Received: 19 June 2026 | Revised: 13 August 2026 | Accepted: 14 September 2026 | Published online: 15 October 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA620210968

BON VIEW PUBLISHING

A Hybrid RAG-Based Framework for Improved Information Retrieval from Long Documents via Summarization and Question Answering

Ahmed Sami Jaddoa[1, 2,*], Jaber Karimpour[1] and Pedram Salehpour[3]

[1] *Department of Computer Science, University of Tabriz, Iran*

[2] *Department of Information Systems Management, University of Information Technology and Communications, Iraq*

[3] *Department of Computer Engineering, University of Tabriz, Iran*

Abstract: We present an integrated system, which is a mix of summarization as well as Question Answering (QA) to boost the performance regarding information retrieval (IR) for long and domain-intensive documents. This work creates a synergistic relationship between both tasks to enhance retrieval efficiency and quality of answers, unlike previous studies focusing separately on either summarization or QA. The framework proposed is comprised of three stages. First, it involves summarization that is used for condensing documents, maintaining their key semantic content. Second, it employs a retrieval-augmented generation (RAG) architecture, where the all-MiniLM-L6-v2 model is used for the retrieval stage, while T5 and BART-Large-CNN models are utilized for the generation stage. Lastly, answers generated are put into comparison with a baseline system which works with complete document processing without summarization. Results at the NIPS dataset reveal that integrating summarization with QA significantly improves each summary informativeness and answer accuracy. To investigate the impact of summarization inside the retrieval pipeline, two complementary eventualities are explored. First, Proactive Summarization as an Indexing Tool (Summarization-First) plays summarization earlier than the indexing and retrieval stages to reduce record size and improve retrieval performance while keeping semantic relevance. Second, Query-Guided Contextual Summarization (Retrieval-First Summarization) applies summarization after the retrieval level, wherein only the retrieved relevant passages are summarized in step with consumer question for generating more targeted and context-conscious solutions. These two scenarios enable a comparative analysis of the way the position of summarization within RAG pipeline impacts retrieval efficiency as well as solution era quality.

Keywords: SBERT, information retrieval, retrieval-augmented generation, question answering

1. Introduction

Nowadays, information retrieval (IR) is of high importance in individuals' daily lives since it is integrated into many helpful features, such as personal assistants, digital libraries, internet browsing, chatbots, along with question-answering systems. Finding and retrieval information associated with user's request is the main goal. The results are often ordered based on how comparable they are to user's query because several items could be relevant. Conventional systems of text retrieval are primarily dependent on matching terms between documents and queries at the beginning of the field of IR. Yet, lexical gaps, synonymy, and polysemy are problems with such term-based retrieval systems that may limit their efficacy [1].

The IR systems attempt to overcome the era of information overload where users attempt to obtain information more effectively and efficiently. Integrating large language models (LLMs) has radically changed the way IR systems are viewed, such as the addition of the LLM-generated data as new data source in IR systems, the substitution of passive retrieval to the proactive generation as basic paradigms, and the use of the LLM as a results evaluator of IR systems [2].

Important information is lost or overlooked because of the continuously increasing and huge size regarding textual corpora on internet. However, it takes a lot of time and effort for human professionals to summarize such resources. This means that the task must be automated. Natural language processing (NLP) can be defined as multidisciplinary topic of study combining elements and methods from linguistics, computer science, and AI.

***Correspondence:** Ahmed Sami Jaddoa, Department of Computer Science, University of Tabriz, Iran and Department of Information Systems Management, University of Information Technology and Communications, Iraq. Email: ahmed.sami@uoitc.edu.iq

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

It handles developing procedures that effectively and meaningfully evaluates enormous volumes of textual data. The technique of automatically producing concise as well as fluid summary which encapsulates the key concepts and subjects of one or more papers is known as text summarization, basic NLP subtask [3, 4]. One of the major significant applications is TS. The user might quickly identify the more relevant documents without having to read the entire document thanks to TS systems, which provide succinct information from source content. Various systems, such as search engines, and a lot of individuals, including scholars, can benefit from TS systems. They enable them to do their duties more effectively [5].

Depending on the required type of summary and input, there are various summarization types. Abstractive and extractive summarizations are the two main categories of summarizations. Text segments, just like the most sizable terms or sentences, are obtained from the supply text without being altered and concatenated for generating the summary in extractive summarization. New sentences created through reformulating or paraphrasing extracted content make up an abstract textual content summary type. As abstractive summaries extra evidently deliver the which means and content regarding unique textual content, they typically resemble human-written summaries [6]. Abstractive summarization comprises synthesizing information from original documents, rephrasing it, along with producing a summary like those created manually, extractive summarization focuses on finding as well as choosing informative sentences for summary generation [7].

The exercise of extracting formerly unknown data through mining facts from textual assets is referred to as text mining (TM), and it was to start with delivered through Feldman. It uses methods and tactics from several specialized fields, such as information extraction (IE), natural language processing (NLP), information mining (DM), machine studying (ML), data retrieval (IR), and records. IR, IE, and DM are the three key levels regarding textual content mining. IR is a method that essentially corresponds to the congregation of information following filtering the relevant documents. Explicit information regarding predetermined categories of relation and entities is preferred by IE approaches. DM is a method that finds subtle correlations between well-known refined data [8].

The purpose regarding QA systems is to locate responses to natural language questions in large collections of documents. IR is usually used to do that. NLP in addition to linguistic sources are frequently 4 utilized by QA structures. In addition, QA structures are designed for automatically providing specific, pertinent, and succinct answers to their questions, that are normally made in herbal language. Those systems play a pretty essential component in the bridging of the conversation gaps between machine intelligence and human language comprehension, permitting the interplay of human beings with computers greater intuitively and certainly. The QA systems are used in many fields such as IR, education, customer support, virtual assistants among others [9].

The rapid growth of scientific literature has made accessing accurate information from long documents increasingly difficult. Although retrieval-augmented generation (RAG) has significantly improved question answering by combining information retrieval with large language models, its effectiveness is still limited when processing lengthy scientific documents. Existing retrieval methods often return large amounts of redundant or partially relevant text, increasing the context provided to the generation model and reducing the accuracy, coherence, and efficiency of the generated answers. Extractive summarization presents an effective mechanism for lowering record length even as maintaining critical facts. However, most current procedures rent summarization either as an impartial mission or as a hard and fast preprocessing step, without investigating how different summarization strategies affect retrieval and question answering overall performance inside a RAG framework. In addition, question-independent summaries may also omit information required to reply to specific consumer questions, while retrieval-only methods often introduce pointless context that degrades era best. Furthermore, medical documents incorporate numerous domain-particular abbreviations and complicated terminology which can lessen semantic matching between consumer queries and file content, negatively affecting retrieval accuracy. The selection of a suitable chew length additionally remains a crucial challenge, as overly large chunks introduce beside the point statistics, even as excessively small chunks may additionally fragment the contextual proof required for correct answer era. Therefore, there may be a need for an integrated RAG framework that effectively combines semantic retrieval, extractive summarization, and query answering while investigating alternative summarization strategies for lengthy scientific documents. Such a framework should lessen redundant context, hold question-applicable statistics, enhance retrieval effectiveness, and generate accurate and coherent answers. The proposed framework does not include an OCR module. The input consists of textual research documents stored in CSV format, where the text is already digitally available. Therefore, OCR performance is outside the scope of this work.

The main contributions of the present paper to make the extraction of answers semantically efficient, this paper uses empirical methodology through comparing two structural pipelines which vary in the information flow sequence. The main research question is to determine the best positioning of extractive summarization as a part of the RAG pipeline. Although the raw data are prepared by initial processing, a more important question emerges regarding the temporal placement regarding summarization: can it be more effective to condense the information space before retrieval, to guarantee the speed of computation, or to defer the summarization to after retrieval to maintain query-specific nuances? As a result, two different methodological scenarios have been developed to analyze text condensation vs. retrieval precision trade-off in the following way:

Scenario 1: Proactive Summarization as an Indexing Tool (Summarization-First Approach)

Summarization in this case is considered a pre-processing step for shrinking the size of data prior to the search. It employs a proactive intensified summarization process, in which the whole database is subjected to an extractive summarization process right after the initial processing stage. The purpose of this step is to convert large documents to small knowledge representations that distill the key concepts. On these summaries, an (IR) search is done based on the response to the query of the user. The competitive advantage of this scenario is in its efficiency of retrieval and speed since the system searches for a narrow semantic space, decreasing the number of computations and in-formation noise prior to the query arrives.

Scenario 2: Query-Guided Contextual Summarization (Retrieval-First Summarization)

In this case, summarization is executed as a put-up processing step, in which it only operates on documents which have been decided to be relevant to the question. It is based totally on the retrieval guided contextual summarization principle. The

procedure starts with trying to find relevant documents (IR) in the complete text corpus primarily based on the query of the user. After the identification of the candidate documents, only this group of documents is presented to summarization. The contextual fidelity is maximized with this systemic approach where the summary is done with respect to the query presented so that the system concentrates on information that directly serves as answer (RAG). This order protects the science content against the loss of minor details that could be lost in proactive summarization thereby increasing the quality and consistency of the final answer.

The primary aim of this dissertation is to develop a hybrid retrieval-augmented generation (RAG) framework that enhances information retrieval and question answering from long scientific documents by integrating semantic retrieval with extractive summarization. The proposed framework seeks to lessen redundant contextual facts even as maintaining the maximum applicable evidence required for correct, coherent, and efficient solution generation. To attain this goal, the dissertation pursues the subsequent objectives:

1) Develop a semantic retrieval mechanism able to be figuring out the maximum applicable passages from lengthy scientific files using Sentence-BERT (SBERT).
2) Design and investigate extractive summarization strategies in the RAG framework:
    a. An offline summarization approach based on SBERT and TextRank for producing compact record representations earlier than retrieval; and
    b. An online question-centered summarization technique that selects the most applicable contextual information after retrieval in line with the consumer's question.

The remainder of the present paper is prepared as follows. Section 2 reviews the associated literature. Section 3 introduces the heritage statistics required for understanding the proposed technique. Section 4 describes the substances and methods used in this study. Section five provides the assessment results and their discussion. Finally, Section 6 concludes the paper, and Section 7 outlines destiny studies instructions.

2. Literature Review

Recent studies associated with summarization and QA are reviewed on this segment in a manner that gives an essential background for positioning the proposed gadget.

2.1. Text summarization of long documents

The research is carried out by Hew Zi et al. [10]. They identified the best ML model regarding news article summarization as well as the best text summarization. CNN/Daily Mail database had to be used to compare TS using specific classifiers.

Guo et al. [11] proposed LongT5, an extension of the T5 architecture designed to efficiently process long input sequences by introducing a novel Transient Global (TGlobal) attention mechanism and adopting the PEGASUS pre-training strategy. The model permits green long document know-how while decreasing the computational value related to general self-interest. LongT5 become evaluated on several long-document summarization and question answering benchmarks, consisting of arXiv, PubMed, BigPatent, MediaSum, Natural Questions, and TriviaQA, achieving modern day performance on a couple of datasets. These consequences show the effectiveness of LongT5 for lengthy-context summarization and question answering tasks, making it a sturdy foundation for packages regarding long files.

Jiang et al. [12] proposed Long RAG, a RAG framework designed to improve query answering over long files by means of introducing long retrieval units and long-context readers. Instead of retrieving numerous short textual content chunks, the framework retrieves a small wide variety of semantically entire lengthy gadgets (about 4K tokens or more), reducing retrieval complexity while keeping contextual data. Experimental opinions on Natural Questions, HotpotQA, Qasper, and MultiFieldQA-en prove that LongRAG finished aggressive or superior performance compared with conventional RAG processes, considerably improving retrieval great and query-answering accuracy without requiring extra model first-rate-tuning.

Kietikul Jearanaitanakij al. [13] intended to develop a particular Thai news summarization system that would allow abstractive summaries to be obtained as well as subsequently produced, not to mention highlighting the fundamental concepts. They proposed a two-part model. Regarding the first component, contiguous region containing important sentences has been extracted using Text-Rank approach. In the second, optimized mBART model has been used as LLM. Using the extracted content, abstractive summary was produced. To put it another way, the suggested method first determines a significant news area before sending it to mBART for summarization. The technique resulted in news summaries preserved the syntactic style of natural Thai language and capturing important information.

Bailly et al. [14] proposed a divide-and-summarize approach, known as the Map method, to improve text summarization using small language models (SLMs). The method splits a document into smaller segments, summarizes each segment independently, and, after that, generates a final summary from the intermediate summaries. Experimental results on simulated texts and 587 scientific articles showed that the Map approach had outperformed the traditional Stuff method by improving information retention and reducing "Lost in the Middle" problem. Moreover, the approach enabled SLMs to achieve performance comparable to, and sometimes exceeding, that of GPT-4o in summarization tasks.

2.2. Question answering using RAG

Aljohani and Alsanoosy [15] proposed a hybrid RAG framework to enhance scientific question answering systems by means of combining sparse retrieval (BM25) and dense retrieval (MedCPT) within a modular structure. The framework became designed to address common demanding situations in healthcare applications, together with old understanding, hallucinations, and inadequate area model in large language models (LLMs).

Pujiono al. [16] used retrieval-augmented generation using vector database for evaluating the produced chatbot's performance. This research examined the efficacy of existing LLMs in addressing regulatory concerns pertaining to the public service organizations. LLM used vector database for assessing and responding to questions based on cosine similarity scores.

Binita Saha et al. conducted another research [17]. They introduced new architecture for RAG systems that use target corpus to reinforce question-answering tasks. The models, that is (LLMs), have revolutionized human-like text production as well as analysis. Nevertheless, such models depend on pretrained data and don't offer real-time updates if they are not combined with live data tools. RAG enhances LLMs through generating

contextually appropriate responses through the use of databases as well as internet resources.

The research of Moreno Cediel et al. [18] attempted to find and create better models regarding Spanish Answer Extraction (AE) as well as Question Generation (QG) problems using answer aware technique. There are three distinct datasets used for optimizing 3 multi-lingual models (mT0-base, mT5-base, and BLOOMZ-560-M): SQAC, Spanish dataset; the Spanish version regarding SQuAD dataset; and its combination (SQuAD + SQAC), which yielded somewhat superior research findings.

Meng Wenjun et al. [19]. employed QA system that has been created by using RAG for focusing on dangers to human health. A total of 300 excellent question-answer pairs from six subfields have been generated, based on the introduction of the QA pair generating method. QA system has been created using both advanced and naïve RAG techniques.

Lin et al. [20] proposed a RAG-based question answering system that integrates expertise graphs with large language models to improve response accuracy and decrease hallucination problems. The proposed framework employed a Light RAG-primarily based dual-layer retrieval mechanism, in which both entity-stage and courting-stage facts have been extracted from a understanding graph and integrated into the solution technology procedure.

SeongKu Kang et al. [21] provide a collection of queries which completely covers the concepts in the document. Coverage-based Query Set Generation (CCQGen) framework has been introduced. The capacity of (CCQGen) of the adaptive changing of the generation as a response to the queries that have been created previously was one of its key differentiators. The terms that were not adequately covered in the previous queries have been found by SeongKu Kang et al. and utilized as prerequisites to create new queries.

Hessa A. Alawwad et al. [22] proposed PLRTQA, a framework that combines RAG with the Llama-2 huge language model to enhance textual textbook question answering. The proposed technique addresses challenges related to long contexts, susceptible reasoning, and out-of-domain know-how by retrieving applicable textbook content and integrating it into answer era process. Experimental effects on CK12-QA dataset confirmed advanced performance, reaching a typical accuracy of eighty-four.24% on the take a look at set and outperforming numerous current kingdom-of-the art techniques.

Chengke Wu *etal*. [23] presented a new paradigm called RAG for Construction Management (RAG4CM). The latter consisted of three parts: (i) a pipeline which parsed project papers to hierarchical structures for creating knowledge pool; (ii) innovative RAG search techniques; (iii) a technique for discovering user preferences. To enhance RAG findings and granularity alignment, the first two components combined document-level hierarchical characteristics with raw material. Preference learning produces consistently better results as well as improve user-system interactions. Several experiments have been carried out as well as prototype system was created.

Although previous studies have investigated long-document summarization, document-grounded question answering, and RAG independently, few have integrated summarization as an intermediate representation to improve retrieval and question answering over long scientific documents. Furthermore, most studies evaluate their methods on benchmark datasets, whereas this work focuses on long scientific papers from the NIPS collection and proposes a hybrid RAG framework tailored for efficient information retrieval and answer generation.

A research gap exists in the absence of an integrated system that combines between text summarization and question answering in long documents. The proposed studies address those challenges with the aid of focusing on the subsequent targets:

1) Improving semantic retrieval from long scientific documents.
2) Reducing redundant information through extractive summarization at the same time as retaining essential content.
3) Investigating both offline and online summarization strategies inside an RAG framework.
4) Enhancing solutions of high-quality through question-targeted contextual summarization.
5) Improving semantic matching through abbreviation growth and appropriate textual content chunking.

3. Background

3.1. Text summarization

Text summarization may be distinct as a subset concerning NLP targeting creating short summaries from lengthy texts. Put greater genuinely, it includes condensing huge pieces of text into more concise and insightful summaries [24]. The TS system's purpose is to supply precis for sure files which get rid of redundant fabric and consist of all vital facts [25]. Summaries could take many awesome formats. It is possible to classify summarization techniques based on what they produce and how they produce it. One document or multiple documents could be used for extracting a summary. Making a summary from single document is referred to as single-document summarization. However, constructing single summaries from multiple documents on a similar topic is referred to as multi-document summarization [26].

3.2. IR for QA

IR can be defined as the science and art of looking for information in documents, for documents themselves, information about documents, databases, etc., based on the needs of the user. Retrieving what is helpful and eliminating what is not is the primary goal [27]. IR has been crucial in assisting QA tasks and in managing and processing the rapidly increasing amounts of big data. IR has historically concentrated on obtaining pertinent material from large datasets or corpuses, a task that has grown more difficult due to the exponential expansion of digital data. Effectively querying massive corpora to obtain relevant results is the focus of IR systems [28]. One of the fundamental fields of IR in general and NLP in particular is QA. Both the enterprise and the scientific community have recently paid greater attention to it. QA systems aim to automatically respond to human inquiries in natural languages based on context [29].

3.3. NLP and LLM

NLP and AI have greatly contributed to the evolution of QA systems. The use of LLMs has transformed the NLP and transformed the way we typically engaged with digital content. They can analyze and generate text that is more like human language, which became an inherent part of multiple industries. Although rules-based and retrieval-based QA systems are widely used, they are unable to comprehend complicated questions and deliver contextually appropriate answers. LLMs are powerful AI systems that have been educated on the tremendous text datasets for generating humanlike language. They normally applied in regions

like content material introduction, gadget translation, chatbots, and summarizing data throughout a number of sectors [30]. Text generation, category, and summarization are among those responsibilities and may be nine considered examples of the essential elements of language knowledge and language era structures. With the demands of advanced NLP solutions still surging in many areas such as IR, content generation, and sentiment analysis, the necessity of stable and effective pertained models has never been greater [31].

3.4. RAG technique

Generation and retrieval models are combined to create RAG. It uses LLM for generating text in response to commands and integrates information from a different system of retrieval for improving the quality of output and relevance of context. In case when query is received, RAG determines the most relevant pieces through the calculation of semantic similarity between the query and every piece that already exists. LLM is then responsible for using the obtained pieces for responding to the query. In addition to limiting the range of responses, which lowers the possibility regarding LLM hallucinations as well as improving the quality of responses, RAG reduces the data handled through LLMs and improves the granularity matching between queries and contents [32].

The three main components are indexing, which involves storing documents in database; retrieval, which involves retrieving relevant documents in response to user query; and generation, which involves LM producing responses depending on input prompt, i.e., the combination regarding user query as well as relevant documents. For enabling effective and reliable retrieval of relevant information in RAG framework, indexing is used to transform raw text into a searchable format. Several crucial procedures, including data preprocessing, vectorization, text chunking, and index storage, are used to accomplish this.

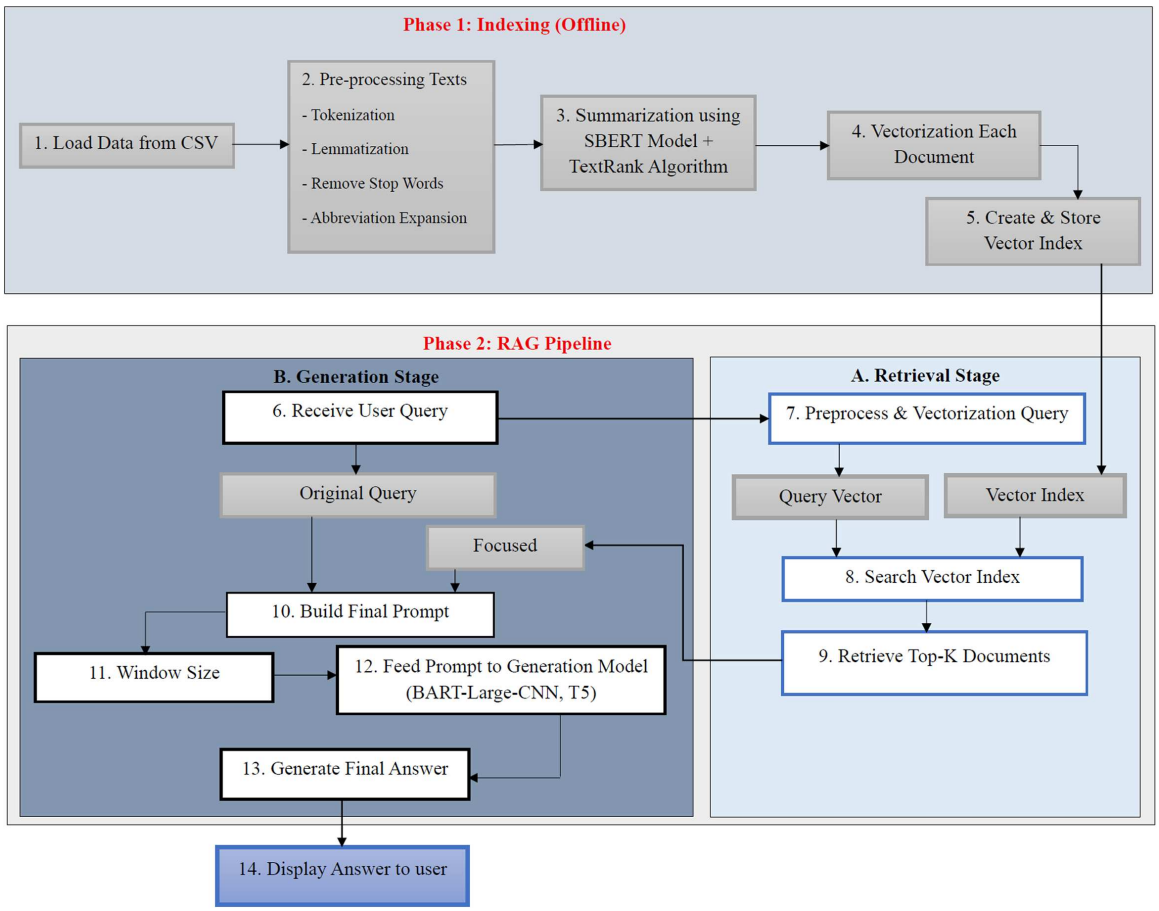
In order model's contextual knowledge, the retrieval phase finds and retrieves the most relevant chunks of text from indexed database. When determining the precision and applicability regarding generated response, retrieval efficiency as well as quality are crucial. The task of synthesizing answers depending on the retrieved data and merging it with LLMs is carried out by RAG generation step. Ensuring that the produced output is consistent, correct, and contextually appropriate is the primary goal of this step [33, 34].

4. Materials and Methods

4.1. System overview

The suggested work is a systematic framework, which combines summarization, IR, and RAG in QA. The methodology will focus on enhancing efficiency and contextual accuracy regarding information extraction as well as answer generation. It is

Figure 1
Overall framework of the proposed summarization-first system



per-formed in three major steps, starting with extractive summarization, then IR based document selection based on the user question, and finally the final answer utilizing RAG approach. To maximize semantic efficiency of answer extraction, the proposed study will use an empirical approach to research where two structural pipelines (differentiated by the sequence of the information flow) will be compared. The main research question is to determine the optimal positioning regarding extractive summarization in RAG pipeline. Although the raw data is prepared by initial processing, a big question then emerges, about temporal placement regarding summarization, i.e., whether it is better to condense the information space before the retrieval operation to allow the computation speed or to deter the summarization till after the retrieval operation to avoid losing query-specific nuances. As a result, there are two different methodological scenarios that are proposed to test the trade-off between the texts. condensation and retrieval preciseness, as follows:

Scenario 1 (offline): SBERT-based query-independent extractive summarization

SBERT (Sentence-BERT) may be effectively employed as a semantic embedding model inside a query-independent extractive summarization framework. In this state of affairs, the summarization manner no longer relies upon a user question or document retrieval. Instead, the objective is to pick out the most vital and representative sentences within a file based totally on their semantic relationships. SBERT is used to generate sentence embeddings, while cosine similarity and TextRank are employed to measure sentence relationships and rank sentences in line with their importance.

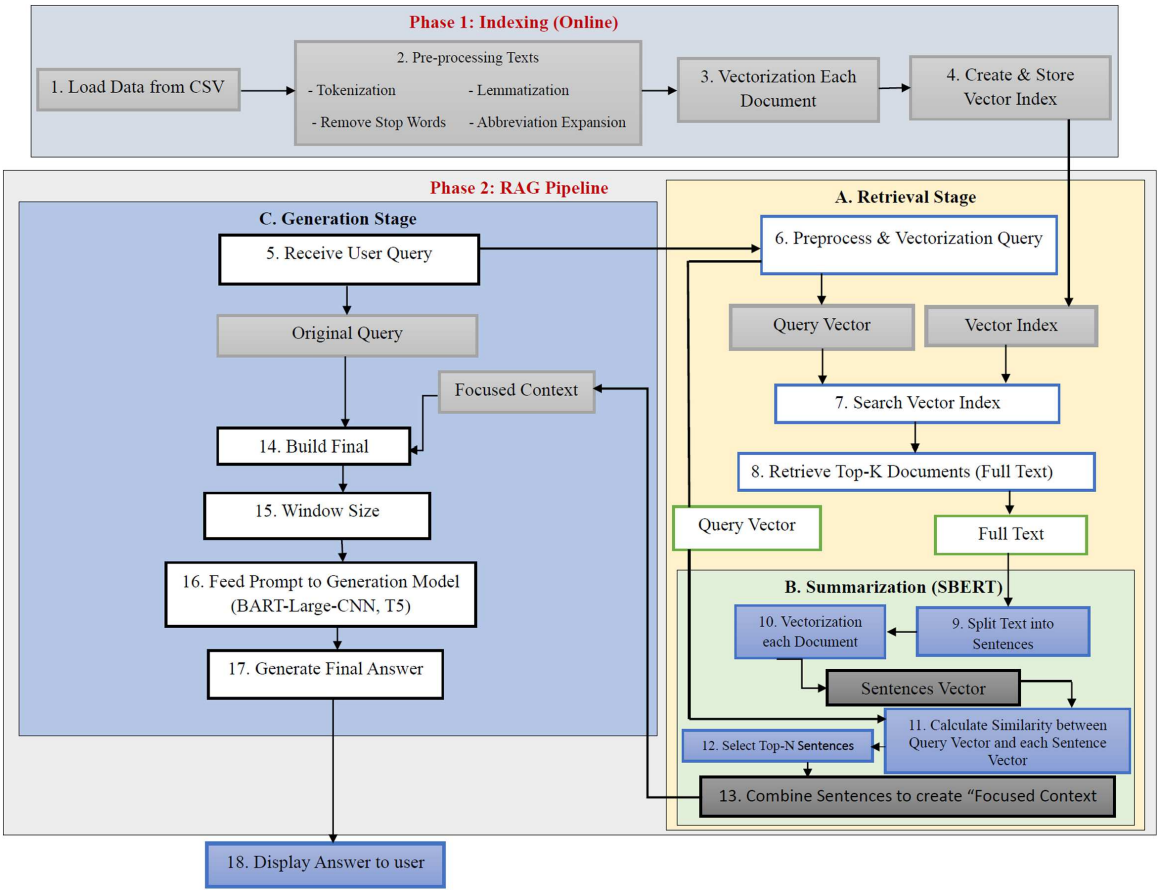
Scenario 2 (online): SBERT-based query-focused extractive summarization

After retrieving documents which can be most relevant to consumer question, a SBERT-based question-centered extractive Summarization level is accomplished to construct a compact and applicable context for the technology model. Unlike the question-impartial summarization scenario, this approach explicitly makes use of the consumer query to decide which sentences should be retained from the retrieved files.

Scenario 1: proactive summarization as an indexing tool (summarization-first approach)

Summarization is treated as pre-processing step for reducing the statistics length before the hunt procedure. This step goals to transform massive documents into concise understanding representations that distil the crucial thoughts. Based on these summaries, an (IR) search is executed in response to the user's question. The machine searches for a targeted semantic area, decreasing computational attempt and putting off statistics noise earlier than the query arrives. Overall Framework of the Proposed Summarization-First System in Figure 1.

Figure 2
Overall architecture of the proposed retrieval-first summarization system



Scenario 2: Query-Guided Contextual Summarization (Retrieval-First Summarization)

Summarization is performed as post-processing step, where it only applies documents that have been proven relevant to the question. The process begins by searching for relevant documents (IR) in the full text corpus based on the user's question. Once the candidate documents are identified, the summarization process begins only for this selected group. Overall Architecture of the Proposed Retrieval-First Summarization System in Figure 2.

4.2. Dataset

Neural Information processing system (NIPS) dataset represents one of the top ML conferences international. It includes the whole thing from laptop vision and deep getting to know to reinforcement studying and cognitive technological know-how. The dataset incorporates titles, abstracts, authors, and extracted text regarding each NIPS look at posted to this point, from the primary convention in 1987 till the contemporary convention in 2016.

4.3. Evaluation metrics

Regarding IR model's performance to be measured objectively, an efficient assessment metrics system is crucial. Cosine Similarity, Precision@k, BERT Score, F1@k, Recall@k, MRR, MAP, and nDCG@k are a few of those metrics. The majority of such measures are essential for evaluating IR systems and gauging the search results quality. The top k results are evaluated by Recall@k, Precision@k, and F1@k, with F1 balancing both, recall emphasizing coverage, and precision emphasizing relevance. MRR is focused on the first relevant result rank, whereas MAP averages average precision over queries to reward systems ranking relevant items high. NDCG@k can be considered a rank-aware metric that handles graded relevance, in which a few outcomes are more significant compared to others, by giving higher weights to the highly suitable items at the top of the list.

1) BERT Score: Score uses pre-trained contextual embedding for identifying semantic similarities *n*-gram overlap evaluations could overlook. BERT score can be defined as automatic evaluation tool for text generation that quantifies semantic similarity between the candidate and reference texts.
2) Precision@k: In relation to a list of recommended items is assessed by such performance metric. It determines the proportion regarding top-*k* sets recommended items that are of relevance to the user. The fraction of top *k* relevant retrieved documents.
3) Recall@k: This metric, which is utilized for evaluation of recommendation systems, is based on the number of the relevant items that have been selected out of all possibly relevant ones. It calculates percentage of relevant items that are found in the top-*k* recommendations for multiple queries. The fraction regarding all relevant documents in top *k*.
4) F1@K (F1 score at *K*): which represents harmonic mean value of precision and recall. Precision@k and Recall@k are balanced through the harmonic mean at specific cut-off point *K*, resulting in single score.
5) Mean Average Precision (MAP): Ranking or recommender systems can lead to re-turning relevant items in Top *K* results as well as prioritizing more relevant items is evaluated using quality metric known as MAP at *K*. Mean Reciprocal Rank (MRR): Statistical metric, which is focused on first-correct answer rank, MRR is utilized for evaluation of recommendation efficiency or IR systems. It can be estimated as average of reciprocal ranks of query results for some specific group of queries. The inverse of the rank of the first accurate item is reciprocal rank with respect to query response. More specifically, in a case where the correct item is at rank *k*, then the reciprocal rank would be 1/*k*. MRR is calculated by the averaging of the reciprocal ranks.
6) (NDCG@K): The ranking quality indicator referred to as NDCG captures effectiveness of ranking algorithms in recommendation systems. It performs the evaluation of how well the expected ranking that is related to the items matches the ideal ranking, taking into account each item-relevance. Rank-aware metrics are used for handling graded relevance. Highly relevant items appearing earlier in the list are given more weight, and the score is normalized to a range of [0].

5. Evaluation Results

The proposed system is evaluated by assessing the effect of the text summarization to the QA pipeline to improve IR performance. The analysis brings out the comparative performances of two different architectures: Scenario 1 (the Summarization-First architecture) that proactively condenses documents with the use of SBERT and then indexing and retrieval using all-MiniLM-L6-v2 and Scenario 2 (the Retrieval-First method) that performs initial context retrieval and then applies query-guided summarization. Finally, the experimental results show how each of the two strategies can effectively optimize the context that is inputted into the generation phase (T5 and BART-large-CNN) which directly impacts the overall accuracy and efficiency of answers that have been generated. Framework configuration comparison between baseline and proposed RAG is shown in Table 1.

Table 1
Framework configuration comparison between baseline and proposed RAG

Generation	Summarization	Retrieval	Framework Configuration
✓	*X*	✓	Traditional RAG (Baseline)
✓	✓	✓	Our Proposed RAG(Enhanced)

5.1. Baseline (traditional RAG)

Its technique involves retrieving texts without summarization using the usage of (all-MinLM-L6-v2 Model) after which sending them directly to generative models the usage of (BART large-CNN, T5 Models). However, the issue is that the prolonged texts determined in NIPS research papers exceed the fashions' maximum word restrict (context window), resulting in a loss of statistics.

5.1.1. Generation part

The results are compared according to how comparable the text chunks are and how well the questions' answers are generated. For evaluating the generation of answers for questions without

Table 2
Results of question similarity and BERT score evaluation for response generation without summarization

	T5 model		BART-large-CNN model	
Q	Score similarity	BERT scores	Score similarity	BERT scores
Q1	0.72	85.5	0.72	84.8
Q2	0.68	86.5	68.3	85.8
Q3	0.69	86.8	69.2	87.3

summarization and to analyze similarity chunk text, we employ RAG for retrieval part with the use of all-MinLM-L6-v2 Model as well as generation part with the use of T5 and Bart-large-CNN models. As shown in Table 2, the question similarity scores and BERT Scores are presented for evaluating the generated responses without summarization using the BART-large-CNN and T5 models.

5.1.2. Retrieval part

We compare the outcomes of IR for QA without summarizing and RAG for generation part utilizing Bart-large-CNN and T5 models and retrieval part utilizing all-MinLM-L6-v2 model. Table 3 presents the retrieval performance metrics for the input question using the T5 model without summarization. Table 4 presents the retrieval performance metrics for the input question using the BART-large-CNN model without summarization.

Table 3
The retrieval performance metrics for the input question using the T5 model without summarization

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	0.80	0.40	0.40	0.53	1.00	0.87
Q2	0.80	0.36	0.32	0.50	1.00	0.83
Q3	0.60	0.37	0.34	0.46	1.00	0.69

Table 4
The retrieval performance metrics for the input question using the BART-large-CNN model without summarization

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	0.80	0.40	0.40	0.53	1.00	0.87
Q2	0.80	0.36	0.32	0.50	1.00	0.83
Q3	0.60	0.37	0.34	0.46	1.00	0.69

5.2. Our proposed RAG

The information retrieval process using (all-MinLM-L6-v2 Model) follows one of two scenarios: the first scenario involves retrieving text from independently summarized documents using (SBERT), then passing it to the generation stage using (BART-large-CNN, T5 Models); the second scenario involves retrieving full texts, passing them to the summarization stage using (SBERT), and subsequently passing the output to the generation stage using (BART-large-CNN, T5 Models).

5.2.1. Scenario 1: Proactive summarization as an indexing tool (summarization-first approach)

1) Generation part

We evaluate the generation of answers for questions with summarization by utilizing RAG for retrieval part utilizing all-MinLM-L6-v2 Model and generation part utilizing T5 as well as Bart-large-CNN models, and we have compared results with chunk text similarity. Table 5 displays Comparison of Question Similarity Scores and BERT Scores for Response Generation Using BART-large-CNN and T5 with Summarization.

2) Retrieval part

We use all-MinLM-L6-v2 Model for retrieval part and T5 and for generation part for comparing the IR performance for QA with summarization utilizing RAG. Table 6 presents the retrieval performance metrics for the summarized input question using the T5 model. Table 7 presents the retrieval performance metrics for the summarized input question using the BART-large-CNN model.

Table 5
Comparison of question similarity scores and BERT scores for response generation using BART-large-CNN and T5 with summarization

	T5 model		BART-large-CNN model	
Q	Score similarity	BERT scores	Score similarity	BERT scores
Q1	0.63	0.86	0.63	0.89
Q2	0.61	0.86	0.60	0.84
Q3	0.48	0.82	0.48	0.89

Table 6
The retrieval performance metrics for the summarized input question using the T5 model

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	0.80	0.40	0.40	0.50	1.00	0.87
Q2	0.60	0.30	0.24	0.40	1.00	0.66
Q3	0.80	0.50	0.47	0.61	1.00	0.85

Table 7
The retrieval performance metrics for the summarized input question using the BART-large-CNN model

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	0.80	0.40	0.40	0.50	1.00	0.87
Q2	0.60	0.30	0.24	0.40	1.00	0.66
Q3	0.80	0.50	0.47	0.61	1.00	0.85

5.2.2. Scenario 2: Query-guided contextual summarization (retrieval-first summarization)

1) Generation part

We evaluate answer generation for questions with summarization utilizing RAG for part of retrieval by utilizing

all-MinLM-L6-v2 Model and the generation part utilizing Bart-large-CNN and T5 models, and we have performed a comparison of results with chunk text similarity. Table 8 displays the Comparison of Question Similarity Scores and BERT Scores for Response Generation Using BART-large-CNN and T5 with Summarization.

Table 8
Comparison of question similarity scores and BERT scores for response generation using BART-large-CNN and T5 with summarization

	T5 model		BART-large-CNN model	
Q	Score similarity	BERT scores	Score similarity	BERT scores
Q1	0.62	0.92	0.62	0.90
Q2	0.70	0.88	0.70	0.89
Q3	0.67	0.89	0.67	0.93

2) Retrieval part

We use all-MinLM-L6-v2 Model for retrieval elements and T5 and Bart-huge-CNN models for the producing component for comparing the information retrieval performance for QA with summarization utilizing RAG. Table 9 presents the retrieval overall performance metrics for the summarized input question the use of the T5 model. Table 10 presents the retrieval performance metrics for the summarized input query using the BART-huge-CNN version.

Table 9
The retrieval performance metrics for the summarized input question using the T5 model

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	1.00	0.50	0.66	0.50	1.00	1.00
Q2	0.80	0.36	0.50	0.36	1.00	0.87
Q3	0.80	0.50	0.62	0.44	1.00	0.83

Table 10
The retrieval performance metrics for the summarized input question using the BART-large-CNN model

Q	*P*@5	*R*@5	MAP	F1@5	MRR	nDCG@5
Q1	1.00	0.50	0.66	0.50	1.00	1.00
Q2	0.80	0.36	0.50	0.36	1.00	0.87
Q3	0.80	0.50	0.62	0.44	1.00	0.83

6. Conclusion

The results of the retrieval and generation phase have shown that the application of summarization had a definite beneficial effect on the enhancement of the performance of both T5 and BART-large-CNN models, both in terms of retrieving relevant information and the generated responses' quality. The retrieval performance indicated a significant increase in the P@5, F1@5, MAP, R@5, and nDCG@5 with the use of summarization especially with BART-large-CNN model that recorded the best results in most of the measures. It had very high scores, up to (1) in certain measures, including P@5, MRR, and nDCG@5, indicating its high efficiency in retrieving and correctly sorting the correct results. During the generation phase, the Score Similarity and BERT Score obtained results that both models could give answers that were like reference answers. The BART-large-CNN model, however, showed comparative advantage to retain meaning and semantic context particularly following the implementation of summarization, but the T5 model also performed similarly and efficiently on certain questions. The findings also showed that the summarization was effective in minimizing the redundant information and the ability of the models to concentrate on the significant information, which had a positive influence on the quality of responses and retrieval accuracy. Overall, it could be concluded that the incorporation of summarization with synthetic intelligence models has extensively helped to growth performance of the query-and-solution system, and that BART-large-CNN version with summarization proved to be the maximum efficient and stable one while compared to the rest of the instances and was the appropriate desire to beautify the performance of IR structures and to produce correct solutions.

Traditional RAG struggles to manner NIPS research papers because of the immoderate duration of scientific texts and wordiness; in evaluation, our summarization-incorporated method correctly filtered the textual content and provided the generative model with a rich, condensed context, resulting in a big boost in solution accuracy.

7. Future work

In future work, the aim is scaling the evaluation model and exploring methods of abstractive summarization for providing more cohesive context to QA pipeline, in addition to the optimization of the strategies of chunking for further refining balance between summary brevity and information density. In addition to that, a key direction will be generalizing the suggested framework across diverse specialized domains, like the medical domain. This expansion would include fine-tuning the model on domain-specific corpora (e.g., PubMed dataset) for the purpose of evaluating its ability to handle complex terminology and provide factually critical answers where it is paramount to have high precision. Lastly, the aim is evaluating other LLMs.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/benhamner/nips-papers?select=papers.csv.

Author Contribution Statement

Ahmed Sami Jaddoa: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review &

Editing, Visualization, Project administration. **Jaber Karimpour:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Supervision, Project administration. **Pedram Salehpour:** Conceptualization, Supervision.

Reference

[1] Hambarde, K. A., & Proença, H. (2023). Information retrieval: Recent advances and beyond. *IEEE Access*, *11*, 76581–76604. https://doi.org/10.1109/ACCESS.2023.3295776

[2] Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., & Xu, J. (2024). Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6437–6447. https://doi.org/10.1145/3637528.3671458

[3] Giarelis, N., Mastrokostas, C., & Karacapilidis, N. (2023). Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, *13*(13), 7620. https://doi.org/10.3390/app13137620

[4] Buthelezi, S., Adeyanju, K. I., Ajagbe, S. A., Akinlade, O. A., & Mudali, P. (2026). mBART50 fine-tuning for translation of South African languages. *Procedia Computer Science*, *283*, 3811–3820. https://doi.org/10.1016/j.procs.2026.06.442

[5] Badry, R., Sharaf Eldin, A, Mohammed., & Saad Elzanfally, D. (2013). Text summarization within the latent semantic analysis framework: Comparative study. *International Journal of Computer Applications*, *81*(11), 40–45. https://doi.org/10.5120/14060-2366

[6] AL-Banna, A. A., & AL-Mashhadany, A. K. (2022). Natural language processing for automatic text summarization [datasets]-survey. *Wasit Journal of Computer and Mathematics Science*, *1*(4), 102–110. https://doi.org/10.31185/wjcm.72

[7] Divya, S., Sripriya, N., Andrew, J., & Mazzara, M. (2024). Unified extractive-abstractive summarization: A hybrid approach utilizing BERT and transformer models for enhanced document summarization. *PeerJ Computer Science*, *10*, e2424. https://doi.org/10.7717/peerj-cs.2424

[8] Peddavenkatagari, C. R. (2024). Empowering information retrieval: A framework for effective data summarization using NLP and SBERT. *International Research Journal of Modernization in Engineering Technology and Science*, *6*(4), 7796–7801.

[9] Farea, A., & Emmert-Streib, F. (2025). Understanding question-answering systems: Evolution, applications, trends, and challenges. *Engineering Applications of Artificial Intelligence*, *156*, 110997. https://doi.org/10.1016/j.engappai.2025.110997

[10] Jian, H. Z., Johnson, O. V., & Wah, K. K. (2022). Text summarization for news articles by machine learning techniques. *Applied Mathematics and Computational Intelligence*, *11*(1), 174–196.

[11] Guo, M., Ainslie, J., Uthus, D. C., Ontanon, S., Ni, J., Sung, Y. H., & Yang, Y. (2022). LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 724–736. https://doi.org/10.18653/v1/2022.findings-naacl.55

[12] Jiang, Z., Ma, X., Chen, W., Wang, Z., Chen, Z., & Yan, Y. (2024). *LongRAG: Enhancing retrieval-augmented generation with long-context LLMs.* arXiv. https://doi.org/10.48550/arXiv.2406.15319

[13] Jearanaitanakij, K., Boonpong, S., Teainnagrm, K., & Thonglor, T. (2024). Fast hybrid approach for Thai news summarization. *Engineering Transactions*, *41*(4). https://doi.org/10.55003/ETH.410307

[14] Bailly, A., Saubin, A., Kocevar, G., & Bodin, J. (2025). Divide and summarize improve SLM text summarization. *Frontiers in Artificial Intelligence*, *8*, 1604034. https://doi.org/10.3389/frai.2025.1604034

[15] Aljohani, B., & Alsanoosy, T. (2026). Enhancing medical question answering with LLMs via a hybrid retrieval-augmented generation framework. *Information*, *17*(2), 133. https://doi.org/10.3390/info17020133

[16] Pujiono, I., Agtyaputra, I. M., & Ruldeviyani, Y. (2024). Implementing retrieval-augmented generation and vector databases for chatbots in public services agencies context. *Jurnal Ilmu Pengetahuan dan Teknologi Komputer*, *10*(1), 216–223. https://doi.org/10.33480/jitk.v10i1.5572

[17] Saha, B., Saha, U., & Malik, M. Z. (2024). QuIM-RAG: Advancing retrieval-augmented generation with inverted question matching for enhanced QA performance. *IEEE Access*, *12*, 185401–185410. https://doi.org/10.1109/ACCESS.2024.3513155

[18] Moreno-Cediel, A., del-Hoyo-Gabaldon, J. A., Garcia-Lopez, E., Garcia-Cabot, A., & de-Fitero-Dominguez, D. (2024). Evaluating the performance of multilingual models in answer extraction and question generation. *Scientific Reports*, *14*(1), 1–17. https://doi.org/10.1038/s41598-024-66472-5

[19] Meng, W., Li, Y., Chen, L., & Dong, Z. (2025). Using retrieval-augmented generation to improve the question-answering system in human health risk assessment: The development and application. *Electronics*, *14*(2), 386. https://doi.org/10.3390/electronics14020386

[20] Lin, Y., Zhang, W., & Lin, J. (2025). Question answering system based on retrieval-augmented generation. In *Proceedings of the 2025 International Conference on Computer Technology, Digital Media and Communication*, 494–499. https://doi.org/10.1145/3783669.3783745

[21] Kang, S. K., Jin, B., Kweon, W., Zhang, Y., Lee, D., Han, J., & Yu, H. (2025). Improving scientific document retrieval with concept coverage-based query set generation. In *WSDM 2025: Proceedings of the 18th ACM International Conference on Web Search and Data Mining*, 895–904. https://doi.org/10.1145/3701551.3703544

[22] Hoang, T. T. U., Rajendran, M., Zhang, K., Wu, Y., & Nguyen, V. A. (2025). PDF retrieval augmented question answering. *arXiv Preprint: 2506.18027.*

[23] Wu, C., Ding, W., Jin, Q., Jiang, J., Jiang, R., Xiao, Q., . . . , & Li, X. (2025). Retrieval augmented generation-driven information retrieval and question answering in construction management. *Advanced Engineering Informatics*, *65*, 103158. https://doi.org/10.1016/j.aei.2025.103158

[24] Bharati, H., Tarange, P., Kumar, A., Mate, S., & Bhagananagre, R. C. (2024). Text summarization using NLP. *International Journal for Research in Applied Science and Engineering Technology*, *12*(1), 803–807. https://doi.org/10.22214/ijraset.2024.56564

[25] Edress, Z., & Ortakci, Y. (2024). Optimizing text summarization with sentence clustering and natural language processing. *International Journal of Advanced Computer Science and Applications*, *15*(10), 1123–1132. https://doi.org/10.14569/IJACSA.2024.01510115

[26] Ozsoy, M., Cicekli, I., & Alpaslan, F. (2010). Text summarization of Turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 869–876.

[27] Ali, L. (2022). Improving information retrieval systems' efficiency. *International Journal of Engineering Research & Technology*, *11*(11), 287–292.

[28] Rayo, J., de La, Rosa., R, & Garrido, M. (2025). A hybrid approach to information retrieval and answer generation for regulatory texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, 31–35.

[29] Le, N. K., Nguyen, D. H., Nguyen, T. T. T., Nguyen, M. P., Le, T., & Nguyen, M. L. (2021). A novel pipeline to enhance question-answering model by identifying relevant information. In K. Yada, Y. Takama, K. Mineshima, & K. Satoh (Eds.), *New frontiers in artificial intelligence* (pp. 296–311). Springer. https://doi.org/10.1007/978-3-031-36190-6_21

[30] Hemmat, A., Sharbaf, M., Kolahdouz-Rahimi, S., Lano, K., & Tehrani, S. Y. (2025). Research directions for using LLM in software requirement engineering: A systematic review. *Frontiers in Computer Science*, *7*, 1519437. https://doi.org/10.3389/fcomp.2025.1519437

[31] Pathak, P., & Rana, P. S. (2024). Comparative analysis of pretrained models for text classification, generation and summarization: A detailed analysis. In A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C. Liu, S. Bhattacharya, & U. Pal (Eds.), *Pattern recognition* (pp. 151–166). Springer. https://doi.org/10.1007/978-3-031-78107-0_10

[32] Muludi, K., Fitria, K. M., Triloka, J., & Sutedi. (2024). Retrieval-augmented generation approach: Document question answering using large language model. *International Journal of Advanced Computer Science and Applications*, *15*(3), 776–785. https://doi.org/10.14569/IJACSA.2024.0150379

[33] Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, *8*, 100417. https://doi.org/10.1016/j.caeai.2025.100417

[34] Yu, W. (2022). Retrieval-augmented generation across heterogeneous knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 52–58. https://doi.org/10.18653/v1/2022.naacl-srw.7

How to Cite: Jaddoa, A. S., Karimpour, J., & Salehpour, P. (2026). A Hybrid RAG-Based Framework for Improved Information Retrieval from Long Documents via Summarization and Question Answering. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA620210968

Appendix A

The following section gives a case study illustrative of proposed framework applied in the presented study for explaining response-generation mechanism in RAG model. It is based upon asking questions as well as retrieving semantically similar texts with cosine similarity scale. Two key scenarios are compared in the following tables: Scenario 1: Proactive Summarization as an Indexing Tool (Summarization-First Approach). Scenario 2: Query-Guided Contextual Summarization (Retrieval-First Summarization). The queries that have been used, texts or responses that have been generated in every case, and quantitative values that have been utilized to evaluate the quality of generation by BERT score are provided in tables A1-A7. The purpose of such comparison is to show summarization effect on enhancing semantic consistency and accuracy of responses that have been generated and clearly analyzing differences between those two scenarios that are involved in the presented study.

Table A1
Example questions and their corresponding reference answers from the dataset

No	Questions	Reference Answers
Q1	Using independent component analysis (ICA) for the removal of artifacts in the EEG recordings.	ICA that is used for artifact removal in the recordings of EEG. It works by decomposing multichannel EEG data into statistically independent components; also, multi-channel EEG epochs have been analyzed with the use of ICA, allowing researchers to separate meaningful brain signals from unwanted artifacts like eye blinks, eye movements, and muscle noise. The method has shown significant success in improving the quality of single-trial EEG, and it removes pervasive artifacts of all types and separates non-event related background EEG activities to separate components.
Q2	What are support vector machines (SVM)?	SVMs have played an important role in the classification problems. SVMs are finding applications in regression estimation, pattern recognition, and operator inversion for ill-posed problems. They suffer from their training algorithm's complexity, which is at least quadratic with respect to the number of the examples. SVMs can in fact suffer in high-dimensional spaces where numerous features are irrelevant.
Q3	What is Gaussian process regression (GPR)?	Gaussian process regression allows a simple analytical treatment of the exact Bayesian inference. It was found to provide good performance yet scales badly with the number of training data. Gaussian process regression (GPR) retains many of the benefits of probabilistic Gaussian process models. GPR has shown excellent performance in several applications.

Table A2
Example of responses generated by the T5 model with BERT Score evaluation for selected questions without summarization (traditional RAG)

No	Questions	Generated Answers	BERT scores
Q1	Using independent component analysis (ICA) for artifact removal in EEG recordings.	ICA attempts to decompose observed data to components, which are as statistically independent from one another as possible. This paper presents a novel approach for the suppression of interference signals and separation of signals from individual evoked sources.	0.86
Q2	What are support vector machines (SVM)?	SVMs implement the following concept: they map input vectors to high-dimensional feature space, in which a maximal margin hyperplane is constructed. in the last few years, there has been a good deal of interest in SVMs. However, SVM use remains limited to a small number of researchers. One of the possible reasons is that training algorithms for SVMs as are slow, particularly for the large problems.	0.87
Q3	What is Gaussian process regression (GPR)?	GPR has demonstrated very good performance in several applications. One unwanted GPR aspect is its scaling behavior with training data set size N. In the direct implementations, the time of training increases as ON3, with a memory footprint of ON2. ARMA models used in time series analysis and spline smoothing correspond to GP model for regression with the training energy estimated based on Eq. (2).	0.87

Table A3
Examples of BART-large-CNN-generated responses and their BERT Scores for selected questions without summarization (traditional RAG)

No	Questions	Generated Answers	BERT scores
Q1	Using independent component analysis (ICA) for artifact removal in the EEG recordings.	ICA has generated a great deal of recent practical and theoretical interest due to its success on a wide range of problems of signal processing. This paper presents an innovative technique for the suppression of interference signals and separation of signals from individual evoked sources. This technique has been based on a new probabilistic graphical model termed stimulus-evoked independent factor analysis (SEIFA).	0.85
Q2	What are support vector machines (SVM)?	In the last few years, there was a surge of interest in SVMs, which have been shown to give good performance in terms of generalization on a broad range of problems. However, SVM use remains limited to a small group of researchers. One of the possible reason is that training algorithms for SVMs are slow, in particular for the large problems. The present paper provides a description of a novel SVM learning algorithm, which is often faster, easier to implement, and has more efficient scaling characteristics when compared to standard SVM algorithm.	0.86
Q3	What is Gaussian process regression (GPR)?	GPR has demonstrated exceptional performance in several applications. One of the unpleasant aspects of GPR is its scaling behavior with training data set size = N. In the direct implementations, the training time increases as ON3, with an ON2 memory footprint. A GP model has been considered for the regression with the training energy given by Eq. (2). In such case, the prior measure could be represented simply by a dimensional Gaussian process prediction. We call the hyperparameters the same as those used in neural networks.	0.87

Table A4
Examples of T5-generated responses and their BERT Scores for selected questions with summarization (enhanced RAG, scenario 1)

No	Questions	Generated Answers	BERT scores
Q1	Using independent component analysis (ICA) for artifact removal in EEG recordings.	The use of ICA for the artifact removal in the EEG recordings. A rigorous notion of the decomposition of a multivariate function into a set of lower dimensional factors is presented.	0.86
Q2	What are support vector machines (SVM)?	RMMS requiring only additional linear constraints in the support vector machine quadratic program leads to a competitively fast implementation. Forward sub-set selection appears to be the approach of choice for the 2 linear problems we tried at varying. On three of those problems, the feature space representation was best, bagging was worse, and support vector regression came in second. SVMs can in fact suffer in high-dimensional spaces where numerous features are not relevant.	0.86
Q3	What is Gaussian process regression (GPR)?	Our approach can also be significantly faster than support vector machine with the sequential minimal optimization algorithm. Among the many proposed sparse Gaussian process approximations, our method is most closely related to. Our method also requires specifying a rejection threshold and is dependent upon how we draw gm jointly from Gaussian process prior, conditioned on all the current latent state. The latent function in these problems is always one dimensional and densely discretized.	0.82

Table A5
Examples of BART-large-CNN-generated responses and their BERT Scores for selected questions with summarization (enhanced RAG, scenario 1)

No	Questions	Generated Answers	BERT scores
Q1	Using Independent component analysis (ICA) for artifact removal in EEG recordings.	Independent component analysis (ICA) can be utilized in blind source separation as in, particularly in cases were fewer sensors exist than sources, whereas the source number is not known. Due to the fact that the components that were separated by sparse representation aren'*t* constrained by the complete independence condition, they could be utilized in analyzing brain synchrony probably with a higher effectiveness compared to the general ICA algorithms.	0.89
Q2	What are support vector machines (SVM)?	Support vector machine and RMM have similar asymptotic behavior. We identified the sensitivity of support vector machines and maximum absolute margin criteria to affine scaling. Forward subset selection seems as the method of choice for the two linear problems that have been tried at varying signal-to-noise ratio values.	0.84
Q3	What is Gaussian process regression (GPR)?	Gaussian process regression (GPR) retains many of the benefits of probabilistic Gaussian process models. Our method can also be significantly faster than support vector machine with the sequential minimal optimization algorithm. It has been found that sub-sets of representer methods can give good and especially fast predictions for datasets with medium and high levels of noise.	0.89

Table A6
Exaples of T5-generated responses and their BERT Scores for selected questions with summarization (enhanced RAG, scenario 2)

No	Questions	Generated Answers	BERT scores
Q1	Using independent component analysis (ICA) for artifact removal in EEG recordings.	Applies a linear decomposition tool, ICA, to multichannel single-trial EEG records. It removes pervasive artifacts of all types and separates non-event related background EEG activities into separate components.	0.92
Q2	What are support vector machines (SVM)?	SVMs have played an important role in the classification problems. Little work went into the incorporation of prior knowledge into SVMs, one of the open questions that remains is how to set "tunable" parameters.	0.88
Q3	What is Gaussian process regression (GPR)?	GPR demonstrated excellent performance in several applications, yet it scales badly with the number of training data. This paper presents a fast approximation approach, based upon kd-trees. It leads to a significant reduction of prediction as well as training times.	0.89

Table A7
Examples of BART-large-CNN-Generated responses and their BERT Scores for selected questions with summarization (enhanced RAG, scenario 2)

No	Questions	Generated Answers	BERT scores
Q1	Using independent component analysis (ICA) for artifact removal in EEG recordings.	Single-trial analysis can possibly reveal considerably richer information on the event-related brain dynamics in the ERPs. It suffers from pervasive artifacts that are associated with blinks, eye movements, and muscle noise. Multi-channel EEG epochs have been analyzed with the use of ICA.	0.90
Q2	What are support vector machines (SVM)?	SVMs are finding applications in regression estimation, pattern recognition, and operator inversion for ill-posed problems. Probabilistic interpretation of SVMs could be used to tackle these problems. The basic SVM idea is to map inputs X onto vectors $c>(x)$ in some high-dimensional feature space.	0.89
Q3	What is Gaussian process regression (GPR)?	GPR allows a simple analytical treatment of exact Bayesian inference. It was found to provide good performance yet scales badly with the number of training data. Fast Gaussian process regression using KD-Trees. Online Gaussian processes achieve a slightly lower level of accuracy.	0.93