extractive Text Summarization', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 5021–5031. Available at: https://doi.org/10.18653/v1/2020.acl-main.451.

Xu, L. *et al.* (2023) 'Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment'. arXiv. Available at: https://doi.org/10.48550/arXiv.2312.12148.

Yang, A. *et al.* (2018) 'Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task', in E. Choi et al. (eds) *Proceedings of the Workshop on Machine Reading for Question Answering*. *ACL 2018*, Melbourne, Australia: Association for Computational Linguistics, pp. 98–104. Available at: https://doi.org/10.18653/v1/W18-2611.

Yasunaga, M. *et al.* (2017) 'Graph-based Neural Multi-Document Summarization', in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Vancouver, Canada: Association for Computational Linguistics, pp. 452–462. Available at: https://doi.org/10.18653/v1/K17-1045.

Yasunaga, M. *et al.* (2019) 'ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks', *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), pp. 7386–7393. Available at: https://doi.org/10.1609/aaai.v33i01.33017386.

Zaheer, M. *et al.* (2021) 'Big Bird: Transformers for Longer Sequences'. arXiv. Available at: http://arxiv.org/abs/2007.14062 (Accessed: 24 July 2023).

van der Zee, T. *et al.* (2017) 'Effects of subtitles, complexity, and language proficiency on learning from online education videos', *Journal of Media Psychology: Theories, Methods, and Applications*, 29(1), pp. 18–30. Available at: https://doi.org/10.1027/1864-1105/a000208.

Zhang, H., Xu, J. and Wang, J. (2019) 'Pretraining-Based Natural Language Generation for Text Summarization'. Available at: http://arxiv.org/abs/1902.09243.

Zhang, J. *et al.* (2020) 'PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization', in *Proceedings of the 37th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, pp. 11328–11339. Available at: https://proceedings.mlr.press/v119/zhang20ae.html (Accessed: 11 July 2024).

Zhang, R. *et al.* (2013) 'Automatic Twitter Topic Summarization With Speech Acts', *Audio, Speech, and Language Processing, IEEE Transactions on*, 21, pp. 649–658. Available at: https://doi.org/10.1109/TASL.2012.2229984.

Zhang, T. *et al.* (2020) 'BERTScore: Evaluating Text Generation with BERT'. arXiv. Available at: https://doi.org/10.48550/arXiv.1904.09675.

Zhang, X. *et al.* (2022) 'GreaseLM: Graph REASoning Enhanced Language Models for Question Answering'. arXiv. Available at: https://doi.org/10.48550/arXiv.2201.08860.

Zhao, J., Yang, L. and Cai, X. (2022) 'HetTreeSum: A Heterogeneous Tree Structure-based Extractive Summarization Model for Scientific Papers', *Expert Systems with Applications*, 210, p. 118335. Available at: https://doi.org/10.1016/j.eswa.2022.118335.

Zhao, X., Wang, T. and Rios, A. (2024) 'Improving Expert Radiology Report Summarization by Prompting Large Language Models with a Layperson Summary'. arXiv. Available at: https://doi.org/10.48550/arXiv.2406.14500.

Zheng, L. *et al.* (2023) 'Judging LLM-as-a-judge with MT-Bench and Chatbot Arena'. arXiv. Available at: http://arxiv.org/abs/2306.05685 (Accessed: 24 July 2023).

Zhong, L. *et al.* (2019) 'Automatic summarization of legal decisions using iterative masking of predictive sentences: 17th International Conference on Artificial Intelligence and Law, ICAIL 2019', *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019*, pp. 163–172. Available at: https://doi.org/10.1145/3322640.3326728.

Zhong, M. *et al.* (2020) 'Extractive Summarization as Text Matching', in D. Jurafsky et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 6197–6208. Available at: https://doi.org/10.18653/v1/2020.acl-main.552.

Zhou, J. *et al.* (2020) 'Graph neural networks: A review of methods and applications', *AI Open*, 1, pp. 57–81. Available at: https://doi.org/10.1016/j.aiopen.2021.01.001.