STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, Month 2025, pp 2529–2559.
Published online in International Academic Press (www.IAPress.org)

From Extraction to Reasoning: A Systematic Review of Algorithms in Multi-Document Summarization and QA

Emmanuel T. Efosa-Zuwa*, Olufunke Oladipupo, Jelili Oyelade

Department of Computer and Information Sciences Covenant University,Ota, Nigeria

Abstract Multi-document summarization and question-answering (QA) have become pivotal tasks in Natural Language Processing (NLP), facilitating information extraction and decision-making across various domains. This systematic review explores the evolution of algorithms used in these tasks, providing a comprehensive taxonomy of traditional, modern, and emerging approaches. We examine the progression from early extractive methods such as TF-IDF and TextRank, to the advent of neural models like BERT, GPT, and T5, and the integration of retrieval-augmented generation (RAG) for QA. Hybrid models combining traditional techniques with neural approaches and graph-based methods are also discussed. Through a detailed analysis of algorithmic frameworks, we identify key strengths, weaknesses, and challenges in current methodologies. Additionally, the review highlights recent trends such as unified models, multimodal algorithms, and the application of reinforcement learning in summarization and QA tasks. We also explore the real-world relevance of these algorithms in sectors such as news, legal, medical, and education. The paper concludes by outlining open research directions, proposing new evaluation frameworks, and emphasizing the need for cross-task annotations and ethical considerations in future algorithmic development.

Keywords Mult document; Summarization; Question Answering; Language Models.

DOI: 10.19139/soic-2310-5070-2398

1. Introduction

In an age of rapid and exponential creation of textual data, the need for advanced techniques for extracting, processing and reasoning over information cannot be overemphasized. Of the core natural language processing (NLP) tasks multi document summarization, and question answering (QA) have become indispensable in both research and application [1][2]. In the former, information from multiple textual sources is synthesized into a coherent and concise summary to aid users gathering much needed insights from very large documents [3]. Similarly, QA systems seek to deliver accurate, context appropriate answers to user posed questions, constituting as a vital information retrieval tool in applications ranging from health care to education as well as business intelligence [4]. These tasks work together to provide the backbone of many of today's AI driven systems that are going to help us augment human decision making and help us get access to information.

This domain ranges from traditional rule based and statistical approaches to modern deep learning and hybrid approaches. There are a bunch of information extraction and summarization algorithms that are based on, for example, graph-based methods or latent semantic analysis and helped people nail the problem of returning a unified answer from multiple heterogeneous sources [5] [6]. Although this worked well, these methods, more often than not, had difficulty in capturing semantic nuances, or reasoning over complicated relationships present in the given text. With the advent of deep learning and transformer-based architectures (BERT, GPT and T5) models are now able to contextualize, summarize and reason over textual data with tremendous accuracy [7]. The latest trends in this area are emerging, like hybrid approaches that integrate both symbolic reasoning with neural networks and the creation of Explainable AI (XAI) models that help predict and reason about possible futures [8].

ISSN 2310-5070 (online) ISSN 2311-004X (print)

However, the domain has its challenges. One major challenge in multi document summarization is to ensure that the produced summary is coherent and non-redundant across a variety of sources that differ in style, structure and content [9]. Challenges such as handling ambiguous or incomplete queries, reasoning over multiple passages to find accurate answers, and making sure the responses are always resilient to adversarial inputs are present in QA systems [10]. Moreover, both tasks suffer from the scalability problem when handling big data and demand for domain adaptation to serve specific domains such as law, medicine.

By innovative techniques, many of these challenges have been recently overcome. In recent years, Transformer based models have introduced self-attention mechanisms that enable better treatment of long sequences of text and in general provide better contextual understanding and reasoning [11]. In order to generate summaries with higher coherence and higher informativeness than existing retrieval–based methods, reinforcement learning and reward–based optimization have been applied [12]. Multi hop reasoning Models and retrieval Augmented Generation (RAG) Frameworks have improved ability to reason over disperse and complex information for QA [13] [14]. On top of this, domain specific fine tuning and few shot learning strategies have further propelled these systems to learn for specialized applications with very little labeled data.

The objectives of this systematic review are as follows:

- To provide a taxonomy of algorithms used in multi-document summarization and QA.
- To highlight trends in the development and application of these algorithms.
- To discuss the challenges associated with existing approaches.

This review aims to explore the progression from extraction to reasoning in multi-document summarization and QA, focusing on the algorithms that drive these tasks. By analyzing the evolution of techniques, identifying persistent challenges, and highlighting recent trends, this review seeks to provide a comprehensive understanding of the state of the art and propose directions for future research.

2. Background and Evolution

Over the years, the fields of multi-document summarization and question answering (QA) have been transformed from simple extraction approaches to detecting complex reasoning strategies [1] [2]. This evolution is due to an increase of task complexity and demand for systems that are able to process broad and large volume textual data.

2.1. Early Techniques in Summarization

In the cases of summarization, early approaches tended to be extractive, with a focus on performing the task of identifying the sentences or phrases containing the most responsible information from the source documents [15]. Early research was dominated by techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), lexical chains, or graph-based ranking methods, e.g. TextRank [16] [17]. Each of these methods used statistical properties of the text, like word frequency and sentence similarity, as the key means of information identification. These techniques proved good at creating summaries with the same words as the original but their ability to provide coherency and retain the semantic relationships among sentences was diminished. Since extractive methods have their limits, abstractive summarization moved in as an alternative, by generating summaries by rephrasing and synthesizing of information all by way of mimicry in the human-like summarization style [18].

2.2. Evolution of Question-Answering Systems

The evolution of QA systems from more rule-based approaches to advanced neural architectures is observed, while presenting different types of pitfalls and suggestions [16]. Early QA systems were restricted to selecting from the resources of structured databases or predefined templates by using handcrafted rules and ontologies [19]. These were in domain specific and unsuitable for processing unstructured data or ambiguous queries. But it was the advent of deep learning that really changed QA by teaching statistical methods how to generalize better across domains, using statistical methods that the machine learning pioneers had introduced before [20]. Systems that use

sequence to sequence architectures together with attention mechanisms based on neural networks could understand and draw logical conclusions from raw text [21]. Even more, these models build upon transformer-based models, such as BERT, RoBERTa, and GPT, which have achieved state of the art results in many downstream applications, providing powerful contextual representation, multi hop reasoning ability, and better handling of complex queries [7].

2.3. Role of Large-Scale Datasets in Algorithm Development

Both summarization and QA have had to develop large-scale, high-quality datasets, to further their advancements. Most neural summarization datasets (e.g. MultiNews and CNN/Daily Mail) have provided diverse, multi-document contexts to train and evaluate the modern algorithms. Then datasets similar to SQuAD (Stanford Question Answering Dataset), Natural Questions, and HotpotQA have seen other benchmarks be set in the QA space such that different models that are very good at comprehension and reasoning can be made [22] [23]. Not only have these datasets been used to train through, but they have also become a foundation for shared tasks/leaderboards and a breeding ground for friendly competition and innovation.

3. Algorithmic Frameworks

Various algorithmic approaches has been tried for multi-document summarization and for question answering (QA) as shown in Figure 1. Broadly speaking, these may be classified as traditional methods, e.g. techniques that are statistical, rule based, or graph based, or as more recent machine-learning, or deep-learning ones [16]. In this section, we will discuss traditional techniques more detailed, such as extractive summarization methods, rule-based QA systems, and graph-based techniques for reasoning and contextual representation.

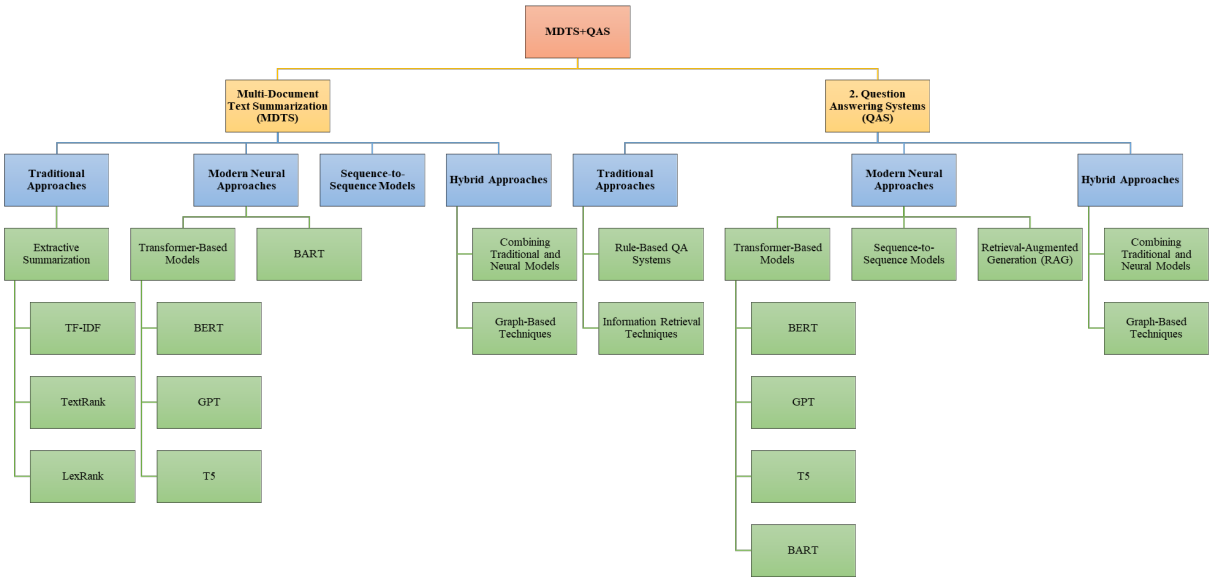

Figure 1. Taxonomy of the MDTS and QAS Frameworks.

3.1. Extractive Summarization Techniques

Extractive summarization focuses on selecting the most relevant sentences or passages from a set of documents to create a concise summary as shown in Figure 2. A widely used method is TF-IDF (Term Frequency-Inverse Document Frequency), which evaluates a word's importance by combining its frequency in a document (TF) with its rarity across the corpus (IDF). This approach identifies terms that are frequent within a document but uncommon across the dataset, making them more informative. Term-based methods, often employing the Bag-of-Words (BOW) model with TF-IDF or its variations, are foundational in extractive summarization.

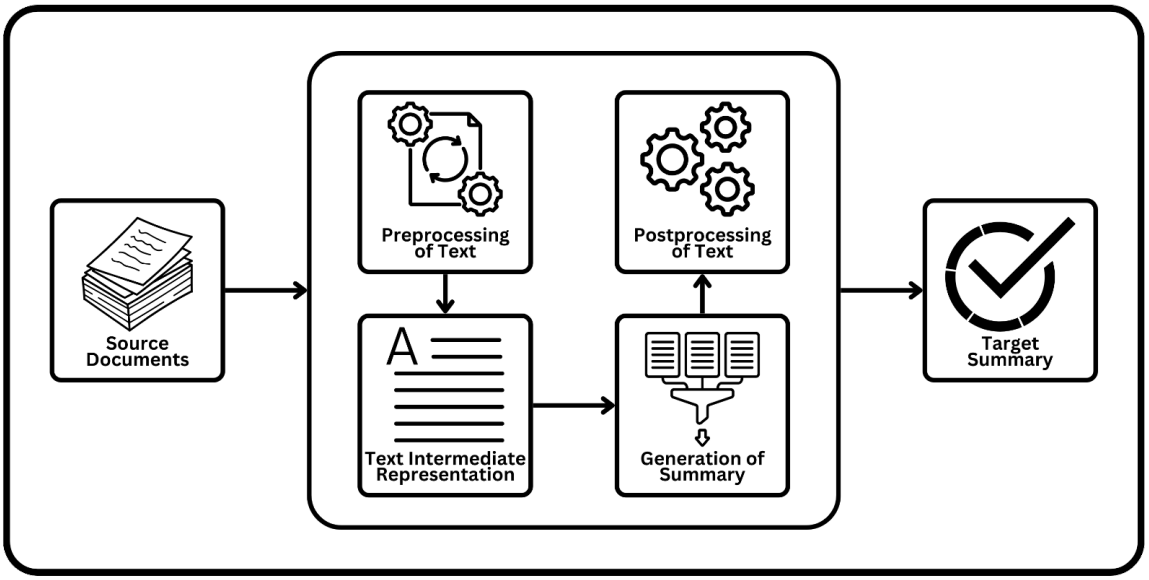

Figure 2. General Architecture of Extractive Summarization Techniques.

Advanced methods like Maximum Coverage and Less Redundancy (MCLR) [24] address optimization challenges by formulating multi-document summarization as a quadratic boolean programming problem, balancing content coverage and redundancy [25]. Additionally, a bottom-up strategy proposed by [26] organizes sentences using criteria such as chronology, topical relevance, precedence, and sequence to improve coherence and ordering [27].

TF-IDF assigns a higher score to words that are frequent in a specific document but rare across the corpus, allowing the summarization model to prioritize sentences containing such words. While effective for identifying important terms, TF-IDF-based methods are limited by their inability to capture the semantic meaning and interrelations of sentences in the document.

Another notable technique is TextRank, a graph-based approach similar to the PageRank algorithm used by search engines. TextRank builds a graph where each sentence in the document is a node, and edges between nodes represent the similarity between sentences. The graph is iteratively ranked, and the sentences with the highest scores are selected for the summary. The similarity between two sentences is typically calculated using cosine similarity on the word embeddings or term frequency vectors.

TextRank operates as follows: construct a graph of sentences, calculate sentence similarity scores using cosine similarity, rank the sentences using the PageRank algorithm, and select the top-ranked sentences for the summary. Similarly, LexRank is another graph-based method, but it uses the cosine similarity of sentence vectors for creating sentence clusters. It builds a weighted graph where edges represent the similarity between sentences, and sentences are ranked based on centrality within the graph. Mihalcea and Tarau [17] were the first to introduce TextRank, demonstrating its effectiveness in unsupervised extractive summarization tasks by leveraging graph-based ranking for identifying key sentences. Erkan and Radev [28] extended this concept with LexRank, which incorporates the

idea of eigenvector centrality to identify the most informative sentences in a graph. [29] further improved these methods by integrating global and local information into the ranking process, enhancing the quality of summaries. These advancements underscore the versatility of graph-based methods in text summarization.

3.2. Rule-based QA Systems and Information Retrieval Techniques

Early on QA systems were largely rule based, in that system was bound to predefined rules and templates that allowed the question and its answer from structured or semi structured data as shown in Figure 3. These would be natural language query processing systems, with a set of heuristics (or logic rules) that map query ambiguous queries into sensible responses. The drawback of the rule-based systems is that they cannot handle uncertain or complex queries that cannot match predefined templates. But in practice, they commonly use manually constructed knowledge bases or ontologies that take too long to build and are too inflexible for dynamic, unstructured data. However, several studies[30][31] [32] [33] primarily adopted traditional approaches like monolithic architecture, morphological analyzers, and binary expression for question answering. Techniques such as JAVELIN [32] and Passage retrieval methods [34] were explored in 2007 and 2008, showcasing the foundational exploration of rule-based and information retrieval-based methods in the field

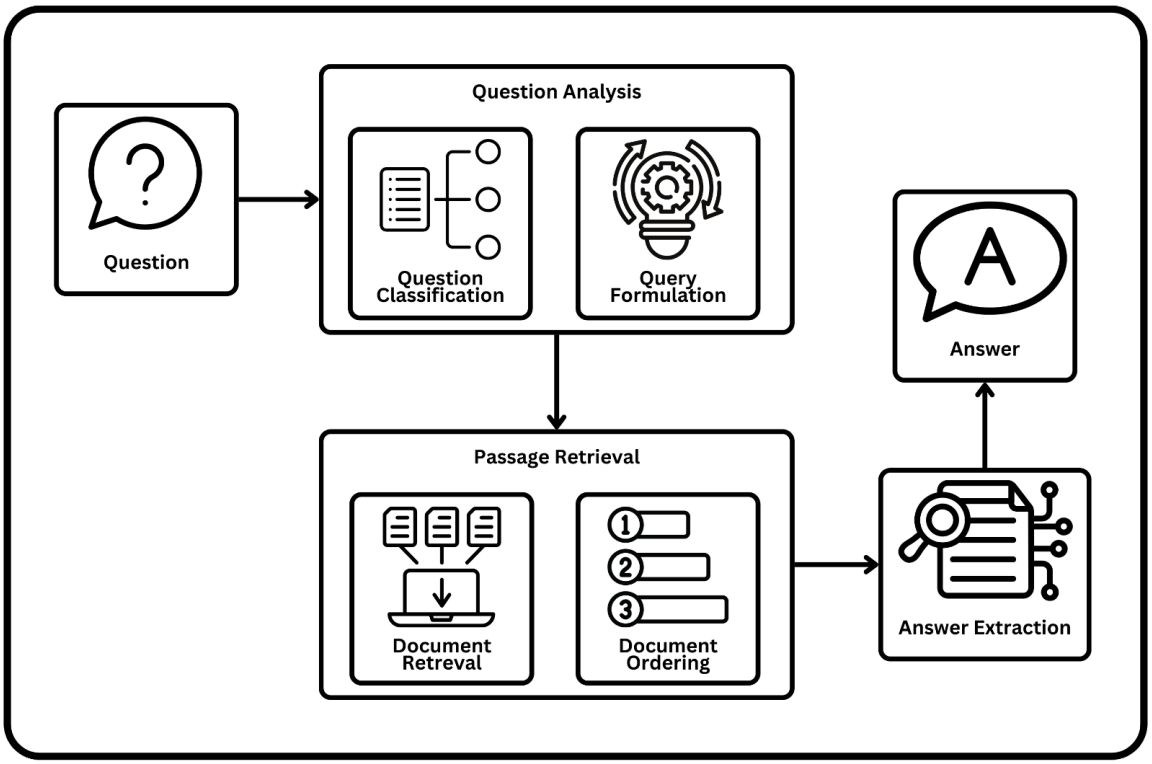

Figure 3. General Architecture of Rule Based Techniques.

Information retrieval (IR) is a key component of rule-based QA systems, conducting a search of a large corpus for a collection of relevant documents or passages to answer a user query. The typical workflow in an IR-based QA system is as follows:

1. Parse the query to understand the intent and entities.
2. Retrieve candidate documents or passages based on keyword matching or semantic similarity.
3. Extract the most relevant information from the retrieved texts to generate an answer.

In simple rule-based IR systems, techniques such Boolean retrieval and vector space model were frequently used such that the query is exactly matched to the documents based on the presence of keywords or the documents and the query is represented as the vector in multidimensional space respectively. However, most of these methods do not capture deeper content meaning or relationships between entities.

3.3. Graph-based Techniques for Reasoning and Contextual Representation

The prominent methods in summarization and QA since they help represent various relationships between entities &concepts using graphs. Hence, multi-hop QA tasks require reasoning over multiple information sources and graph-based reasoning turns out to be most reliable. Text can also be put into a form of a graph in which nodes are entities, sentences or concepts whereas edges represent relations or similarities between the nodes. It contains realizations of techniques for traversal of these graphs, for instance, breadth-first search (BFS), depth-first search (DFS), to answer queries or for creating summaries. Some works in this area include; [35], which put forward LexRank for extractive summarization, where the idea is based on the eigenvector centrality of the sentences in the similarity graph, and [28], were they introduced LexRank with special focus on the computation of the relevance values of the sentences.

However, [17] used TextRank that uses the graph ranking methodology for text processing such as text summarization. These pioneering studies provided a basis for using graph structures to model document relationships efficiently and make summaries comprehensive of various content from many documents. These methods produce accurate and clear multi-document summaries as they postulate that related sentences are connected in the graph.

3.4. Modern Neural Approaches

Around the advent of neural networks, most notably deep learning models, the landscape of multi document summarization and question answering (QA) has evolved. These approaches use large scale datasets and sophisticated architectures to obtain extraordinary performance on confronting complex language tasks. In this section, we take a closer look at some of the most modern neural approaches: Transformer based models, sequence to sequence models, and retrieval-augmented generation (RAG).

3.4.1 Transformer-Based Models

Natural language processing (NLP) saw a revolution with transformer-based models that are able to capture contexts and model long range dependencies. These models are built around the block of the of self-attention, which computes over all tokens in a sequence at the same time and hence can be efficiently parallelized and be more robust at getting contextual information.

- BERT (Bidirectional Encoder Representations from Transformers):

BERT is a bidirectional transformer model pre trained on large corpus with masked language modeling and next sentence prediction tasks. It also does extremely well on contextual understanding, and it takes into account what has come before as well as what comes after a sentence. BERT fine –tunes on datasets like SQuAD for QA to extract specific answers from a passage. For example, Japanese how-to tip QA [36] dataset was developed to address the scarcity of data in non-factoid QA, with potential applications to other languages. BERT [37] [38] [39] has shown remarkable improvements in various tasks such as passage re-ranking, document retrieval, and passage retrieval. However, it is important to acknowledge its limitations, including input length restrictions and reliance on weak supervision. Furthermore, the combination of BERT with traditional models like BM25 [10], has provided both precise answers and paragraph-sized summaries for different types of questions, emphasizing the need for further investigation into incorporating deep learning and enhancing entity relationships. MD-NFQA [40] addressed a crucial resource gap in Multilingual Non-Factoid QA, although it had limitations in capturing annotator behavior due to constraints in the evaluation interface.

- GPT (Generative Pre-trained Transformer):

GPT is a unidirectional transformer model optimized for text generation. It predicts the next token in a sequence, making it particularly effective for abstractive summarization and generative QA. Unlike BERT, GPT processes input text in a left-to-right manner, focusing on autoregressive modeling. Its architecture allows it to generate coherent and contextually relevant text, making it a preferred choice for tasks requiring natural language generation. Radford et al. [41] introduced GPT, showcasing its ability to generate human-like text by leveraging unsupervised pretraining on large corpora. Brown et al. [42] further expanded on this with GPT-3, demonstrating its remarkable capacity for few-shot learning and its effectiveness across a wide range of language tasks, including summarization and QA. These contributions underscore GPT's transformative impact on natural language processing.

- T5 (Text-to-Text Transfer Transformer):

T5 formulates all input/output in text-to-text format by converting all inputs and outputs into string of text. For summarization and QA, it is evident that T5 outperforms all the model by transforming them into sequence-to-sequence problems that provide the input documents and questions and get the output summaries and answers, respectively. For instance, Raffel et al.[43] proposed T5 that it is a text-to-text transfer transformer that can be applied to a wide variety of benchmarks. Khashabi et al. [44] investigated its applicability to open-domain QA and demonstrated how, depending on those datasets, a PTM fine-tuned model is the most effective. Further, Lewis et al. [14] put forward another strategy called retrieval-augmented generation (RAG) that combines retrievers and transformer models such as T5 for QA tasks need factual contents from outside world.

- BART (Bidirectional and Auto-Regressive Transformers):

BART combines the strengths of BERT and GPT, employing a denoising autoencoder structure. It is pre-trained by corrupting text and learning to reconstruct it, making it robust for abstractive summarization and QA. BART's encoder-decoder architecture enables it to process input text bidirectionally while generating output text autoregressively. For example, [45] presents an extract-then-abstract Transformer framework for Multi-Document Summarization (MDTS). It uses pre-trained language models such as BART to create a hierarchical extractor and abstractor for salient sentence selection and content rewriting.

3.4.2 Sequence-to-Sequence Models for Summarization and QA

Seq2Seq models form the backbone of many summarization and QA systems, leveraging an encoder for input sequence processing and a decoder for output generation. These models often integrate attention mechanisms, enabling the decoder to focus on the most relevant parts of the input during generation as shown in Figure 4. Early Seq2Seq architectures, such as LSTM-based models, achieved notable success in summarization tasks. However, modern transformers like BART and T5 have surpassed them, offering superior handling of long and complex contexts. In QA, Seq2Seq models combined with reranking are employed to generate abstractive answers, even when the exact answer is not explicitly stated in the text. For instance, See et al. [46] enhanced Seq2Seq summarization by introducing pointer-generator networks to address out-of-vocabulary words and improve factual consistency. Similarly, Raffel et al. [43] demonstrated the versatility of T5 by reframing NLP tasks, including summarization and QA, into a unified text-to-text format, achieving state-of-the-art results across multiple benchmarks. These advancements underscore the evolution and adaptability of Seq2Seq models in NLP tasks.

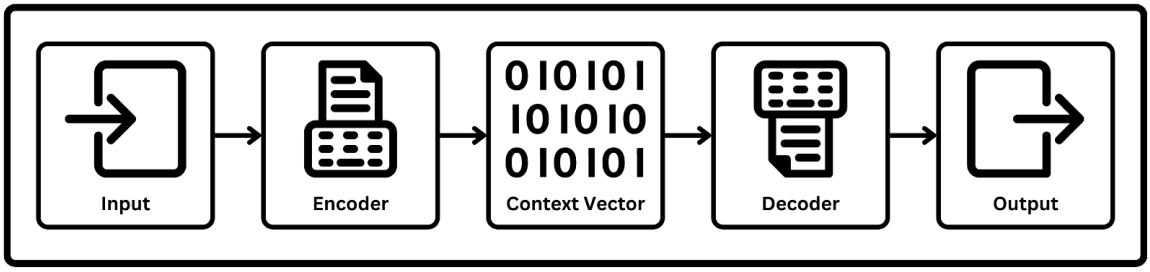

Figure 4. General Architecture of Sequence-to-Sequence Models for Summarization and QA.

3.4.3 Retrieval-Augmented Generation (RAG)

Recognizing the limits of the QA capacity of generative models, Retrieval-Augmented Generation (RAG) combines information retrieval with generative models to enhance performance in open-domain QA tasks. RAG systems consist of two main components: the retriever, which fetches relevant passages from a large corpus using vector embeddings like Dense Passage Retrieval (DPR) for efficient semantic similarity searches, and the generator, which synthesizes answers by passing the retrieved passages and input query to a generative model such as BART or T5 as shown in Figure 5. By grounding generated answers in retrieved evidence, RAG mitigates the risk of hallucinating incorrect information, ensuring relevance and accuracy.

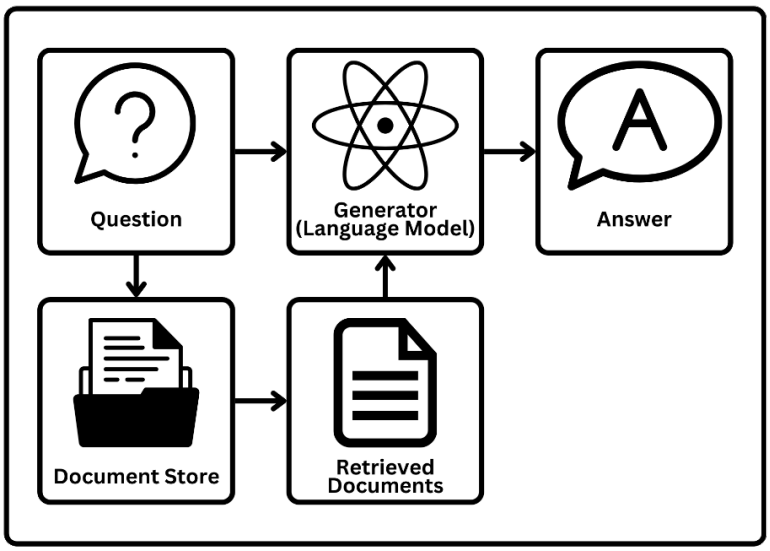

Figure 5. General Architecture of RAG based Models Summarization and QA.

Notable contributions include Lewis et al. [14] who introduced the RAG framework, demonstrating its superior performance by integrating dense retrieval and generative capabilities for grounded QA. Similarly, Karpukhin et al. [47] developed DPR, which significantly improved the retriever's efficiency and accuracy by leveraging bi-encoder architectures for passage ranking. These works collectively highlight the robustness of RAG in addressing the limitations of standalone generative models.

3.5. Hybrid Approaches

Hybrid approaches are a synthesis of traditional methods and neural models trying to exploit both paradigms' advantages. In particular, these approaches are extremely effective in situations where neural models have contextual understanding that can help the structured reasoning of traditional algorithms. The rest of this section describes how these methodologies are combined, and in particular examines graph based reasoning and contextual representation tools.

3.5.1 Combining Traditional Methods with Neural Models

Meanwhile, hybrid frameworks combine exploratory techniques such as rule based algorithms or a statistical technique as a preprocessing or feature engineering stage and serve enriched data to a neural approach for better performance results as shown in Figure 6. In particular, this integration improves on tasks where structured information and deep contextual understanding are both important such as multi document summarization and question answering (QA).

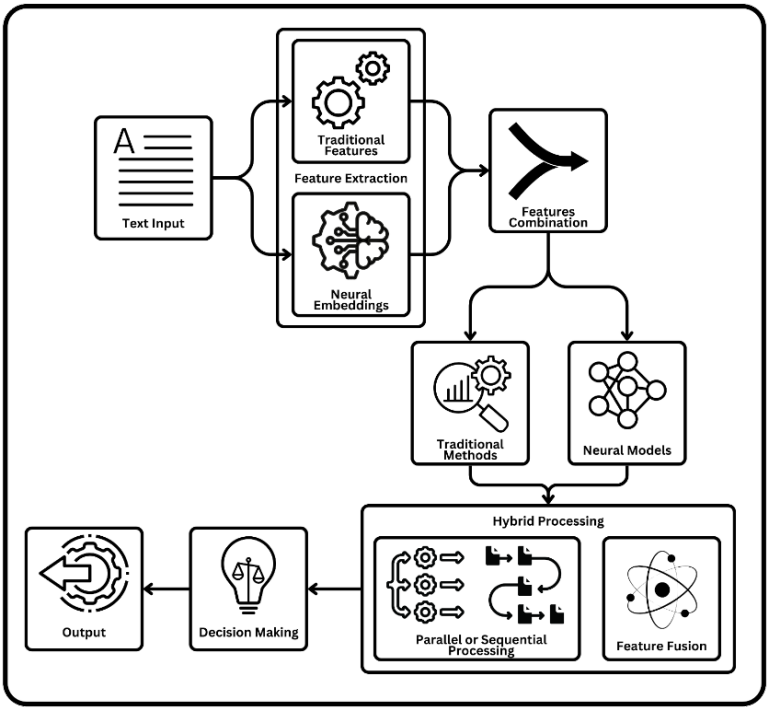

Figure 6. General Architecture of Neural Networks and Traditional Models.

- Feature Augmentation:

In traditional approach, it is common to employ methods like TF-IDF or Part of Speech (POS) tagging to extract certain important information features from a particular document which we consider important based on our ratio formula. These features could be further complemented by embeddings from other neural language models such as BERT or GPT to improve the quality of the summaries. This work by Mikolov et al. [48] also presents Word2Vec a neural model for word embeddings which overcame the limitations of the traditional features map enhancement by also capturing semantic information. Incorporating this notion further, with a model as powerful as BERT, [49] worked with transformer that has the capacity to produce contextualized embeddings for the text leading to the enhancement of text summarization and QA tasks. Liu and Lapata [50] elaborately explained how the use of BERT-based models scores higher than prior techniques on both extractive as well as abstractive summarization due to the integration of pre-learned representations as well as per-sentence features.

- Pipeline Architectures:

Neural models are provided passages by hybrid QA systems based on the traditional information retrieval methods for the BM25. This helps in maintaining more relevancy context specific that actually helps the neural model in reducing computational complexity and enhances the results accuracy. For example, a hybrid pipeline may apply TF-IDF for identifying suitable passage(s) from two or more documents, and then incorporate a transformer model like BERT to fine-tune the process and generate a summary or an answer. Robertson & Zaragoza [51] did the basic work on BM25 in detail and described how it achieves high ranking precision through term frequency and inverse document frequency. Chen et al. [52] showed that incorporating TF-IDF with neural readers can be used in open-domain QA, and incorporating systems such as DrQA that are retrieval along with neural models that could be used for end-to-end answering. Subsequently the authors [47] proposed Dense Passage Retrieval (DPR), in which, hybrid QA was stepped up by using dense embeddings with a transformer-based retriever in getting better candidate selection and response formulation.

3.5.2 Graph-Based Techniques for Reasoning and Contextual Representation

Graph based techniques have emerged as a powerful tool for summarization of multi document as well as QA as shown in Figure 7. This class of methods represents documents, sentences, and entities as graph nodes and relationships between them as graph edges, e.g., when two items cooccurred, have semantic similarity, or are logically connected.

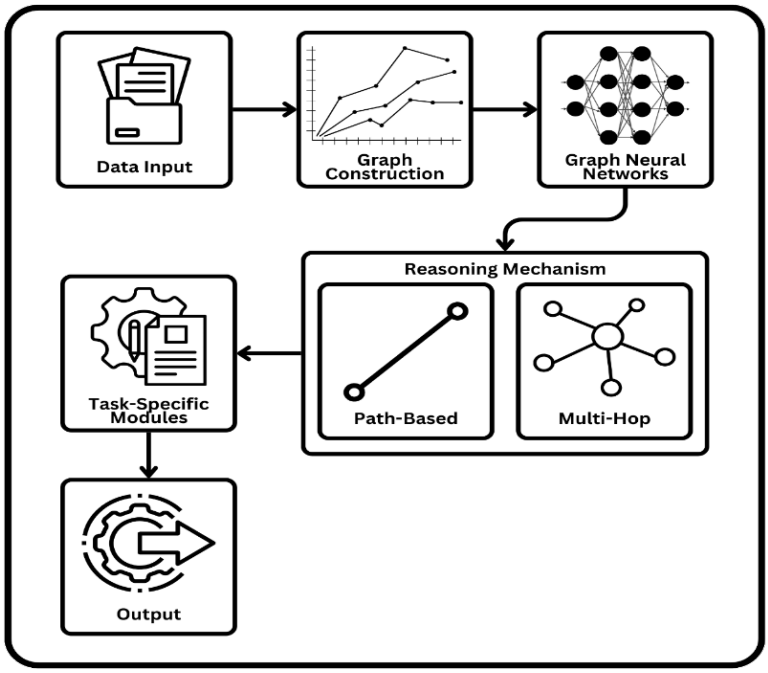

Figure 7. General Architecture of Graph Based Techniques.

- Graph Construction:

We then build the graph with the nodes as the individual sentences, phrases or entities and the edges as the cosine similarity between word-to-word embeddings or the dependency relations which are parsed. Coherent outputs involve this graph structure where reasoning and summarization processes operate. [17] projected a graph-based approach, the TextRank, which models text as a graph and incorporates resemblance of sentences for weights when ranking. Knight and Marcu [53] suggested dependency-based approaches to text abstracting concentrating on the syntactic dependencies as semantic associations. Furthermore, Yasunaga et al. [54] indicated how GCNs could easily diffuse and reason about evidence about relationships in given structural data of textual properties to improve on graph-based documents, summarization as well as question answering through complex relationships. The strength and equations of these techniques are shown in Table 1 and Table 2.

Technique	Method Type	Key Strengths	References
TF-IDF	Extractive	Simple, interpretable, works well for term importance	[47]
TextRank	Extractive (Graph)	Captures sentence similarity and centrality in a document; effective for short summaries	[17]
LexRank	Extractive (Graph)	Sentence clustering ensures coherent and contextually relevant summaries	[28]
Rule-based QA	Rule-based	Effective for structured queries with fixed patterns	[30] [31] [32] [33]
Information Retrieval	Extractive (QA)	Fast retrieval of relevant documents based on keywords or semantic similarity	[16][14]
Graph-based Reasoning	Graph-based	Excellent for modeling complex relationships between entities and multi-hop reasoning tasks	[55][28][56]
BERT	Transformer (Encoder)	Bidirectional context understanding, robust for extractive QA	[36][37][38] [39][10]
GPT	Transformer (Decoder)	Effective for generative tasks, excels in abstractive summarization	[41][42]
T5	Transformer (Seq2Seq)	Unified text-to-text framework, versatile for summarization and QA	[43][44][14]
BART	Transformer (Seq2Seq)	Combines bidirectional encoding with autoregressive decoding, robust for abstractive tasks	[45]
Seq2Seq Models	Encoder-Decoder	Effective for abstractive summarization and QA	[43][46]
RAG	Retrieval + Generation	Combines retrieval for evidence with generation for fluency, excellent for open-domain QA	[14][47]
Traditional + Neural	Combines rule-based/statistical methods with neural embeddings	Balances structured reasoning with deep contextual understanding	[48][49][50]
Graph-Based Techniques	Represents documents and relationships as graphs	Captures complex dependencies and logical relationships across documents	[17][53][54] [57][58]
Pipeline Architectures	Uses traditional IR for retrieval and neural models for refinement	Reduces computational overhead, improves focus on relevant information	[47][51][52]

Table 1. MDTS and QAS Techniques and Their Strengths

- Graph Neural Networks (GNNs):

GNNs are used to transmit information throughout the acquired graph, taking into account both the proximity and more extensive topological interactions. For instance, a GNN can transduce information from neighboring nodes to favor the representation of each node, thereby becoming contextually superior in representing an associated node. Surprisingly, the authors [57] presented a considerable extension known as Graph Convolutional Network that can collect information from local graph neighbors efficiently and has become the cornerstone of many GNN

applications. Subsequently, [58] introduced Graph Isomorphism Networks (GIN) as a more expressive model of graph structures integrated with multilayer perceptron for node aggregation. Moreover, in multi-document summarization and QA tasks, Yasunaga et al. [54] also claimed how GNNs can improve the reasoning process across scattered textual information within the graphs formed from text. The suitable development environment of these techniques is shown in Table 3.

Technique	**Equation**
TF-IDF	$TF - IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$
TextRank	$S_i = \frac{1}{d_i} \sum_{j \in N(i)} \binom{w_{ij}}{d_j} \times S_j$
LexRank	$S_i = \sum_{j=1}^{n} \frac{A_{ij}}{\sum_{k=1}^{n} A_{ik}} \times S_j$
Rule-based QA	Answer = Rule(Q)
Information Retrieval	$Score(d,q) = \sum_{t \in q} TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$
Graph-based Reasoning	$R = \sum_{i \in V} Weight(i) \times NeighborSum(i)$
BERT	$BERT(Q) = Encoder(Q)$
GPT	$GPT(Q) = Decoder(Q)$
T5	$T5(Q) = Seq2Seq(Q)$
BART	$BART(Q) = Seq2Seq(Q)$
Seq2Seq Models	$Seq2Seq(Q) = Encoder(Q) \rightarrow Decoder(Q)$
RAG	$RAG(Q) = Retriever(Q) \rightarrow Generator(Q)$
Traditional + Neural	$Traditional + Neural(Q) = Rulebased(Q) + Neural(Q)$
Graph-Based Techniques	$G = \sum_{i=1}^{n} GraphWeight(i) \times Dependency(i)$
Pipeline Architectures	$Pipeline(Q) = IR(Q) + NeuralModel(Q)$

Table 2. MDTS and QAS Techniques and Their General Equations

Technique	**Method Type**	**Tools/Software**	**Implementation Environment**	**Suitable Datasets**	**Key Limitations**	**Proposed Solution**
TF-IDF	Extractive	Scikit-learn, NLTK	Python, Jupyter Notebooks	Multi-News, CNN /DailyMail	Does not capture semantic meaning or sentence relationships	Combine TF-IDF with word embeddings like Word2Vec or GloVe to incorporate semantic understanding.
TextRank	Extractive (Graph)	Gensim, NetworkX	Python, Jupyter Notebooks	Multi-News, CNN /DailyMail	Assumes graph structure may not always align with the narrative	Use contextual embeddings (e.g., BERT embeddings) to refine graph weights and improve narrative alignment.
LexRank	Extractive (Graph)	Gensim, NetworkX	Python, Jupyter Notebooks	Multi-News, WikiSum	May struggle with highly diverse or long documents	Introduce topic segmentation or clustering to handle diversity and improve coherence.
Rule-based QA	Rule-based	SpaCy, NLTK, Prolog	Python, Prolog	SQuAD, Natural Questions	Limited flexibility, poor at handling ambiguous or novel queries	Combine rule-based methods with neural embeddings for better generalization and flexibility.

Information Retrieval	Extractive (QA)	Elasticsearch, Whoosh, Solr	Python, Java, Elasticsearch Server	SQuAD, TriviaQA	Relies heavily on keyword matching, often misses context	Use semantic search models like DPR or BM25 with contextual embeddings for better context capture.
Graph-based Reasoning	Graph-based	NetworkX, DGL	Python, DGL, PyTorch, TensorFlow	Multi-News, WikiSum	Can be computationally expensive for large graphs	Optimize graph structures with pruning techniques and leverage distributed computing frameworks.
BERT	Transformer (Encoder)	Hugging Face Transformers, TensorFlow	Python, PyTorch, TensorFlow	SQuAD, Natural Questions	Limited generative capabilities	Fine-tune BERT with a generative decoder for abstractive tasks or combine with GPT models.
GPT	Transformer (Decoder)	OpenAI GPT API, Hugging Face Transformers	Python, PyTorch, TensorFlow	XSum, Gigaword	Unidirectional context modeling, prone to hallucination	Use bidirectional models like T5 for context modeling and introduce grounding mechanisms to reduce hallucination.

T5	Transformer (Seq2Seq)	Hugging Face Transformers, TensorFlow	Python, PyTorch, TensorFlow	XSum, SQuAD	Computational-ly expensive for large inputs	Use model distillation or quantization to reduce compu-tational overhead.
BART	Transformer (Seq2Seq)	Hugging Face Transformers, PyTorch	Python, PyTorch, TensorFlow	XSum, CNN /DailyMail	Requires substantial compu-tational resources	Implement sparse attention mechanisms or low-rank approxi-mations to improve efficiency.
Seq2Seq Models	Encoder-Decoder	TensorFlow, Keras, Hugging Face	Python, Tensor-Flow, Keras	SQuAD, Gigaword	Struggle with long input sequences without attention mechanisms	Use transformers or hierarchical attention mechanisms to handle long inputs effectively.
RAG	Retrieval + Genera-tion	Hugging Face Transformers, FAISS	Python, PyTorch, TensorFlow, FAISS	SQuAD, TriviaQA	Dependent on retriever quality, may retrieve irrelevant passages	Train with hard negatives and cross-encoder models for better relevance.
Traditional + Neural	Combines rule-based /statistical methods with neural embed-dings	Scikit-learn, SpaCy, Hugging Face	Python, Jupyter Notebooks, TensorFlow, PyTorch	Multi-News, SQuAD	May require extensive fea-ture engineer-ing and man-ual tuning	Automate feature selection using neural architecture search or embedding-based techniques.

Graph-Based Techniques	Graph-based	NetworkX, DGL, PyTorch	Python, PyTorch, TensorFlow, DGL	WikiSum, Multi-News	Computationally intensive, especially for large graphs	Use graph sampling techniques like GraphSAGE or mini-batching for scalability.
Pipeline Architectures	Uses traditional IR for retrieval and neural models for refinement	Hugging Face Transformers, Elasticsearch	Python, TensorFlow, PyTorch, Elasticsearch Server	SQuAD, Multi-News	Dependent on retrieval quality, may propagate errors from earlier stages	Incorporate feedback loops and end-to-end fine-tuning for error mitigation across pipeline stages.

Table 3. MDTS and QAS Algorithms and Software Packages/Tools, Their Limitations, and Proposed Solutions

4. Challenges in Algorithm Design

Design and implementation of algorithms for multi-document summarization (MS) and question answering (QA) are challenged by many tasks' dependent ones and also those for both the tasks. The inherent complexity of natural language, diversity of real-world data, and computational demands of modern techniques result in these challenges. First, it discusses general challenges in summarization and QA and cross cutting issues affecting both domains.

4.1. Summarization-Specific Challenges

Algorithms for multi-document summarization are tasked with condensing large volumes of information into concise and meaningful summaries. However, this process is fraught with several key challenges:

4.1.1 Redundancy

One of the recurrent challenges in extractive summarization techniques is redundancy that is, information from several documents, which is either similar or overlapping, can be included in the final summary [59] [60]. Moreover, this does not only shorten the utility of the summary, but rather makes it unnecessarily longer. By addressing redundancy but often run into short and broad pipeline take in generating concise abstractive methods.

4.1.2 Coherence

This is another big challenge, and particularly so when pulling together data from other sources. Summaries have to have logical flow, and retain a consistent narrative structure [53]. Unfortunately, extractive methods generally don't do this, only piecing together sentences from the source documents, which create disjointed outputs.

4.1.3 Factual Consistency

In abstractive summarization, factual correctness is a major concern: neural models can inadvertently introduce incorrect facts or hallucinate information that is not in the source [22]. This makes the summaries unreliable and threatens to raise stakes in domains like healthcare or finance where accuracy is everything.

4.2. QA-Specific Challenges

Question-answering systems face unique challenges that stem from the need to understand and reason over complex queries and datasets.

4.2.1 Ambiguity

It is fraught with a great degree of ambiguity. Multiple valid interpretations of a single query exist, which the QA system must know to disambiguate [61]. For example, “What is the capital?”, In some sense could mean financial capital or, rather, geographical capital.

4.2.2 Multi-Hop Reasoning

QA systems are forced to synthesis (information) from multiple sources or (synthesis) in multiple steps to answer (complex) queries in Multi-Hop Reasoning [13]. An example is answering ”Which country’s leader won the Nobel Peace Prize in 2020?” identification of the leader, the country that gave them the award and the award details. This requires great advanced reasoning and great contextual understanding.

4.2.3 Hallucination

A critical problem that we encounter in neural QA models is hallucination, where models generate plausible sounding answers that are factually incorrect or unsupported by the given data [62]. As well as eroding the credibility of the system, it exposes the system to risk in sensitive applications.

4.3. Cross-Cutting Issues

Certain challenges are common across both summarization and QA due to their reliance on similar data and computational frameworks.

4.3.1 High-Dimensional Data

Is a fundamental challenge in NLP, where algorithms must process large volumes of text containing diverse linguistic patterns, entities, and relationships [63]. Managing this complexity requires sophisticated models and efficient preprocessing techniques.

4.3.2 Resource Intensiveness

We share a similar concern when it comes to Resource Intensiveness, especially for modern neural architectures such as transformers that require a lot of compute and memory [64]. However, training and deploying these models at scale still seems prohibitive to many researchers and organizations.

4.3.3 Bias

Bias in training data and algorithms is a pervasive issue that affects the fairness and generalizability of summarization and QA systems [61]. Bias in summarization and QA systems can manifest in various forms, including dataset bias, algorithmic bias, and societal bias, each contributing to the reinforcement of inequities in real-world applications. Dataset bias is present when the data used to train these models is an expression of

societal prejudice, like a healthcare QA system trained on one demographic group's data and whose outputs do not generalize to other social groups. These problems are too great, and algorithmic bias makes them worse, since the design and architecture of the model itself amplifies inherent biases in the training data [65]. For instance, if the data used in legal summarization tools is biased to a legal corpus that is mostly homogeneous, then the assumption of the algorithms can result in favoring some genders, races, or social groups. As for the data and the algorithmic design, they are influenced by both societal bias, which is determined by the norms and values of a society, and the lack of data. For example, job related QA systems are based on a societal bias of making gendered or culturally based recommendations that recommend male dominated roles for male users and female dominated roles for female users. [66][79]

Organizations working in sensitive domains, like healthcare and law, these biases have a high actual impact. We also observe in healthcare that biased summarization tools can mislead healthcare practitioners, which results in minority populations receiving unequal treatment in clinical trial data, as is the case with clinical trial data that underrepresents some demographic populations. Well, in legal environment application also, biased QA or summarization system might misguide the legal practitioners and enforce the racial or gender biases in legal judgments. Moreover, these biases hurt individual as also violate the fairness and integrity of these systems. Furthermore, these stereotypes could perpetuate other biases in other sectors, such as in terms of education and employment. To address these issues, these systems need to be designed carefully based on the data and the corresponding algorithmic design and the values of this society to make them fruitful for all users.[68][69]

4.4. Metric Standardization and Reproducibility

A major limitation of the use of evaluation metrics such as ROUGE and BLEU is the unavailability of standardization of their application, which negatively affects the use of consistent and fair comparisons amongst studies. Even though these metrics are widely used in the multi-document summarization and question-answering tasks, there is much divergence on how they are implemented between disparate research works. For an example, many different variants of ROUGE (ROUGE-1, ROUGE-L etc.) are employed without a rationale as to which to choose. Because it is not possible to directly compare studies results, with independent evaluation methods, this inconsistency makes it difficult to present results in a direct manner. Moreover, some of the studies focus on certain aspects, such as factuality, coherence or relevance, and not all of these can be accounted by the traditional metrics. Results cannot be interpreted because there is no standardization in the selection of metrics, thus making this more complicated[70] [71].

A second is because replicability remains another big problem, especially with neural models because the hyperparameters, the training protocols and the dataset splits which are often reported inconsistently or omitted altogether. Lack of information about these aspects makes it difficult for researchers to reproduce the experiment or to confirm the result, thus withdrawing progress in the field. If a result cannot be reproduced by other researchers, the results they're reporting are untrustworthy and their work is nearly impossible to build on. These problems need to be mitigated by the timely need for standardized reporting practices, including specifications of hyperparameters, training procedure and dataset split. In practice, such measures would not only guarantee reproducing experiments, but also encourage the transparency and collaboration, which makes research more reliable. This would ensure that better comparisons and validations can take place across studies, thus aiding the progress of the multi-document summarization and QA systems.[72]

4.5. Real-World Deployment Challenges

Summary and QA models are deployed in real world applications and there are inherent problems involving scalability, latency, and integration with existing workflow. However, transformer-based models such as BART and T5 are quite effective, yet they are quite resource intensive, which means they are not suitable for large scale deployment. However, since real time systems like virtual assistants call for very low latency response times, but autoregressive models are inherently delayed, where to put the delay becomes a question. Techniques like quantization as well as model additional can reduce inference time while slightly affecting output quality [73][74].

If the workflow being considered is already integrated with legacy pipelines, for example, in healthcare or finance, there are additional complications regarding integration with the regulatory constraints and the nature

of the structured decision-making pipelines. For example, there is little feasibility with deploying models on low resource edge devices that serve in emergency healthcare. Combination of rule-based heuristics with machine learning models can also be used as the hybrid approach to integration and to comply. Future work related to optimized model for constrained environment and adaptive architecture which trade off between accuracy and efficiency should be considered [75][76].

4.5.1 Dynamic Information Adaptation

Several commonly used summarization and QA models assume that document collections are static, and do not achieve good performance on dynamic environments where the information changes frequently. For instance, news articles, medical guidelines, financial reports need to be updated frequently, whilst models should be able to evolve under new data while remaining stable [77]. This further complicates the reliability of these models where the data has statistical properties that change over time. To tackle this issue, methods that include continual learning, online updating mechanisms and adaptive retraining strategies are necessary so that models can leverage continuously incoming information incrementally [78][79] [80].

One of the best approaches to solving this problem is by integrating continuous learning frameworks that allow models to continue learning as time goes by. Elastic Weight Consolidation (EWC) or memory-augmented networks may provide solutions that enables previously learned knowledge to be preserved when new data do exist. For instance, EWC mitigates catastrophic forgetting by penalizing changes in important weights that occur during learning, and thus retains critical knowledge. On the contrary, memory-augmented networks rely on external memory storage to be able to dynamically adapt to new tasks and information by flexibly retrieving and updating the model[81] [82].

It also alleviates the difficulty of incrementally updating knowledge bases used by retrieval-augmented generation (RAG) systems with new documents. Given that real time data is evolving, these pipelines make it possible for systems to update their knowledge base with no need for retraining from scratch. To enable a model to adapt to evolving datasets, it is possible to incrementally fine tune on datasets and use uncertainty aware learning strategies. With focus on such continuous learning and incremental updating of AI systems, AI systems can choose to improve their system reaction to real world data, while accuracy and freshness are kept [83] [84] [85].

4.6. Green AI and Efficiency: Addressing Environmental Impact

Large scale transformer-based models, such as GPT-3 require these large computation costs, which means they have high energy consumption and carbon emission. Training these models are energy intensive, that often rely on non renewable energy, and thus have a large carbon footprint. The concern over how the environment is impacted from increasingly AI adoption across industries is an issue to keep an eye on. Since real time inference operations are added to trainers, their sustainability is greatly affected. In real time continuous processing scenarios or when frequently communicating with cloud, for instance for deployment of models on low resource devices, energy usage is further pushed up [86] [87].

4.6.1 Energy-Efficient Architectures

Since the computational load has been increasing in large scale models, it is very important to promote an energy efficient architecture to minimize environmental impacts from AI systems. Model distillation, quantization, and sparse attention mechanisms are very useful techniques that can greatly decrease the computation overhead without hampering the model performance. Model distillation transmits knowledge from a large and cumbersome model to a small and energy efficient one. On the contrary, the precision of the parameters of a model is reduced in quantization, saving on the computational resources, whilst keeping the effectiveness of the model. Attention mechanisms that are sparse allow models to only concentrate on the most important bits of the input so that computations are minimized and energy is saved [88] [89] [90].

4.6.2 Transparency and Reporting of Environmental Costs

Apart from developing energy efficient architectures it is also important that the inclusion of energy consumption metrics, such as Flops and Carbon Emissions in the research publications. This reporting also helps increase awareness of environmental costs involved with training and deploying large models. Transparency is critical to encourage more green AI technologies being developed and thus, allowing researchers and practitioners to better inform their decisions in the tradeoff between model performance and the environmental impact [91][92] [93].

4.6.3 Adaptive Computation and Green AI Practices

In order to address the energy concerns, researchers are studying the adaptive computation strategies where the models refine their complexity according to the properties of the tasks. There should be some initiatives regarding carbon-aware scheduling and green AI practices spread into model development workflows, to decrease a bigger part of the environmental footprint of AI systems. All of these adoption, energy efficient hardware, renewable cloud infrastructure, eco-conscious scheduling practices, can in total lessen the carbon emissions in the AI deployment [94] [95] [96]

5. Trends and Innovations

In recent years, there has been significant progress in natural language processing (NLP); with unprecedented impact on multi document summarization and question (QA) answering systems as shown in Table 4. The increasing power of large-scale language models, integration of multimodal data and advanced learning paradigms have all led to these innovations [97]. We discuss the latest established trends in the area, in particular, unified models, multimodal algorithms, reinforcement learning, and methods built for a low resource scenario or cross lingual transfer [98][99].

5.1. Recent Developments in Summarization and Question Answering

The rapid advancement of AI models has significantly influenced the fields of text summarization and question answering (QA). While this paper discusses general trends in these domains, a more in-depth examination of recent state-of-the-art models further highlights the ongoing progress and their impact [100]. In particular, GPT4, and PaLM have sophisticated architectures of increased summarisation and QA related gains, as they have a bigger training dataset scale and much better reasoning capabilities [101] [102].

One hundred years ago, Sir Arthur Conan Doyle and Sherlock Holmes were also able to understand what collecting context entails and turned the whole thing into a proper summary. It is different to previous exemplications in the sense that it offers better alignment with human intent, yielding results more concisely and accurately over different domains such as news, legal and medical texts. This model can then use such prompts to understand the subtle prompt and give domain specific summaries, a useful tool when one wants precise and precision domain specific summarization in such applications[103] [104]. On settings similar to that, its performance is also strong, exploiting the fact that it was trained over a vast amount of text across various linguistic resources as was the case with Google's PaLM. It can be used for cross lingual summarization and QA tasks such that broader language accessibility is considered as well as dealing with underrepresented language challenge [102].

Along with this, these models have also enhanced contextual understanding and reasoning for QA systems further. PaLM and especially GPT-4 are more capable of answers to a multi turn question with better responses for extended coherence dialogues. In addition, the facts are leveraged for a stronger fact based answering through retrieval augmented generation (RAG) where external knowledge sources help minimally risk hallucination events and ensure that fact are consistent [105] [106].

5.2. Unified Models for Summarization and QA

A trend in NLP (Natural Language Processing) is unified models like GPT-4, T5, BART. These models are meant to operate a single framework under which they can tackle other tasks, such as summarization, QA, etc [107].

A case in point is the T5 model (Text-to-Text Transfer Transformer) which casts all NLP tasks as text-to-text transformations. To take just two examples, summarization problems become problems of generating a shorter version of the input text, while QA is a problem of generating an answer based on question and the related context [43]. This unification makes training and fine tuning much easier and lets models benefit from shared representations across tasks. The T5 training objective is defined as:

GPT-4 and similar models further extend this paradigm by supporting multimodal inputs and generating context-aware responses across diverse domains. These unified models reduce the need for task-specific architectures and enable efficient transfer learning.

5.3. Multimodal Algorithms

Text, image, and video integration offers new ground on which summarization and QA can happen. The goal of multimodal algorithms is to process and reason about heterogeneous data sources leading to richer and/or more informative outputs.

For instance, multimodal summarization systems create summaries that combine textual content with matching images or video, and are both information rich and visually appealing [108]. Likewise, multimodal QA systems make use of visual input, say diagram or charts to answer questions that textual input alone could not. Conventional architecture of multimodal systems has separate encoders for each modality and a fusion layer to combine their representations [109]. Their fusion is then processed by a decoder which results in the desired output.

5.4. Reinforcement Learning for Optimizing Responses

Optimizing summarization and QA systems is a risky proposition, and reinforcement learning (RL) has been a promising technique for this. RL provides a way for models to learn policies that serve to maximize task specific rewards (such as informativeness, coherence, or factual accuracy) [110].

RL is often applied to manage some other metric like ROUGE or METEOR in summarization. In interactive setting where user feedback is available RL can be used to increase answer relevance and correctness for QA [111].

5.5. Few-Shot and Zero-Shot Learning

Few shots and zero shot learning tackle the problem of low resource tasks and languages by allowing a model to generalize from little to exactly zero task specific data. Rather, these approaches rely on pre trained language models and prompt engineering to generalize to new tasks with little additional training [112].

As an example, GPT-4 is able to perform few shots learning by conditioning on a small set of task examples given as part of the input prompt [61]. In contrast, zero shot learning requires the model to output for unseen tasks based only on the model's general knowledge.

These techniques are most applicable to low-resource languages, where annotated datasets are sparse. With access to cross-lingual transfer and shared representations, few-shot, and zero-shot models can achieve robust performance across a wide range of linguistic settings.

5.6. Cross-Lingual and Multi-Language Algorithms

In order to have summarization and QA capabilities across multiple languages and underrepresented languages, both cross-lingual and multi-language algorithms are used [113]. They use these models to best exploit multilingual word embeddings and transfer learning to solve the task in new linguistic contexts. For example, multilingual BERT (mBERT) and XLMMagento RoBERTa, are pre-trained on text taken from many different languages and understand and produce in multiple language settings [7]. Cross-lingual transfer techniques further enhance performance through using data from high resource to high resource languages for the low resource language.

Further work on data scarcity in underrepresented languages especially in cross lingual contexts needs to be approached with initiatives that foster collaboration and data sharing [114]. These challenges are addressed by promising crowd and the crowd sourced annotation platforms. Thus, these platforms allow the annotation of data by a large variety of contributors (native speakers and linguists) increasing data diversity and representativeness of training datasets. The result of this helps develop more inclusive AI models, which also helps in preserving and growing linguistic diversity [115]. People are engaged with global communities and linguists, to make AI systems available and applicable in different cultural contexts around the world.

Technique	**Key Features**	**Advantages**	**Limitations**	**Suitability**
Recent Advanced Models (GPT-4, PaLM)	Large-scale pre-trained language models	Strong contextual understanding; improved accuracy	Requires extensive computational resources	High-performance summarization and QA
Retrieval-Augmented Generation (RAG)	Enhances QA with external knowledge	Improves factual consistency; reduces hallucinations	Relies on quality of retrieval sources	Fact-based QA, long-form summarization
Transformer-Based Architectures	Self-attention for contextual learning	Captures long-range dependencies effectively	High inference time; large memory requirements	Summarization, QA, language modeling
Unified Models	Single framework for multiple tasks	Simplifies training; efficient transfer learning	Requires large computational resources	Summarization, QA, multi-task scenarios
Multimodal Algorithms	Integrates text, images, and videos	Richer outputs; supports diverse data sources	Complex architecture; limited benchmarks	Multimodal summarization and QA
Reinforcement Learning	Optimizes task-specific rewards	Direct metric optimization; interactive learning	Reward design is challenging; high training cost	Summarization and interactive QA
Few-Shot and Zero-Shot Learning	Adapts to new tasks with minimal data	Low-resource applicability; versatile	Sensitive to prompt design; inconsistent results	Low-resource languages and tasks
Cross-Lingual Algorithms	Supports multiple languages	Expands reach to underrepresented languages	Limited by pre-training data diversity	Multi-language summarization and QA

Table 4. Comparison of Advanced Techniques for Summarization and Question Answering (QA) with Key Features, Advantages, Limitations, and Suitability

6. Applications and Real-World Relevance

Multi-document summarization and question answering systems have a variety of potential practical applications in many domains that dramatically highlight the potential transformation that they can bring to solving real world complex problems. These technologies, which provide concise, context aware information and aid interactive inquiry, are now critical in news, legal, medical domains; virtual assistants; in research tools; and in education [116]. In particular, the integration of summarization and QA capabilities to decision support systems is becoming increasingly important in critical decision making.

6.1. Summarization in News, Legal, and Medical Domains

Multi document summarization has enjoyed extensive use in domains where there is a need to synthesize large amount of information.

- **News**: Summary systems in journalism help generate summaries of breaking news stories by aggregating information from multiple sources. These summaries help readers stay informed about key developments since their last reading, while maintaining factual consistency. Systems like Google News present a unified view of current events [117]. To enhance the relevance and accuracy of these systems, partnerships with journalists and news editors are essential. Collaboration with domain experts can ensure that summarization tools are fine-tuned to reflect the nuances of news reporting and meet the evolving demands of the media industry.
- **Legal**: Summarization tools are used by legal professionals to read case law, contracts, and legal briefs. These systems extract critical information such as precedents and clauses, allowing legal professionals to review documents in less time and with more efficiency [107]. In the legal domain, summarization plays a key role in deriving essential information, aiding in litigation and compliance processes. Close collaboration with legal experts is vital to co-design the datasets and evaluation criteria, ensuring that the summaries produced are legally accurate, relevant, and aligned with the specific needs of the legal field.
- **Medical**: In the medical field, summarization systems assist clinicians by aggregating patient records, research articles, and clinical guidelines. These tools are particularly valuable in decision-making, especially in the high-pressure environment of emergency care [1] [107]. Medical professionals also benefit from summaries of the latest research, helping them stay informed about new developments in medical science. Collaborating with clinicians, researchers, and healthcare professionals to co-design these datasets ensures that the summarization models are accurate, contextually relevant, and tailored to the needs of practitioners, thus supporting better decision-making and patient care.

6.2. QA in Virtual Assistants, Research Tools, and Education

Question-answering systems are integral to enhancing user interaction and providing precise information across various applications.

- **Virtual Assistants**: QA systems power virtual assistants like Siri, Alexa, and Google Assistant, enabling them to respond to user queries with contextually relevant answers [118]. These systems rely on advanced language models to process natural language inputs and retrieve accurate responses.
- **Research Tools**: Researchers utilize QA systems to query scientific databases and retrieve information from large corpora of academic papers. For example, tools like Semantic Scholar integrate QA capabilities to allow users to extract insights from research articles efficiently [119].
- **Education**: In education, QA systems serve as interactive learning aids. By answering student queries and providing explanations, these systems enhance the learning experience. They also support educators by generating quiz questions and summarizing lesson content [120].

6.3. Combining Summarization and QA for Decision-Support Systems

The integration of summarization and QA capabilities into decision-support systems is a growing trend in various industries.

- **Healthcare**: Summarization and QA are very important in healthcare and decision support systems to inform clinicians with inference which helps them make more informed decisions, e.g. summarizing patient histories to answer diagnostic queries [121]. But such systems deal with sensitive and confidential information which means that privacy and security issues have to be seriously addressed. A privacy violation can result in legal repercussions when personal data, which could be patient information in healthcare summaries, are accidentally disclosed. To mitigate these risks, techniques like federated learning and differential privacy can be employed.

- **Federated Learning**: Federated learning enables models to be trained across decentralized data sources without transferring sensitive data, thus preserving privacy. In this approach, data remains local to its source, such as hospital records, while only model updates are shared across devices or nodes, minimizing the risk of data breaches [122].
- **Differential Privacy**: Differential privacy introduces noise to the data, ensuring that individual information cannot be extracted from the model's outputs. This further enhances data protection in healthcare summarization and QA systems, preventing the accidental exposure of private information [123].

- **Business**: In the corporate sector, decision-support systems assist executives by summarizing market trends and answering strategic questions [51]. This integration streamlines decision-making processes and enhances organizational efficiency. The use of privacy-preserving techniques, such as federated learning, becomes essential to ensure sensitive corporate data, such as financial records, remains protected throughout model training and deployment.

- **Public Policy**: Policymakers leverage decision-support systems to analyze reports, summarize legislative documents, and answer policy-related questions [124]. These capabilities facilitate the development of informed and effective policies. As these systems handle sensitive governmental data, privacy-preserving methods like federated learning and differential privacy are crucial to safeguarding confidential information, such as internal policy drafts or demographic data.

7. Open Research Directions

The area of multi-document summarization and question answering (QA) is very dynamic and changing very rapidly, hence there are lot of opportunities to push the research. Much progress has been made, but certain critical gaps with areas not yet explored are still present. This section provides open research directions in holistic evaluation frameworks, generation of comprehensive datasets in this context, as well as ethical implications and trustworthiness in automated systems.

7.1. Frameworks for Holistic Algorithm Evaluation

A major challenge in the area is the absence of standardized and wide-ranging frameworks for evaluating algorithms between tasks. Many existing evaluation metrics concentrate on partial aspects: for summarization tasks, like ROUGE or BLEU are based on lexical overlap, for QA tasks Exact Match or F1 Score are accuracy based [1] [111]. These approaches, while valuable, fail to capture nuanced aspects such as coherence, factual consistency, reasoning depth, and adaptability across diverse scenarios.

A proposed direction involves the development of multi-dimensional evaluation frameworks that integrate quantitative and qualitative measures. Such frameworks should assess algorithms on a range of criteria, including:

- **Semantic Fidelity**: Evaluating the extent to which generated summaries or answers preserve the intended meaning of the source content [125].
- **Contextual Relevance**: Measuring the ability of algorithms to adapt to varying contexts and user intents.
- **Efficiency**: Incorporating computational efficiency and scalability as critical evaluation dimensions, especially for real-world applications [126].
- **User-Centric Metrics**: Proposing metrics such as usability surveys and task completion rates to assess real-world utility and operational efficiency, focusing on how well systems serve user needs in practical settings. These metrics help in understanding user satisfaction, especially in applications like customer support and virtual assistants [127].
- **Domain-Specific Benchmarks**: Developing tailored evaluation metrics for specialized contexts, such as legal coherence scores for legal documents and medical factual consistency checks for healthcare-related summaries. These metrics would align evaluation with the unique requirements of different domains [104].

Furthermore, advocating for standardized evaluation protocols is crucial for ensuring reproducibility and enabling fair comparisons across studies. Shared tasks with fixed datasets and consistent metrics, such as NIST's TREC challenges, offer a valuable framework for comparing system performance in a controlled, transparent manner. These shared tasks promote consistent methodologies, help identify best practices, and foster meaningful advancements in summarization and QA research by ensuring that experiments can be reliably replicated across different research teams and environments.

Holistic evaluation frameworks would provide a more complete understanding of algorithm performance and foster innovation by setting benchmarks that align with both practical and theoretical objectives [128].

7.2. Comprehensive Datasets with Cross-Task Annotations

The development of robust algorithms for summarization and QA heavily depends on the availability of high-quality datasets [129]. While large-scale datasets such as SQuAD, Multi-News, and CNN/Daily Mail have been instrumental, they are often task-specific and lack annotations that enable cross-task learning.

Future datasets should include:

- **Cross-Task Annotations**: Integrating labels for both summarization and QA tasks within the same dataset. For instance, a dataset could include document summaries alongside corresponding question-answer pairs derived from the same content [130]. This would enable the development of unified models capable of excelling in both tasks.
- **Domain-Specific Data**: Expanding datasets to include specialized domains, such as law, medicine, and finance, to address the unique challenges of these fields [131].
- **Multimodal Data**: Incorporating text, images, and videos into datasets to facilitate the development of multimodal algorithms that can process diverse information sources [108].

7.2.1 Bias Mitigation and Ethical AI

At the same time, automated summarization and question answering (QA) systems are becoming more and more dependent for which raises serious ethical issues related to biases in the training data and the model outputs. The biases can arise from linguistically imbalanced datasets, and cultural imbalances. If left unaddressed, they may lead to misinformation, reinforce stereotypes, or disproportionately disadvantage to certain groups. Thus, transparency in development and evaluation of the model is required to mitigate these risks [132].

Another important way of ensuring fairness in AI systems involves conducting bias audits when datasets are understood or trained. Fairness Indicators and tools like AI Fairness 360 can help us identify and quantify the (existing) biases in the data or the model, so as to understand that the training dataset is representative, and that the trained models do not carry bias. For all kinds of sensitive applications, including those in healthcare, finance, or legal domain, bias in output is extremely significant. Incorporating bias assessments into training and into each run, we can identify and remedy the errors before these models reach customers' hands, delivering models that work and are fairer respectively [133] [134].

While bias audits should be encouraged as a means to fulfill basic ethical requirements, a just and honest approach is also to encourage transparency reports for deployed AI systems. The development of these reports should include details of sources of training data, any potentially biased findings and specific aspects of ethics review processes undertaken. Such insights will help to build trust in AI systems by increasing accountability and users can understand what the models are capable of [135].

Moreover, it is important to integrate diverse feedback of the stakeholders in designing ethical systems for AI development. Involving ethicists, domain experts, and even affected communities in the design process of the algorithm will serve as a way to catch the unintended harms and check whether the standard is met. Bringing in the insight from those who can understand the societal implications of AI and those whose lives are directly touched by the outcomes of it, developers can find proactive ways of addressing the risk of discrimination and other harmful consequences of AI in the early stages of design [136].

Moreover, explainability and interpretability measures are very important for the accountability. These approaches allow researchers as well as practitioners to explore model decisions and fix un-intended biases thus increasing transparency and fairness.

7.3. Standardized Documentation and Toolkits

As AI systems continue to evolve and play a significant role across various domains, ensuring transparency and accountability in their development is crucial. One important aspect of this is the standardization of documentation.

7.3.1 Model Cards and Dataset Datasheets

We argue that dataset datasheets and model cards should be accompanied with a unified purpose to control model building, to bring a consistency for describing model and dataset limitations, biases, and intended use cases. As introduced by the AI community, the model cards are a detailed report about the capabilities, constraints and ethical considerations of machine learning models. Likewise, the datasheet of a dataset provides key information regarding the origin, composition, expected use, and possible bias or assumptions to which the developers and end-users need to be aware in deployment of these systems. The standardized documents are useful as these documents help foster trust and make sure that the models are used responsibly, particularly in the high stakes environments where the stakes of misinterpretation or bias would be significant [137] [138].

7.3.2 Open-Source Toolkits

To decrease the barriers to entry for hybrid approaches merging traditional and neural techniques, development of open-source toolkits proves to be necessary in addition to standardized documentation. In addition, open source helps us provide a unified framework for the implementation of traditional and deep learning techniques by providing comprehensive libraries, such as Hugging Face integrations. They act as toolkits to make the mental and practical experimentation process more viable so that practitioners can implement and adapt hybrid solutions without having to build comprehensive systems from scratch. As well, with the open source platform, it also foster collaboration in the AI community as it provides a platform for exchange of ideas and breakthroughs from researchers and developers in the field to make AI a sound area of research [139][140][141][142].

8. Conclusion

In the last ten years, multi-document summarization and question–answering (QA) have made tremendous progress thanks to improvements in algorithm design and the proliferation of large-scale datasets. The domain started with traditional approaches, including extractive summarization and rule-based QA systems, which provided simplistic, yet effective solutions to early challenges in the domain. While sufficient over time to get us this far, however, the advent of neural and hybrid methods has fundamentally changed the landscape of the field by enabling reasoning beyond simple rule-based methods and contextual understanding, with the ability to perform a greater variety of tasks and in a greater variety of domains. However, these achievements are accompanied by important challenges such as redundancy, coherence, ambiguity, as well as computational needs of high dimensional data.

Although we are far from a perfect system, summarization and QA have huge potential for research and development in the future. The future capabilities of these systems will be shaped by unified models which naturally combine summarization and QA capabilities, incorporate multimodal data, and improve on few-shot and zero-shot learning. However, how systems will address challenges such as bias, factual consistency, ethical implications of automating machines, will be important to their reliability and societal acceptance. Comprehensive evaluation frameworks and cross task annotated datasets show that there is a need for refining methodologies to conform to real world requirements.

Full realization of summarization and QA systems will require interdisciplinary collaboration. Some contributions focus on the technical aspects of this field, while others address ethical and practical considerations.

By fostering collaboration among researchers, ethicists, and practitioners, the research community can develop solutions that are not only technologically innovative but also equitable, transparent, and impactful. As summarization and QA boundaries expand, these systems will play a crucial role in determining how information is retrieved, understood, and leveraged.

A key challenge in this field is the need for robust methods to mitigate biases in automated summarization and QA systems. This can be addressed through fairness-aware training techniques and diverse dataset curation to ensure balanced and representative model training. Another important challenge is enhancing the interpretability of neural models, where techniques such as attention visualization and explainability frameworks can provide deeper insights into model decision-making. Additionally, developing hybrid approaches that effectively integrate traditional and neural methods is essential for improving accessibility and usability, which can be facilitated through standardized toolkits and model documentation. Future research should also prioritize multilingual and low-resource language processing to ensure inclusivity in AI applications. Finally, a systematic quantitative analysis of algorithm performance across various tasks and datasets would be a valuable extension of this work, offering deeper insights into the comparative strengths and weaknesses of different techniques.

REFERENCES

1. H. Asgari, B. Masoumi, and O. S. Sheijani, *Automatic text summarization based on multi-agent particle swarm optimization*, 2014 Iran. Conf. Intell. Syst. ICIS 2014, pp. 0–4, 2014, doi: 10.1109/IranianCIS.2014.6802592.
2. V. Bolotova, V. Blinov, Y. Zheng, W. B. Croft, F. Scholer, and M. Sanderson, *Do People and Neural Nets Pay Attention to the Same Words: Studying Eye-tracking Data for Non-factoid QA Evaluation*, Int. Conf. Inf. Knowl. Manag. Proc., pp. 85–94, 2020, doi: 10.1145/3340531.3412043.
3. A. T. R. Devi, K. J. Sathick, A. A. A. Khan, and L. A. Raj, *A novel framework using zero shot learning technique for a non-factoid question answering system*, Int. J. Web-Based Learn. Teach. Technol., vol. 16, no. 6, pp. 1–13, 2021, doi: 10.4018/IJWLTT.20211101.oa12.
4. L. W. Y. Yang et al., *Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study*, Front. Public Heal., vol. 11, 2023, doi: 10.3389/fpubh.2023.1063466.
5. L. Yao et al., *Accelerating the Discovery of Anticancer Peptides through Deep Forest Architecture with Deep Graphical Representation*, Int. J. Mol. Sci., vol. 24, no. 5, 2023, doi: 10.3390/ijms24054328.
6. C. Li, Y. Liu, F. Liu, L. Zhao, and F. Weng, *Improving multi-documents summarization by sentence compression based on expanded constituent parse trees*, EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 691–701, 2014, doi: 10.3115/v1/d14-1076.
7. M. Lewis et al., *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 7871–7880, 2020, doi: 10.18653/v1/2020.acl-main.703.
8. K. Mohiuddin et al., *Attention Is All You Need*, Int. Conf. Inf. Knowl. Manag. Proc., no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.
9. T. Uçkan and A. Karcı, *Extractive multi-document text summarization based on graph independent sets*, vol. 21, pp. 145–157, 2020, doi: 10.1016/j.eij.2019.12.002.
10. M. Sarrouti and S. Ouatik El Alaoui, *SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions*, Artif. Intell. Med., vol. 102, no. November 2019, p. 101767, 2020, doi: 10.1016/j.artmed.2019.101767.
11. S. Mitrović, D. Andreoletti, and O. Ayoub, *ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text*, pp. 1–11, 2023.
12. A. Shrestha and A. Mahmood, *Review of deep learning algorithms and architectures*, IEEE Access, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.
13. R. S. Roy and A. Anand, *Multi-Hop Question Answering*, 2022, doi: 10.1007/978-3-031-79512-1_11.
14. P. Lewis et al., *Retrieval-augmented generation for knowledge-intensive NLP tasks*, Adv. Neural Inf. Process. Syst., vol. 2020-Decem, 2020.
15. S. Mohammed, *A Survey of Text Summarization Extractive Techniques*, J. Emerg. Technol. Web Intell., vol. 5, no. 1, p. 1, 2013, doi: 10.4304/jetwi.5.1.1-1.
16. B. Mutlu, E. A. Sezer, and M. A. Akcayol, *Multi-document extractive text summarization: A comparative assessment on features*, Knowledge-Based Syst., vol. 183, 2019, doi: 10.1016/j.knosys.2019.07.019.
17. R. Mihalcea and P. Tarau, *TextRank: Bringing order into texts*, Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2004 - A Meet. SIGDAT, a Spec. Interes. Gr. ACL held conjunction with ACL 2004, vol. 85, pp. 404–411, 2004.
18. A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, *Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model*, ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 1074–1084, 2020, doi: 10.18653/v1/p19-1102.
19. V. Lopez, V. Uren, M. Sabou, and E. Motta, *Is question answering fit for the semantic web?: A survey*, Semant. Web, vol. 2, no. 2, pp. 125–155, 2011, doi: 10.3233/SW-2011-0041.
20. N. K. Tran and C. Niederée, *A Neural Network-based Framework for Non-factoid Question Answering*, Web Conf. 2018 - Companion World Wide Web Conf. WWW 2018, vol. 2018-Janua, pp. 1979–1983, 2018, doi: 10.1145/3184558.3191830.

21. W. Lei, X. Jin, Z. Ren, X. He, M. Y. Kan, and D. Yin, *Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures*, ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 1437–1447, 2018, doi: 10.18653/v1/p18-1133.
22. L. Wang and W. Ling, *Neural network-based abstract generation for opinions and arguments*, 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf., pp. 47–57, 2016, doi: 10.18653/v1/n16-1007.
23. Y. Yu, W. Zhang, K. Hasan, M. Yu, B. Xiang, and B. Zhou, *End-to-End Answer Chunk Extraction and Ranking for Reading Comprehension*, no. 1, 2016.
24. R. M. Alguliev, R. M. Aliguliyev, and M. S. Hajirahimova, *GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy*, Expert Syst. Appl., vol. 39, no. 16, pp. 12460–12473, 2012, doi: 10.1016/j.eswa.2012.04.067.
25. E. Canhasi and I. Kononenko, *Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization*, Expert Syst. Appl., vol. 41, no. 2, pp. 535–543, 2014, doi: 10.1016/j.eswa.2013.07.079.
26. D. Bollegala, N. Okazaki, and M. Ishizuka, *A bottom-up approach to sentence ordering for multi-document summarization*, COLING/ACL 2006 - 21st Int. Conf. Comput. Linguist. 44th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., vol. 1, no. July, pp. 385–392, 2006, doi: 10.3115/1220175.1220224.
27. J. P. Qiang, P. Chen, W. Ding, F. Xie, and X. Wu, *Multi-document summarization using closed patterns*, Knowledge-Based Syst., vol. 99, pp. 28–38, 2016, doi: 10.1016/j.knosys.2016.01.030.
28. G. Erkan and D. R. Radev, *LexRank: Graph-based lexical centrality as salience in text summarization*, J. Artif. Intell. Res., vol. 22, pp. 457–479, 2004, doi: 10.1613/jair.1523.
29. X. Wan and J. Xiao, *Single document keyphrase extraction using neighborhood knowledge*, Proc. Natl. Conf. Artif. Intell., vol. 2, pp. 855–860, 2008.
30. T. Mori, M. Sato, M. Ishioroshi, Y. Nishikawa, S. Nakano, and K. Kimura, *A Monolithic Approach and a Type-by-Type Approach for Non-Factoid Question-answering*, Proc. NTCIR-6 Work. Meet., 2007.
31. M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara, *A System for Answering Non-Factoid Japanese Questions by Using Passage Retrieval Weighted Based on Type of Answer*, Proc. NTCIR-6 Work. Meet., pp. 477–482, 2007.
32. H. Shima and T. Mitamura, *JAVELIN III: Answering Non-Factoid Questions in Japanese*, Proc. NTICIR-6 Work., pp. 464–468, 2007.
33. J. Fukumoto, *Question Answering System for Non-factoid Type Questions and Automatic Evaluation based on BE Method*, Proc. NTCIR-6 Work. Meet., pp. 441–447, 2007.
34. M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara, *Non-factoid Japanese question answering through passage retrieval that is weighted based on types of answers*, IJCNLP 2008 - 3rd Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., vol. 2, pp. 727–732, 2008.
35. D. R. Radev, H. Jing, M. Styś, and D. Tam, *Centroid-based summarization of multiple documents*, Inf. Process. Manag., vol. 40, no. 6, pp. 919–938, 2004, doi: 10.1016/j.ipm.2003.10.006.
36. T. Chen, H. Li, M. Kasamatsu, T. Utsuro, and Y. Kawada, *Developing a how-to tip machine comprehension dataset and its evaluation in machine comprehension by BERT*, Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 26–35, 2020, doi: 10.18653/v1/2020.fever-1.4.
37. Mass, Yotam and Roitman, Hagai and Erera, Shai and Rivlin, Oren and Weiner, Ben and Konopnicki, David, *A Study of BERT for Non-Factoid Question-Answering under Passage Length Constraints*, In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4510–4520, 2019, doi:10.18653/v1/P19-1441.
38. M. Worring, *NLQuAD: A Non-Factoid Long Question Answering Data Set*, pp. 1245–1255, 2021.
39. L. Vikraman, A. Montazeralghaem, H. Hashemi, W. B. Croft, and J. Allan, *Passage Similarity and Diversification in Non-factoid Question Answering*, ICTIR 2021 - Proc. 2021 ACM SIGIR Int. Conf. Theory Inf. Retr., pp. 271–280, 2021, doi: 10.1145/3471158.3472249.
40. V. Bolotova-Baranova, V. Blinov, S. Filippova, F. Scholer, and M. Sanderson, *WikiHowQA: A Comprehensive Benchmark for Multi-Document Non-Factoid Question Answering*, vol. 1, pp. 5291–5314, 2023, doi: 10.18653/v1/2023.acl-long.290.
41. Alec Radford and K. Narasimhan, *Improving language understanding by generative pre-training*, Homol. Homotopy Appl., vol. 9, no. 1, pp. 399–438, 2007, doi: 10.4310/HHA.2007.v9.n1.a16.
42. Brown, T. B. and Mann, B. and Ryder, N. and Subbiah, M. and Kaplan, J. and Dhariwal, P. and Neelakantan, A. and Shyam, P. and Sastry, G. and Askell, A. and Agarwal, S. and Herbert-Voss, A. and Krueger, G. and Henighan, T. and Child, R. and Ramesh, A. and Ziegler, D. M. and Wu, J. and Winter, C. and Hesse, C. and Chen, M. and Sigler, E. and Litwin, M. and Gray, S. and Chess, B. and Clark, J. and Berner, C. and McCandlish, S. and Radford, A. and Sutskever, I. and Amodei, D., *Language Models are Few-Shot Learners*, In *Advances in Neural Information Processing Systems*, vol. 2020, 2020.
43. C. Raffel, A. Roberts, M. Matena, and P. J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, vol. 21, pp. 1–67, 2020.
44. D. Khashabi et al., *UNIFIEDQA: Crossing format boundaries with a single QA system*, Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020, pp. 1896–1907, 2020, doi: 10.18653/v1/2020.findings-emnlp.171.
45. Y. Z. Song, Y. S. Chen, and H. H. Shuai, *Improving Multi-Document Summarization through Referenced Flexible Extraction with Credit-Awareness*, NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 1667–1681, 2022, doi: 10.18653/v1/2022.naacl-main.120.
46. A. See, P. J. Liu, and C. D. Manning, *Get to the point: Summarization with pointer-generator networks*, ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 1073–1083, 2017, doi: 10.18653/v1/P17-1099.
47. V. Karpukhin et al., *Dense passage retrieval for open-domain question answering*, EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.
48. T. Mikolov, *Efficient estimation of word representations in vector space*, arXiv Prepr. arXiv1301.3781, vol. 3781, 2013.
49. J. Devlin, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv Prepr. arXiv1810.04805, 2018.
50. Y. Liu and M. Lapata, *Text summarization with pretrained encoders*, EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 3730–3740, 2019, doi: 10.18653/v1/d19-1387.

51. S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.
52. D. Chen, *Reading Wikipedia to answer open-domain questions*, arXiv Prepr. arXiv1704.00051, 2017.
53. K. Knight and D. Marcu, *Statistics-Based Summarization - Step One: Sentence Compression*, Proc. 17th Natl. Conf. Artif. Intell. 12th Conf. Innov. Appl. Artif. Intell. AAAI 2000, pp. 703–710, 2000.
54. M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. Srinivasan, and D. Radev, *Graph-based neural multi-document summarization*, CoNLL 2017 - 21st Conf. Comput. Nat. Lang. Learn. Proc., pp. 452–462, 2017, doi: 10.18653/v1/k17-1045.
55. K. Ganesan, C. X. Zhai, and J. Han, *Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions*, Coling 2010 - 23rd Int. Conf. Comput. Linguist. Proc. Conf., vol. 2, no. August, pp. 340–348, 2010.
56. A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, *Automatic sentiment-oriented summarization of multi-documents using soft computing*, Soft Comput., vol. 23, no. 20, pp. 10551–10568, 2019, doi: 10.1007/s00500-018-3653-4.
57. T. N. Kipf and Max Welling, *SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS*, Iclr, pp. 1–11, 2017.
58. K. Xu, S. Jegelka, W. Hu, and J. Leskovec, *How powerful are graph neural networks?*, 7th Int. Conf. Learn. Represent. ICLR 2019, pp. 1–17, 2019.
59. R. Ferreira et al., *A multi-document summarization system based on statistics and linguistic treatment*, Expert Syst. Appl., vol. 41, no. 13, pp. 5780–5787, 2014, doi: 10.1016/j.eswa.2014.03.023.
60. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, no. Mlm, pp. 4171–4186, 2019.
61. Hadi, M. Usman and others, *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*, Authorea Preprint, 2024.
62. Rawte, Vishal, *A Survey of Hallucination in 'Large' Foundation Models*, arXiv preprint arXiv:2309.06555, 2023.
63. I. Lauriola, A. Lavelli, and F. Aiolli, *An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools*, Neurocomputing, vol. 470, no. xxxx, pp. 443–456, 2022, doi: 10.1016/j.neucom.2021.05.103.
64. A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, *A survey of the recent architectures of deep convolutional neural networks*, Artif. Intell. Rev., vol. 53, no. 8, pp. 5455–5516, 2020, doi: 10.1007/s10462-020-09825-6.
65. S. Shaier, K. Bennett, L. Hunter, and K. Kann, *Emerging Challenges in Personalized Medicine: Assessing Demographic Effects on Biomedical Question Answering Systems*, pp. 540–550, 2024, doi: 10.18653/v1/2023.ijcnlp-main.36.
66. Deroy, Alexis and Maity, Sandeep, *Questioning Biases in Case Judgment Summaries: Legal Datasets or Large Language Models?*, arXiv preprint arXiv:2312.01423, 2023.
67. S. Zhang and P. Kuhn, *Understanding Algorithmic Bias in Job Recommender Systems: An Audit Study Approach*, no. 17, 2022.
68. S. Matthews, J. Hudzina, and D. Sepehr, *Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings*, Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022, vol. 36, pp. 12026–12033, 2022, doi: 10.1609/aaai.v36i11.21461.
69. S. Dori-Hacohen et al., *Fairness via AI: Bias Reduction in Medical Information*, vol. 1, no. 1, pp. 1–5, 2021.
70. T. Scialom et al., *QuestEval: Summarization Asks for Fact-based Evaluation*, EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 6594–6604, 2021, doi: 10.18653/v1/2021.emnlp-main.529.
71. J. Liu, Z. Shi, and A. Lipani, *SummEQuAL: Summarization Evaluation via Question Answering using Large Language Models*, Proc. 2nd Work. Nat. Lang. Reason. Struct. Explan. (@ACL 2024), pp. 46–55, 2024.
72. S. Kapoor et al., *REFORMS: Consensus-based Recommendations for Machine-learning- based Science*, Sci. Adv., vol. 10, no. 18, pp. 1–17, 2024, doi: 10.1126/sciadv.adk3452.
73. Stream, *How to Achieve a 9ms Inference Time for Transformer Models*, GetStream.io, 2023.
74. Z. Li et al., *DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization*, Proc. Annu. Meet. Assoc. Comput. Linguist., vol. 2, pp. 203–211, 2022, doi: 10.18653/v1/2022.acl-short.22.
75. LeewayHertz, *Hybrid AI: Components, Applications, Use Cases and Development*, LeewayHertz Blog, 2025.
76. Yang, Soyeon and Kim, Daehyun and Lee, Seungwon, *HYLR-FO: Hybrid Approach Using Language Models and Rule-Based Systems for On-Device Food Ordering*, Electronics, vol. 14, no. 4, pp. 775, 2025, doi:10.3390/electronics14040775.
77. C. DeChant, I. Akinola, and D. Bauer, *Learning to summarize and answer questions about a virtual robot's past actions*, Auton. Robots, vol. 47, no. 8, pp. 1103–1118, 2023, doi: 10.1007/s10514-023-10134-4.
78. L. Korycki and B. Krawczyk, *Class-incremental experience replay for continual learning under concept drift*, IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work., pp. 3644–3653, 2021, doi: 10.1109/CVPRW53098.2021.00404.
79. Zhang, Sisi and Liu, Jianwei and Zuo, Xin, *Adaptive Online Incremental Learning for Evolving Data Streams*, arXiv preprint arXiv:2201.01633, 2022.
80. F. Lyu et al., *Overcoming Domain Drift in Online Continual Learning*, vol. 14, no. 8, pp. 1–13, 2024.
81. J. Kirkpatrick et al., *Overcoming catastrophic forgetting in neural networks*, Proc. Natl. Acad. Sci. U. S. A., vol. 114, no. 13, pp. 3521–3526, 2017, doi: 10.1073/pnas.1611835114.
82. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, *Meta-Learning with Memory-Augmented Neural Networks*, in 33rd International Conference on Machine Learning, ICML 2016, 2016, pp. 2740–2751.
83. Fan, Yuchen, *Research on the Online Update Method for Retrieval-Augmented Generation (RAG) Model with Incremental Learning*, arXiv preprint arXiv:2501.07063, 2025.
84. R. Krishnan, P. Khanna, and O. Tickoo, *Enhancing Trust in Large Language Models with Uncertainty-Aware Fine-Tuning*, pp. 1–22, 2024.
85. Ms-johnalex and Ktoliver and Mcleanbyron, *Build Advanced Retrieval-Augmented Generation Systems — Microsoft Learn*, Microsoft Learn, 2024.
86. Zilliz, *What is the carbon footprint of NLP models?*, Zilliz, 2025.
87. S. Mehta, *How Much Energy Do LLMs Consume? Unveiling the Power Behind AI*, AdaSci. Accessed: Feb. 10, 2024.
88. G. Hinton, O. Vinyals, and J. Dean, *Distilling the Knowledge in a Neural Network*, pp. 1–9, 2015.

89. Jacob, Benoit and Kligys, Skirmantas and Chen, Bo and Zhu, Menglong and Tang, Matthew and Howard, Andrew and Adam, Hartwig and Kalenichenko, Dmitry, *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*, arXiv preprint arXiv:1712.05877, 2017.
90. Child, Rewon and Gray, Scott and Radford, Alec and Sutskever, Ilya, *Generating Long Sequences with Sparse Transformers*, arXiv preprint arXiv:1904.10509, 2019.
91. State of the Planet, *AI's Growing Carbon Footprint*, Columbia Climate School, 2023.
92. Bloomberg, *What is the Carbon Footprint of AI and Deep Learning?*, Bloomberg, 2023.
93. S. A. Budennyy et al., *eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI*, Dokl. Math., vol. 106, pp. S118–S128, 2022, doi: 10.1134/S1064562422060230.
94. Graves, Alex, *Adaptive Computation Time for Recurrent Neural Networks*, arXiv preprint arXiv:1603.08983, 2016.
95. D. Patterson et al., *Carbon Emissions and Large Neural Network Training*, pp. 1–22, 2021.
96. E. Strubell, A. Ganesh, and A. McCallum, *Energy and policy considerations for modern deep learning research*, AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 1393–13696, 2020, doi: 10.1609/aaai.v34i09.7123.
97. J. Wang et al., *A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks*, vol. 18, no. 9, pp. 1–21, 2024.
98. Lupascu, Marian and Rogoz, Ana-Cristina and Stupariu, Mihai Sorin and Ionescu, Radu Tudor, *Large Multimodal Models for Low-Resource Languages: A Survey*, arXiv preprint arXiv:2502.05568, 2025.
99. G. Moro, N. Piscaglia, L. Ragazzi, and P. Italiani, *Multi-language transfer learning for low-resource legal case summarization*, Artif. Intell. Law, vol. 32, no. 4, pp. 1111–1139, 2023, doi: 10.1007/s10506-023-09373-8.
100. F. Le Bronnec et al., *LOCOST: State-Space Models for Long Document Abstractive Summarization*, EACL 2024 - 18th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf., vol. 1, pp. 1144–1159, 2024.
101. OpenAI, *GPT-4 Technical Report*, arXiv preprint arXiv:2303.08774, 2023.
102. Anil, Rohan and Dai, Andrew M. and Firat, Orhan and Johnson, Melvin and Lepikhin, Dmitry and Passos, Alexandre and Shakeri, Siamak and Taropa, Emanuel and Bailey, Paige and Chen, Zhifeng and others, *PaLM 2 Technical Report*, arXiv preprint arXiv:2305.10403, 2023.
103. M. Akter, E. Çano, E. Weber, D. Dobler, and I. Habernal, *A Comprehensive Survey on Legal Summarization: Challenges and Future Directions*, vol. 1, no. 1, 2025.
104. A. Mullick et al., *On The Persona-based Summarization of Domain-Specific Documents*, Find. Assoc. Comput. Linguist. ACL 2024, pp. 14291–14307, 2024.
105. Zhao, S. and Yang, Y. and Wang, Z. and He, Z. and Qiu, L. K. and Qiu, L., *Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely*, arXiv preprint arXiv:2401.12345, 2024.
106. R. Yang et al., *Retrieval-augmented generation for generative artificial intelligence in health care*, npj Heal. Syst., 2025, doi: 10.1038/s44401-024-00004-1.
107. D. Khurana, A. Koli, K. Khatter, and S. Singh, *Natural language processing: state of the art, current trends and challenges*, Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, 2023, doi: 10.1007/s11042-022-13428-4.
108. Kumar, Abhishek and Mittal, Trisha and Manocha, Dinesh, *MCQA: Multimodal Co-attention Based Network for Question Answering*, arXiv preprint arXiv:2004.12238, 2020.
109. S. Piva, C. Bonamico, C. Regazzoni, and F. Lavagetto, *A flexible architecture for ambient intelligence systems supporting adaptive multimodal interaction with users*, Ambient Intell., pp. 97–120, 2005.
110. Z. Wen, D. O'Neill, and H. Maei, *Optimal demand response using device-based reinforcement learning*, IEEE Trans. Smart Grid, vol. 6, no. 5, pp. 2312–2324, 2015, doi: 10.1109/TSG.2015.2396993.
111. A. Haghighi and L. Vanderwende, *Exploring content models for multi-document summarization*, NAACL HLT 2009 - Hum. Lang. Technol. 2009 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Conf., no. June, pp. 362–370, 2009, doi: 10.3115/1620754.1620807.
112. J. Wei et al., *Finetuned Language Models Are Zero-Shot Learners*, ICLR 2022 - 10th Int. Conf. Learn. Represent., pp. 1–46, 2022.
113. A. Jiang and A. Zubiaga, *Cross-lingual Offensive Language Detection: A Systematic Review of Datasets, Transfer Approaches and Challenges*, vol. 37, no. 4, 2023.
114. E. A. Stepanov et al., *Cross-language transfer of semantic annotation via targeted crowdsourcing: task design and evaluation*, Lang. Resour. Eval., vol. 52, no. 1, pp. 341–364, 2018, doi: 10.1007/s10579-017-9396-5.
115. Khalilia, Hadi and Otterbacher, Jahna and Bella, Gianluca and Noortyani, Reza and Darma, Suryanto and Giunchiglia, Fausto, *Crowdsourcing Lexical Diversity*, arXiv preprint arXiv:2410.23133, 2024.
116. T. Mori and M. Nozawa, *Multi-document summarization using a questionanswering engine*, Proc. 4th NTCIR Work., vol. 4, no. April 2003, pp. 305–320, 2004.
117. Haque, Md. Majharul and Pervin, Sharmin and Begum, Zubaida, *Literature Review of Automatic Multiple Documents Text Summarization*, International Journal of Innovation and Applied Studies, vol. 3, no. 1, pp. 121–129, 2013.
118. P. M. Mah, I. Skalna, and J. Muzam, *Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0*, Appl. Sci., vol. 12, no. 18, 2022, doi: 10.3390/app12189207.
119. S. Ghosh, S. Ghosh, and D. Das, *Complexity Metric for Code-Mixed Social Media Text*, Comput. y Sist., vol. 21, no. 4, pp. 693–701, 2017, doi: 10.13053/CyS-21-4-2852.
120. A. A. Y. Al-Qahtani and S. E. Higgins, *Effects of traditional, blended and e-learning on students' achievement in higher education*, J. Comput. Assist. Learn., vol. 29, no. 3, pp. 220–234, 2013, doi: 10.1111/j.1365-2729.2012.00490.x.
121. P. E. Miller et al., *Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database*, J. Card. Fail., vol. 25, no. 6, pp. 479–483, 2019, doi: 10.1016/j.cardfail.2019.01.018.
122. Y. Cheng, Y. Liu, T. Chen, and Q. Yang, *Federated learning for privacy-preserving AI*, Commun. ACM, vol. 63, no. 12, pp. 33–36, 2020, doi: 10.1145/3387107.
123. A. Dyda et al., *Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality*, Patterns, vol. 2, no. 12, p. 100366, 2021, doi: 10.1016/j.patter.2021.100366.

124. D. Luitse and W. Denkena, *The great transformer: Examining the role of large language models in the political economy of AI*, Big Data Soc., vol. 8, no. 2, 2021, doi: 10.1177/20539517211047734.
125. G. Shi et al., *A new communication paradigm: from bit accuracy to semantic fidelity*, pp. 1–8, 2021.
126. N. S. Leng2, R. C. M. Yusoff1, G. N. Samy1, and S. M. and Z. I. R. 1Advanced R. Ibrahim1,*, *Special Issue Efficiency and Scalability of Advanced Machine Learning and Optimization Methods for Real-World Applications*, J. Fundam. Appl. Sci., vol. 4, no. 1, pp. 9–10, 2018, doi: 10.13140/RG.2.2.19611.89128.
127. J. M. Ferreira et al., *Impact of usability mechanisms: An experiment on efficiency, effectiveness and user satisfaction*, Inf. Softw. Technol., vol. 117, no. November 2018, p. 106195, 2020, doi: 10.1016/j.infsof.2019.106195.
128. L. Alzubaidi et al., *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021, doi: 10.1186/s40537-021-00444-8.
129. L. Al-Alawi, J. Al Shaqsi, A. Tarhini, and A. S. Al-Busaidi, *Using machine learning to predict factors affecting academic performance: the case of college students on academic probation*, Educ. Inf. Technol., vol. 28, no. 10, pp. 12407–12432, 2023, doi: 10.1007/s10639-023-11700-0.
130. M. Islam and F. Shehzad, *A Prediction Model Optimization Critiques through Centroid Clustering by Reducing the Sample Size, Integrating Statistical and Machine Learning Techniques for Wheat Productivity*, Scientifica (Cairo)., vol. 2022, 2022, doi: 10.1155/2022/7271293.
131. Cui, Jiaxi and Zhang, Wentao and Tang, Jing and Tong, Xudong and Zhang, Zhenwei and Amie and Wen, Jing and Wang, Rongsheng and Wu, Pengfei, *AnyTaskTune: Advanced Domain-Specific Solutions through Task-Fine-Tuning*, arXiv preprint arXiv:2407.07094, 2024.
132. M. Gray et al., *Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science*, Clin. Pharmacol. Ther., vol. 115, no. 4, pp. 687–697, 2024, doi: 10.1002/cpt.3117.
133. Google Research, *Fairness Indicators: Scalable Infrastructure for Fair ML Systems*, Google Research Blog, 2019. Accessed: 2025-03-06.
134. R. K. E. Bellamy et al., *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*, IBM J. Res. Dev., vol. 63, no. 4–5, 2019, doi: 10.1147/JRD.2019.2942287.
135. P. Radanliev, *AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development*, Appl. Artif. Intell., vol. 39, no. 1, 2025, doi: 10.1080/08839514.2025.2463722.
136. E. Kallina and J. Singh, *Stakeholder Involvement for Responsible AI Development: A Process Framework*, Proc. 4th ACM Conf. Equity Access Algorithms, Mech. Optim. EAAMO 2024, 2024, doi: 10.1145/3689904.3694698.
137. M. Mitchell et al., *Model cards for model reporting*, FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar., no. Figure 2, pp. 220–229, 2019, doi: 10.1145/3287560.3287596.
138. C. Garbin and O. Marques, *Assessing Methods and Tools to Improve Reporting, Increase Transparency, and Reduce Failures in Machine Learning Applications in Health Care*, Radiol. Artif. Intell., vol. 4, no. 2, pp. 1–9, 2022, doi: 10.1148/ryai.210127.
139. V. Bengani, *Hybrid Learning Systems: Integrating Traditional Machine Learning with Deep learning Techniques*, no. May, pp. 0–122, 2024, doi: 10.13140/RG.2.2.34709.54248/1.
140. W. Jiang et al., *An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry*, Proc. - Int. Conf. Softw. Eng., pp. 2463–2475, 2023, doi: 10.1109/ICSE48619.2023.00206.
141. C. H. Nwokoye, M. J. P. Peixoto, A. Pandey, L. Pardy, M. Sukhai, and P. R. Lewis, *A Survey of Accessible Explainable Artificial Intelligence Research*, pp. 1–13, 2024.
142. K. S. Glazko et al., *An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility*, vol. 1, no. 1. Association for Computing Machinery, 2023, doi: 10.1145/3597638.3614548.