BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Retrieval-Augmented Generation (RAG): Enhancing LLMs with External Knowledge

Samiksha Patil

Dept. of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, India
Email: samiksha.patil@dypic.in

Prof. Ashima Mahendra

Dept. of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, India
Email: ashimamahendra@dypic.in

Abstract

Large Language Models (LLMs) generate fluent answers, but they depend mainly on their training data. As a result they can give wrong or outdated responses, and they may lack the specific information a question needs. Retraining a model again and again to fix this takes a great deal of time. Retrieval-Augmented Generation (RAG) addresses the problem by retrieving relevant information from external knowledge sources, such as documents, databases and the web, before a response is generated. This paper presents a literature survey and problem understanding for RAG. It explains the retrieve–augment–generate workflow, states the problem and six objectives, and relates RAG to current technology trends: real-time information, enterprise AI, agentic AI, cost-effective scalability and reduced hallucination. Eight studies from 2020 to 2026 are compared on methodology, key findings and limitations. The analysis reveals that the quality of information retrieval, as well as the effectiveness, accuracy and assessment, still remain unstudied topics. Four research gaps are identified: high computational resources, irrelevant or incomplete retrieval, limited effectiveness in specialised domains, and the need for better evaluation of factuality, relevance and reliability.

Keywords : retrieval-augmented generation, large language models, hallucination, dense retrieval, vector databases, knowledge-intensive NLP, literature survey

I. INTRODUCTION

Nowadays, Large Language Models (LLMs) are associated with tasks related to question answering, summarization and conversational agents. However, such models use static knowledge acquired at the point of training. Therefore, for instance, in case of a question about a recent event or a specialized topic that was poorly represented in the training dataset, LLMs will generate the answer that will be outdated, inaccurate or even imaginary. One of the reasons for that is that updating the whole model whenever there are changes in facts is a time-consuming and expensive process. Retrieval-Augmented Generation (RAG) is a Generative AI framework that combines the capabilities of LLMs with an external knowledge

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

source [1]. Instead of relying only on information learned during training, RAG retrieves relevant and up-to-date information from documents, databases or the web before generating a response. This improves the accuracy, relevance and reliability of AI-generated answers while reducing hallucinations. RAG is now widely used in chatbots, enterprise search, healthcare, legal assistance, customer support and education.

This paper reports the literature survey and problem-understanding stage of a technical seminar on RAG. Its contributions are: (i) a clear statement of the problem and six objectives; (ii) a discussion of why RAG matters for current technology trends; (iii) a comparison of eight studies [1, 2, 3, 4, 5, 6, 7, 8] by methodology, findings and limitations; and (iv) four research gaps that follow from the comparison. Section II explains the RAG workflow, Section III states the problem and objectives, Section IV covers relevance to current trends, Section V presents the literature survey, Section VI discusses problem understanding and the research gap, and Section VII concludes.

II. BACKGROUND: THE RAG WORKFLOW

A RAG system adds a retrieval step in front of the language model. Fig. 1 shows the five stages of the workflow.

1. **Question:** the user asks a question.
2. **Retrieve:** the system forms a retrieval query and searches the knowledge base for relevant information.
3. **Augment:** the most relevant retrieved texts are added to the prompt.
4. **Generate:** the LLM generates a response from the question and the retrieved context.
5. **Response:** the user receives an accurate, context-aware answer.

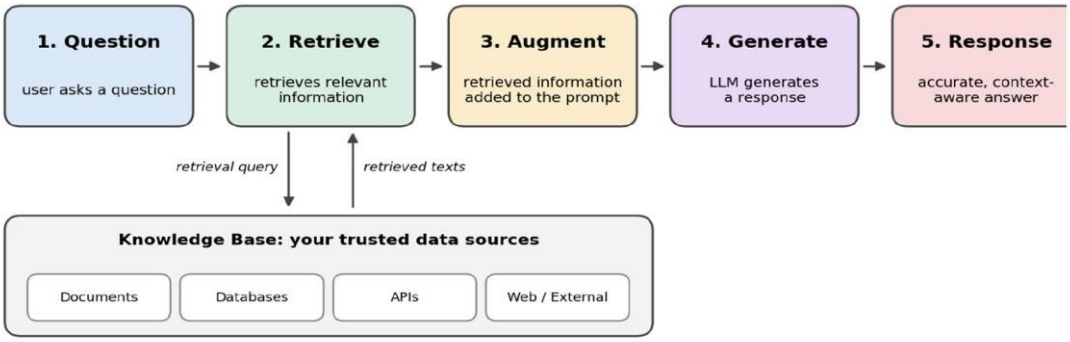


Figure 1. Workflow of Retrieval-Augmented Generation: a question triggers retrieval from the knowledge base, the retrieved texts augment the prompt, and the LLM generates the response.

The knowledge base holds the trusted data sources of the system: documents, databases, APIs and web or external content. Because the knowledge lives outside the model, it can be changed without modifying the model itself. The original RAG framework pairs a pretrained sequence-to-sequence generator with a dense-vector Wikipedia index and a neural retriever [1].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

III. PROBLEM STATEMENT AND OBJECTIVES

A. Problem Statement

LLMs can give wrong or outdated answers because they depend mainly on their training data. They may not have access to the latest or specific information when answering a question, and updating the entire model repeatedly requires a lot of time. RAG addresses these challenges of hallucinated, inaccurate or outdated information by retrieving relevant information from external knowledge sources before generating responses. This improves the accuracy, reliability, transparency and trustworthiness of AI-generated answers.

B. Objectives

The study is guided by six objectives.

1. **Improve accuracy:** retrieve relevant information from external sources to generate more accurate and contextually relevant responses.
2. **Minimize the occurrence of errors:** Create anlin the answers provided by grounding them on actual and trusted information and reliable sources.
3. **Keep information updated:** gain access to new data from external databases or documents without a need for LLM constant retraining.
4. **Support domain-specific knowledge:** let LLMs answer using specialised information from fields such as healthcare, finance, education and technology.
5. **Enhance reliability and trust:** make responses more dependable by connecting generated answers with supporting information.
6. **Lower the cost:** Use an external knowledge base rather than going through the whole process of retraining.

IV. RELEVANCE TO CURRENT TECHNOLOGY TRENDS

- **Real-time and updated information.** RAG retrieves the latest information from databases, documents and knowledge bases, which suits applications where information changes frequently.
- **Enterprise AI.** This system could be a great resource of different papers analyzing LLMs, RAG architecture, dense retrieving, vector databases, and knowledge inside AI.
- **Agentic AI.** RAG is becoming important in AI agents, because agents can retrieve information from several sources and use it to perform complex tasks and make decisions.
- **Cost-effective and scalable.** Organisations can update the knowledge base instead of retraining a large model, which makes the system easier and cheaper to maintain.
- **Reduced hallucination.** Retrieving supporting information before answering helps reduce false or made-up output.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

V. LITERATURE SURVEY

A. Review Approach

Analyses of scientific papers on LLMs, RAG architectures, dense retrieval systems, vector databases, and knowledge-enhanced AI were carried out. Existing RAG frameworks and retrieval techniques were then studied and compared on architecture, retrieval method and generation quality. The eight studies selected for the survey matrix are summarised in Table 1 and placed in time in Fig. 2.

B. Foundational Work

The authors Lewis etal[1] proposed the RAG framework whereby they combine the pre-trained sequence-to-sequence generator with the dense vector index from Wikipedia and use a neural retriever for better performance. Once retrieval errors happen, they will continue to produce wrong generation since the original system depends highly on one fixed knowledge base.

Izacard and Grave [2], on the other side, retrieve several passages at once independently of each other and fuse them in an abstract generator-decoder that will lead to improvements beyond expectation in question answering. The use of lots of retrieved data can lead to increased cost and processing requirements.

Atlas[3] is another example of a pre-trained model

C. Adaptive and Structured Retrieval

Self-RAG utilizes an adaptive retrieval approach combined with generation and self-reflection analysis via reflection tokens. The system determines when a retrieval operation is needed, which makes it possible to enhance factuality and accuracy of citations; however, the downside is the advanced system complexity and specialized training.

Edge et al. introduce graph representation-based knowledge and community summarization for retrieval purposes.

D. Surveys and Evaluation

Chen et al. [6] carry out a thorough evaluation of knowledge-based retrieval, generating mechanisms, multimodal data and reasoning aspects. The authors mention such an important sphere as knowledge incorporation, interpretability, efficiency of retrieval and adaptability to various contexts; unfortunately, matching information and widely applying external sources still proves to be hard.

Wu et al. [7] analyse RAG architectures, retrieval strategies, generation, evaluation and applications for natural language processing. They show the rapid evolution of RAG and mark retrieval quality, efficiency, factuality and robustness as major research areas. As a survey, the work introduces no single new architecture.

Upadhyay et al.[8] describe the TREC 2025 RAG Track which assesses systems with extensive descriptive queries and metrics like relevancy, precision, and attribution. Although this formulation brings new points of view into evaluation, it can still prove to be ineffective in terms of evaluating complicated reasoning techniques and attribution problems.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

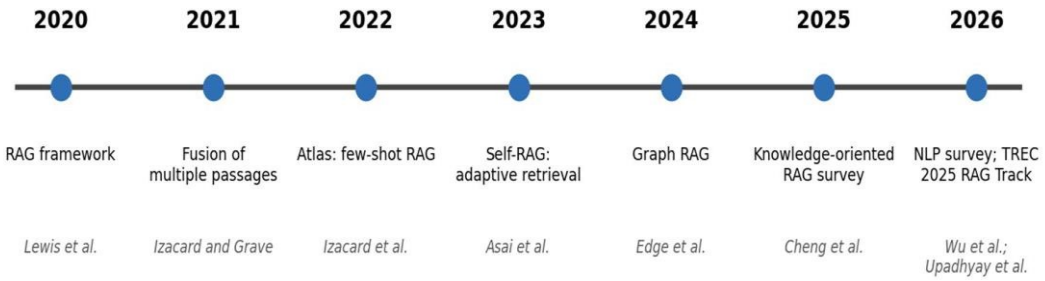


Figure 2. Timeline of the eight surveyed studies, from the original RAG framework (2020) to recent surveys and evaluation work (2026).

TABLE 1

LITERATURE SURVEY MATRIX

Author/Year	Methodology used	Key findings	Limitations
Lewis et al. (2020) [1]	Combines a pretrained seq2seq generator with a dense-vector Wikipedia index and neural retriever	Established the RAG framework; improved factuality and performance on knowledge-intensive NLP tasks	Retrieval errors can propagate into generation; relies heavily on a fixed knowledge index
Izacard & Grave (2021) [2]	Retrieves multiple passages independently and fuses them through a generative encoder-decoder	Combining information from multiple passages substantially improves open-domain QA	Processing many passages increases computational and memory requirements
Izacard et al. (2022) [3]	Pre-trained retrieval-augmented LM for few-shot knowledge-intensive tasks	Strong few-shot QA performance with far fewer parameters	Requires a high-quality retrieval index and substantial pre-training
Asai et al. (2023) [4]	Adaptive retrieval combined with generation and self-critique using reflection tokens	Model decides when retrieval is needed; improves factuality and citation accuracy	Requires specialised training; increases system complexity
Edge et al. (2024) [5]	Graph-based representations and community summaries used to retrieve information	Improves global sensemaking over large document collections compared with chunk-based retrieval	Graph construction can be expensive and requires additional processing

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Author/Year	Methodology used	Key findings	Limitations
Cheng et al. (2025) [6]	Reviews knowledge-oriented RAG, retrieval mechanisms, generation, multimodal data and reasoning	Identifies knowledge integration, interpretability, retrieval efficiency and domain adaptation as important directions	Knowledge alignment and efficient integration of diverse external sources remain difficult
Wu et al. (2026) [7]	Comprehensive analysis of RAG architectures, retrieval strategies, generation, evaluation and applications	Shows rapid evolution of RAG; identifies retrieval quality, efficiency, factuality and robustness as major research areas	Survey-based; does not introduce a single new RAG architecture
Upadhyay et al. (2026) [8]	Evaluates RAG systems using long narrative queries and metrics such as relevance, completeness and attribution	Moves RAG evaluation toward complex real-world information needs and citation-supported answers	Evaluation remains challenging for complex reasoning and attribution

VI. PROBLEM UNDERSTANDING AND RESEARCH GAP

A. Problem Understanding

The survey reveals a development. It can be shown that the RAG method and the use of multi-passages illustrate that information from the outside world has a positive impact on answering factual questions. According to Atlas, externally-accessed information can provide adequate results for small systems. After that, further research improved the process of retrieving information by allowing the systems to choose the most appropriate time to access the information or organizing data as a graph. In recent studies, the authors turn to the very basics and analyze the development of the field under realistic information retrieval demands. Two ideas are found in the limitations column of Table 1. First, the process of retrieving or generating something usually requires a higher amount of computation, training, or complexity of the system.

B. Research Gap

The comparison leads to four research gaps, which are mapped to the surveyed evidence in Table 2.

1. RAG systems require high computational resources for retrieval, indexing and generation.
2. Irrelevant or incomplete retrieved information can lead to inaccurate or hallucinated responses.
3. Existing RAG models have limited effectiveness in specialised domains without domain-specific knowledge.
4. Better methods are needed to evaluate the factuality, relevance and reliability of RAG-generated answers.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE 2

RESEARCH GAPS AND SUPPORTING EVIDENCE

Research gap	Evidence in surveyed studies
High computational resources	Many passages raise compute and memory [2]; graph construction is expensive [5]; large index and pre-training needed [3]
Irrelevant or incomplete retrieval	Retrieval errors propagate into generation [1]; adaptive retrieval and self-critique add complexity [4]
Limited domain effectiveness	Knowledge alignment and domain adaptation remain difficult [6]
Evaluation of factuality, relevance and reliability	Evaluation of reasoning and attribution is challenging [8]; factuality and robustness remain open [7]

VII. CONCLUSION

.This paper has completed the literature review and the phase of understanding the problem for RAG. The topic of the seminar and the statement of the problem were determined. The papers on LLMs, RAG architectures, dense retrieval, vector databases and knowledge-backed AI technologies were studied, and the existing RAG technologies and retrieval approaches have been compared in terms of the type of architecture, retrieval technique and the generation quality. The matrix of literature reviews was created analyzing the ways and findings and limitations of eight studies. From the matrix, four gaps were detected: high expenses, retrieval irrelevant or incomplete results, limited efficiency in definite domains and evaluation of factuality, relevance and reliability.

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” 2020.

[2] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in Proc. EACL, 2021, pp. 874–880.

[3] Izacard et al., “Atlas: Few-shot learning with retrieval augmented language models,” 2022.

[4] Asai et al., “Self-RAG: Learning to retrieve, generate, and critique through self-reflection,” 2023.

[5] Edge et al., “From local to global: A Graph RAG approach to query-focused summarization,” 2024.

[6] Cheng et al., “A survey on knowledge-oriented retrieval-augmented generation,” 2025.

[7] Wu et al., “Retrieval-augmented generation for natural language processing: A survey,” Artificial Intelligence Review, 2026.

[8] Upadhyay et al., “Overview of the TREC 2025 Retrieval Augmented Generation (RAG) Track,” 2026.