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 for Enhancing Large Language Models with External Knowledge: A Critical Theoretical Review

[1]Dhruv Mandal, [2]Arvind R. Gautam, [3]Dr. Vishal Puranik

[1]Student, [2]Assistant Professor - Seminar Guide, [3]Assistant Professor - Seminar Guide
[1]Department of Artificial Intelligence and Data Science,
[1]Ajeenkya DY Patil School of Engineering, Pune, India

Abstract

During their training, Large Language Models (LLMs) learn their knowledge in the form of parameters. As a result, LLMs can be prone to hallucination (generation of misinformation) and have outdated information. They are also limited in terms of what they can understand and how well they can operate. Retrieval-Augmented Generation (RAG) is a proposed solution that aims to address these limitations. RAG includes an additional text generator that can query an external knowledge base to access and include up-to-date information in the model's generated response. The paper includes 15 citations and covers the theory behind RAG. The authors have not designed or developed any new system, and they have not achieved any gains in system performance. Instead, they provide the definition and design of RAG and discuss the technologies associated with RAG including dense passage retrieval, retrieval-augmented pre-training, and decoding using a fused generator. The authors address the various methods for retrieval and context integration and assess the potential for RAG in the legal and medical domains, along with various design decisions. Finally, they list open issues and research directions. They conclude that retrieval methods have enabled better system performance in knowledge-based tasks; however, retrieval methods alone are not able to completely eliminate hallucination in system outputs. The identified issues have existing solutions such as multi-hop reasoning, selecting answers based on evidential documents, latency and trade-offs of system coverage and breadth, and adapting to a target domain. Nevertheless, there are still research problems that need to be addressed.

Keywords: Retrieval-Augmented Generation, Large Language Models, Natural Language Processing, Dense Passage Retrieval, Fusion-in-Decoder, Hallucination Mitigation, Knowledge-Intensive NLP.

I. Introduction

Large language models store what they know in their parameters, which are fixed when training ends [1], [2]. Consequently a standalone model cannot reliably report events after its training cutoff, cannot cite the source of a statement, and may lack coverage of specialised or private material. Lewis et al. point out that these models have little predictive power and are difficult to update or extend. [1]. They may also produce fluent statements that are not supported by facts, a behaviour commonly called hallucination [3]. Updating parametric knowledge by retraining is costly, which motivates ways to supply knowledge from outside the model.

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), introduced by Lewis et al. [1], combines a pre-trained generator (parametric memory) with a searchable document index (non-parametric memory). Related lines of work include retrieval-augmented pre-training [2], dense retrievers [4], and generators that fuse many passages [5]. Two surveys organise the area [6], [7]. Because the field has grown quickly, claims about what RAG achieves are often stated in general terms, and the supporting evidence is spread across tasks, models, and metrics that are not directly comparable.

The technical seminar synopsis on which this paper is based identifies four gaps: multi-hop reasoning, standardised measurement of hallucination reduction, trade-offs among retrieval quality, response quality, and latency, and limited coverage of domain-specific applications. This paper examines each gap against the published record, and distinguishes what is established, what is partially addressed, and what remains open.

The study is purely theoretical. It does not build a RAG pipeline, collect data, run experiments, or produce benchmark scores. Its contributions are: (i) a unified formal description of RAG components; (ii) a comparative analysis of retrieval and context-integration strategies; (iii) an appraisal of application evidence where the retained literature supports it; (iv) an evidence-graded analysis of open challenges; and (v) a compact reference base of 15 sources, each marked as peer-reviewed or preprint.

The review focuses on four key questions. RQ1: how do RAG architectures couple retrieval with generation, and how has the design evolved? RQ2: what does the literature show about hallucination mitigation and factual grounding? RQ3: what trade-offs link retrieval quality, answer quality, and latency? RQ4: which research gaps are open, and which have been partly resolved?

A. Review Approach and Scope

This is a narrative critical review, not a formal systematic review, and no review protocol was registered. The reference base is deliberately limited to 15 sources. It contains the eight sources listed in the synopsis (Lewis et al., Karpukhin et al., Guu et al., Izacard and Grave, Gao et al., Sharma, and the two 2026 preprints on federated RAG and RAG security) and seven further sources. These were added only where a specific analytical need was not covered by the synopsis: a hallucination taxonomy, in-context retrieval, adaptive retrieval, evaluation of RAG abilities, the position of evidence in context, and application evidence in medicine and law. Peer-reviewed venues were preferred. Preprints are used where they are the only or the most informative source, and are labelled as preprints in Table I, in the reference list, and in the text. Findings from 2026 preprints are treated with extra caution because they have not yet been through peer review.

Topics named in the synopsis that the retained sources do not evidence, such as vector database products, education, finance, and enterprise support, are not appraised, and no claim is made about them. Section III presents the literature from the most recent work to the oldest, and Table I follows the same order. The reference list follows the standard IEEE convention of numbering by first citation, so reference numbers do not run in year order.

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)

II. BACKGROUND AND FUNDAMENTALS

A. Parametric Knowledge and Its Limits

A pre-trained language model estimates a distribution over text and stores what it has learned in its weights. This is efficient for fluency and broad reasoning but has four consequences for knowledge-intensive use [1]. First, knowledge is static, so facts that change after training are missing. Second, coverage is uneven, since rare or private facts are poorly represented. Third, answers carry no explicit provenance. Fourth, once an error occurs, it can be difficult to fix without training the model again.

Huang et al. note that there are two main categories of hallucinations: fidelity hallucinations, where the LLM produces a response in conflict with the given instructions and/or context; and factuality hallucinations, where the model produces a response that is in direct contradiction with the known world facts [3]. This distinction is useful in RAG systems, since retrieval primarily deals with providing missing or outdated factual information, and can even supply the information to evaluate the fidelity of the generated response. However, it is important to understand that retrieval cannot guarantee that the generator will follow the information correctly.

B. Formal View of RAG

Let x be an input query, y be the output and z be a retrieved passage drawn from a corpus. A dense retriever scores a query passage pair based on the inner product of two encodings [4]:

$$\mathrm{sim}(x, z) = E_Q(x)^T E_P(z) \qquad (1)$$

In the original formulation the retrieval distribution is proportional to the exponentiated score [1]:

$$p_\eta(z \mid x) \propto \exp(\, \mathrm{sim}(x, z)\,) \qquad (2)$$

The query and the passage are then used to condition the generator using parameters θ. Lewis *et al.* define two ways of marginalising over the top-*k* passages. The *RAG-Sequence* model uses one passage for the whole output sequence [1]:

$$p(y \mid x) \approx \Sigma_{z \in \text{top-}k}\, p_\eta(z \mid x)\, p_\theta(y \mid x, z) \qquad (3)$$

whereas the *RAG-Token* model allows a different passage for each generated token:

$$p(y \mid x) \approx \Pi_i\, \Sigma_{z \in \text{top-}k}\, p_\eta(z \mid x)\, p_\theta(y_i \mid x, z, y_{<i}) \qquad (4)$$

These equations express the central idea: the output probability is a mixture over retrieved evidence, with the retrieval distribution acting as a latent variable. Later systems relax the strict probabilistic form, for example by placing retrieved text in the prompt of a frozen model [8], but the retrieve-then-generate structure is preserved.

C. Components and Workflow

A RAG pipeline has four conceptual stages (Fig. 1). The process of indexing involves cleaning, segmenting, and encoding source materials into vectors before storing them in a searchable index. During retrieval, the query is encoded and the k most similar chunks are found by approximate nearest-neighbour search; both Lewis et al. and Karpukhin et al. use maximum inner product search over a precomputed index [1], [4]. During augmentation, the retrieved chunks are combined with the query, often with an instruction to rely

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)

on the evidence. During generation, the model produces an answer conditioned on this augmented input. Gao et al. organise the field around these retrieval, augmentation, and generation components [6].

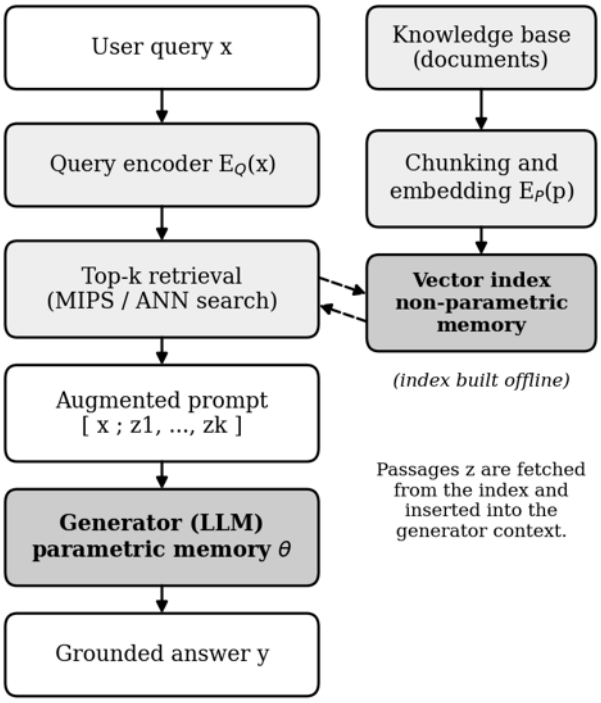


Fig. 1. Conceptual RAG architecture. Dashed arrows indicate access to the non-parametric memory. This is an explanatory illustration prepared by the author. It is not a system that has been implemented or tested.

D. Parametric and Non-Parametric Memory

RAG separates what the model knows how to do from what it can look up [1]. Parametric memory is responsible for language and reasoning. Non-parametric memory is used to store and edit knowledge that is easily accessible and editable by users, such as facts, without editing the model. Some advantages of the parametric approach are being able to easily provide source documents and cheap and fast updates of knowledge. Some disadvantages are the new failure mode of retrieval and an additional dependency. This memory approach also comes with the problem of having conflicting information when retrieved documents contradict information in the parametric memory, which is a facet of the faithfulness problem.

III. LITERATURE REVIEW AND RELATED WORK

The last 15 retained references were reviewed in reverse chronological order, that is, starting from the latest. The rationale for choosing the order of presentation in this section is that it is supposed to follow a forward evolution of the technology starting from the latest state-of-the-art and then backtracking to the earlier foundations of the technology. Table I, shown below, lists the references used in the review study in the chronological order.

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)

A. 2026: Deployment and Security (Preprints)

Two 2026 preprints cited in the synopsis extend the discussion to deployment settings. FD-RAG proposes a federated dual-system design in which lightweight memory matching handles well-covered queries and LLM reasoning is invoked only when necessary, aimed at edge settings where documents cannot be shared [9]. The authors report accuracy gains of up to 7.8% and 8.4 times lower latency than their baselines, and give a convergence analysis for the hypergraph learning [9]. These are the authors' own results from a preprint and have not been independently replicated.

A security survey by Xu et al. organises attacks and defences with a taxonomy along four axes (attack surface, defence layer, objective, and target) over a six-stage knowledge-access pipeline, and argues that earlier work often conflates RAG-specific risks with inherent LLM flaws [10]. It is also a preprint, so its conclusions should be revisited after peer review.

B. 2025: Taxonomies, Hallucination, and Legal Evidence

Sharma provides a taxonomy of retriever-centric, generator-centric, hybrid, and robustness-oriented architectures and underlines recurrent trade offs between retrieval accuracy and generating flexibility, between efficiency and fidelity, and between modularity and coordination [7]. The cited version is an arXiv preprint that states it is under review. Huang et al., in a survey published in a peer-reviewed journal, examine hallucination in LLMs and discuss the limits of retrieval-augmented models in combating it [3].

Magesh et al. conducted a preregistered evaluation of commercial legal research tools built on RAG, whose providers had described RAG as eliminating hallucination. The tools hallucinated between 17% and 33% of the time, which is less than a general-purpose chatbot but far from zero [11].

C. 2024: Adaptive Retrieval, Evaluation, and Context Use

Self-RAG teaches a model to select when to retrieve and to evaluate its own output using unique reflection tokens [12]. The authors report gains over ChatGPT and retrieval-augmented Llama2-chat on several tasks, at the price of special training and extra inference steps. Chen et al. benchmark four abilities that RAG needs, namely noise robustness, negative rejection, information integration, and counterfactual robustness, and find that the evaluated LLMs still struggle with all of them [13].

Xiong et al. introduce MIRAGE, a medical question-answering benchmark, and use it to compare combinations of corpora, retrievers, and LLMs [14]. According to Liu et al., models perform better with information at the beginning and finish of a long context compared to the middle. [15], which affects how retrieved passages should be ordered and how many should be supplied.

D. 2023: Surveys and In-Context Retrieval

Gao et al. organise RAG into naive, advanced, and modular paradigms and summarise tasks and evaluation practices [6]. The cited version is an arXiv preprint. Ram et al. show that retrieved documents can simply be prepended to the input of an unmodified language model, which removes the need to train the generator jointly with the retriever [8].

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)

E. 2021: Fusion-in-Decoder

Izacard and Grave encode each retrieved passage together with the question independently, then let the decoder attend to all encoded passages jointly [5]. Encoder cost grows linearly with the number of passages, and answer accuracy improves as more passages are supplied, which gave state-of-the-art results at publication. The trade-off is that decoder-side attention and memory grow with the number of passages, so latency rises.

F. 2020: Foundational Work RAG.

Lewis et al. combine a BART generator with a DPR retriever over about 21 million Wikipedia passages, fine-tuning the query encoder and generator simultaneously while retaining a fixed document encoder and index [1]. They claim more exact and factual generations than a parametric-only model, as well as state-of-the-art results on a variety of open-domain question answering tasks at the time, such as Natural Questions. The fact that the baselines were task-specific retrieve-and-extract systems and parametric sequence-to-sequence models rather than modern instruction-tuned LLMs is a critical qualification. Statements that RAG "outperforms LLMs" therefore need care.

Dense Passage Retrieval (DPR). Karpukhin *et al.* replaced sparse term matching with a dual-encoder retriever built from two BERT models and trained with a contrastive objective that uses in-batch and BM25-mined negatives [4]. For a question q, a positive passage p^+, and negatives p_j^-, define $s^+ = \mathrm{sim}(q, p^+)$ and $s_j^- = \mathrm{sim}(q, p_j^-)$. The loss is

$$L = -\log [\ \exp(s^+) / (\ \exp(s^+) + \Sigma_j \exp(s_j^-)\ )\ ] \quad (5)$$

On open-domain question answering datasets, the authors report a 9% to 19% absolute improvement in top-20 passage retrieval accuracy over a strong BM25 baseline. They further show in a sample-efficiency experiment that roughly one thousand training questions are already adequate to surpass BM25 [4]. This qualifies the common statement that DPR "requires large labelled datasets": supervision is needed, but the original paper shows good sample efficiency.

REALM. Guu et al. integrate retrieval into language model pre-training. A mask language model which performs end-to-end learning to retrieve documents to fill masked tokens and is updated out-of-band [2]. For open-domain question answering, our results outperform the previous state-of-the-art by 4% to 16% in absolute terms. The problem with end-to-end models with a large corpus is that they become computationally prohibitive to train.

G. Relationships and Evolution of Architectures

The first four studies focus on solving different problems. DPR improves what is retrieved [4], REALM improves how retrieval is learned [2], FiD improves how many passages can be used [5], and RAG provides the general retrieve-then-generate formulation [1]. Each study has its own limitations, so later researchers have started combining their approaches to overcome these gaps.

The earliest retrieved documents show evidence of three important developments. The first is a reduction in coupling. Unlike RAG and REALM, which use end-to-end learning to jointly optimize their components, in-context RALM is completely independent of its language model [8]. Second, there is adaptivity and the ability of a system to check the quality of the retrieved document, as in Self-RAG [12]. Third is modular design where retrieval can be structured in increasingly advanced ways (Gao et al [6]). Edge deployment

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)

[9] and security [10] extend the picture at the system level. Other architectural lines, such as retrieval at trillion-token scale, black-box retriever adaptation, corrective evaluators, and graph-based indexing, are outside the retained reference base and are not assessed here.

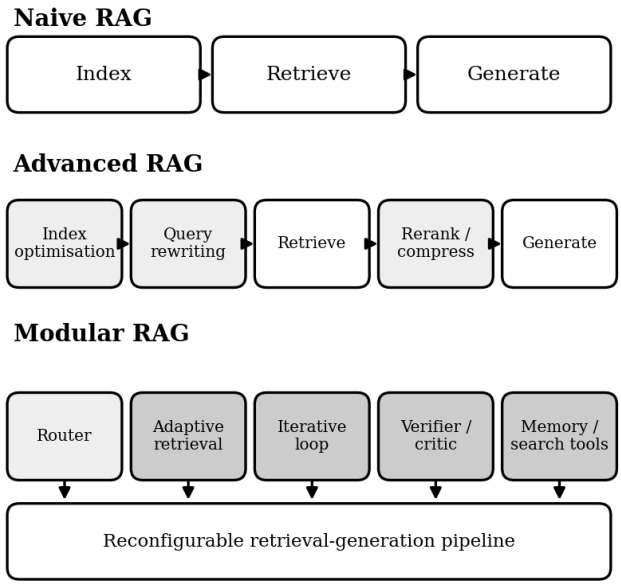


Fig. 2. Conceptual evolution of RAG paradigms following the naive, advanced, and modular categories of Gao et al. [6]. This is an explanatory illustration prepared by the author.

IV. COMPARATIVE ANALYSIS OF EXISTING RAG APPROACHES

TABLE I

CHRONOLOGICAL LITERATURE MATRIX OF THE RETAINED STUDIES (LATEST TO OLDEST)

Study	Year and type	Core idea	Reported finding (by authors)	Limitation / relevance
T. Gao *et al.* [9]	2026, preprint	Federated dual-system RAG with compact QA memories	Up to 7.8% accuracy gain and 8.4x lower latency vs. own baselines	Own results; unreplicated preprint
Xu *et al.* [10]	2026, preprint	Taxonomy of RAG attacks and defences over a knowledge-access pipeline	Organises attacks, defences, and evaluation; names two structural mismatches	Survey; preprint

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)

Study	Year and type	Core idea	Reported finding (by authors)	Limitation / relevance
Sharma [7]	2025, preprint (under review)	Taxonomy of RAG architectures	Identifies recurring trade-offs and open challenges	Survey; not yet peer reviewed
Huang *et al.* [3]	2025, journal (ACM TOIS)	Survey and taxonomy of LLM hallucination	Separates factuality from faithfulness hallucination; discusses limits of retrieval	Survey; not RAG-specific
Magesh *et al.* [11]	2025, journal (J. Empirical Legal Stud.)	Preregistered test of commercial legal RAG tools	Tools hallucinated 17% to 33% of the time, less than a general chatbot	Specific products and time of testing
Asai *et al.* [12]	2024, conf. (ICLR)	Adaptive retrieval with reflection tokens	Reported gains over ChatGPT and retrieval-augmented Llama2-chat on several tasks	Requires special training; added inference steps
Chen *et al.* [13]	2024, conf. (AAAI)	Benchmark of four RAG abilities	LLMs still struggle with noise, rejection, integration, and counterfactuals	Evaluates existing LLMs; not a method
Xiong *et al.* [14]	2024, conf. (Findings of ACL)	MIRAGE medical RAG benchmark	Retrieval surpasses chain-of-thought prompting by up to 18% in six LLMs	Benchmark, not clinical validation
Liu *et al.* [15]	2024, journal (TACL)	Study of position of evidence in long context	Performance is highest at the start or end of context and drops in the middle	Motivates reranking and compression
Gao *et al.* [6]	2023, preprint	Survey with naive, advanced, and modular paradigms	Summarises tasks and evaluation practices	Survey; preprint
Ram *et al.* [8]	2023, journal (TACL)	Prepend retrieved text to a frozen LM	Improves language modelling without changing the LM	Prompt length and ordering sensitivity
Izacard and Grave [5]	2021, conf. (EACL)	Encode passages independently, fuse in decoder	Accuracy rises with number of passages; state of the art at publication	Latency and memory grow with passage count

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)

Study	Year and type	Core idea	Reported finding (by authors)	Limitation / relevance
Lewis *et al.* [1]	2020, conf. (NeurIPS)	DPR retriever plus BART generator; RAG-Sequence and RAG-Token	More factual creation than parametric-only models; state-of-the-art on many open-domain QA tasks	Baselines were not modern LLMs; index fixed during training
Karpukhin *et al.* [4]	2020, conf. (EMNLP)	Dual-encoder BERT retriever, contrastive training	9% to 19% absolute top-20 gain over BM25; surpasses BM25 with about 1,000 training examples	Needs supervision; BM25 stronger on one dataset
Guu *et al.* [2]	2020, conf. (ICML)	Retrieval-augmented masked LM pre-training	4% to 16% absolute gain on open-domain QA	High pre-training cost; large index maintenance

A. Retrieval Methods

Sparse retrieval, of which BM25 is the standard example, scores documents by weighted term overlap and is the baseline against which DPR was evaluated [4]. Dense retrieval maps text to vectors so that semantically related text is close even without shared words [4]. It is not uniformly better: the DPR authors note that BM25 remains stronger on SQuAD, a dataset in which questions share many words with the answer passage [4]. They also report that a combination of sparse and dense scores improves over either alone, which is the basis of hybrid retrieval [4]. Table II summarises the differences. No single method dominates, as their effectiveness varies based on query type, corpus, cost, and the evidence from [4] indicates that dense retrievers should not be assumed to universally replace lexical signals. Retrieval can furthermore be learned jointly with the language objective, as in REALM [2], at the expense of increased training costs.

TABLE II

THEORETICAL COMPARISON OF RETRIEVAL PARADIGMS

Paradigm	Representation	Strength	Weakness
Sparse (BM25) [4]	Term weights	No training; strong with high word overlap	Vocabulary mismatch
Dense dual encoder [4]	One vector per passage	Semantic matching; fast search	Needs training; weaker on SQuAD [4]
Hybrid [4]	Sparse plus dense	Complementary signals	Score fusion and tuning needed
Learned end to end [2]	Retriever trained in pre-training	Retrieval tuned to the language objective	High pre-training cost

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)

B. Embedding Similarity and Vector Search

Dense retrieval reduces to nearest-neighbour search. Besides the inner product in (1), cosine similarity normalises vector length:

$$\cos(u, v) = u \cdot v / (\|u\| \|v\|) \quad (6)$$

A search in a database of N vectors of dimension d requires O(Nd) operations per query, which for large N is prohibitively expensive. Therefore, practical systems use approximate nearest neighbour indexes that sacrifice some recall for significant gains in speed; this is the approach in RAG and DPR pipelines [1] [4]. Commercial and open source vector databases and orchestration frameworks referenced in the synopsis are mostly documented in engineering literature, not in academic papers, so this work does not evaluate them.

C. Context Selection and Generation Strategies

Once evidence is retrieved, the system must decide how much to pass to the generator and in what form. Four patterns appear in the retained literature (Table III). Concatenation in the prompt is simple and works with frozen models [8]. Passages are encoded individually utilizing fusion-in-decoder, which then enables the decoder to fuse them [5]. Marginalisation inside the output probability [1] and retrieved text in the encoder input during pre-training [2] tie evidence to training. Adaptive designs decide whether to retrieve and critique what was retrieved [12].

Context quality is as important as quantity. According to Liu et al., models make better use of information at the start and completion of a long context than information in the middle [15], so adding more passages can reduce accuracy if the key passage is poorly placed. Chen et al. demonstrated that LLMs can be misled by noisy documents and frequently fail to reject questions that the evidence cannot answer [13]. Their results provide additional motivation for adopting reranking, compression, and selective retrieval as core operations, discarding the idea of optional operations, as in the advanced and modular paradigms of [6].

TABLE III

COMPARISON OF CONTEXT-INTEGRATION ARCHITECTURES

Approach	Where evidence enters	Training coupling	Main advantage	Main trade-off
RAG [1]	Passages marginalised in the output probability	Query encoder and generator fine-tuned; index fixed	General, interpretable formulation	Quality bounded by retrieval; fixed index
REALM [2]	Retrieved text in the encoder input during pre-training	End-to-end pre-training with learned retrieval	Retrieval learned from the language objective	High pre-training cost
FiD [5]	Passages encoded separately, fused in decoder	Fine-tuned generator; separate retriever	Scales to many passages	Latency and memory grow with passage count

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)

Approach	Where evidence enters	Training coupling	Main advantage	Main trade-off
In-context RALM [8]	Retrieved text prepended to the prompt	None (frozen LM)	Plug-and-play with any model	Prompt length; position sensitivity [15]
Self-RAG [12]	Adaptive retrieval and critique of evidence	Reflection-token training	Handles poor or unneeded retrieval	Special training; extra inference steps

D. Trade-offs

The comparison shows three tensions, which are the same as those identified by the survey of Sharma [7]. First, precision versus coverage trade-off: retrieving more passages results in a higher probability of containing the answer [5] as well as noise [13], [15]. Second, quality versus latency: critique, iterative retrieval, and reranking improve the quality of evidence, but require sequential steps [12]. As a simplified illustration prepared for this paper, and not taken from a cited study, end-to-end latency can be decomposed as

$$T \approx T_{\text{enc}}(x) + T_{\text{ANN}}(N, k) + T_{\text{rank}}(k) + T_{\text{prefill}}(|x| + k\ell) + T_{\text{dec}}(|y|) \qquad (7)$$

where ℓ is the average passage length. The prefill and rerank terms grow with k, while the ANN term depends on index size and recall settings. Third, modularity versus coordination: modular designs are flexible but errors can propagate across components [7].

V. APPLICATIONS OF RAG

This section provides a summary covering healthcare, finance, legal services, education, and chatbots. The systems described in the papers used for the summary, not the author's created systems, are represented in Table IV. Only healthcare, legal services and decentralized edge deployment are supported by the primary sources. Results in the studies cited in this section were inconsistent. Some sources provided benchmarks with control while others provided reviews on a specific product.

A. Healthcare Information Systems

Xiong et al. analyzed combinations of corpora, retrievers, and LLMs and published MIRAGE, a benchmark comprised of 7,663 questions from five medical question-answering datasets [14]. According to their results, retrieval elevates GPT-3.5 and Mixtral to GPT-4 levels and enhances the accuracy of six LLMs by up to 18% over chain-of-thought prompting. The best arrangement is depending on the combination of retriever and corpus. This supports RAG for medical question answering, but it is a benchmark result, not evidence of safe clinical use.

B. Legal Document Retrieval

Magesh et al. conducted a study on RAG systems in commercial legal research and technology. Before the study, commercial providers marketed RAG systems as a solution to hallucinations. Magesh et al. observed that RAG systems reduce hallucinations, but the hallucinations still exist. During the study, 17-33% of the responses of the RAG systems were hallucinations. Magesh et al.'s findings refute the popular belief that using RAG removes the problem of hallucinations.

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)

C. Decentralised and Edge Settings

FD-RAG is a system that allows documents and raw information to remain on a device. It can still be used for question-answering with better accuracy and faster latency over other systems [9]. Unlike most RAG systems, FD-RAG does not consider a specific field and has an outcome based on an unreplicated preprint.

D. Areas Not Appraised

There are some chatbot domains, for instance, education, finance, enterprise support, and general chatbots that lie within the scope of this survey. However, their primary source studies were excluded from this survey, and consequently, this survey does not comment on the performance of RAG in these areas. The validation for RAG in these domains is important and the reader should refer to [6], [7] for initial research in these areas.

TABLE IV

SUMMARY OF APPLICATION EVIDENCE FROM THE RETAINED LITERATURE

Domain	Study	Reported evidence	Caveat
Healthcare	Xiong *et al.* [14]	Up to 18% accuracy gain over chain-of-thought; lifts some models to GPT-4 level on MIRAGE	Benchmark, not clinical validation
Legal	Magesh *et al.* [11]	Commercial RAG legal tools hallucinate 17% to 33% of the time, less than a general chatbot	Specific products and time of testing
Edge / federated	T. Gao *et al.* [9]	Up to 7.8% accuracy gain, 8.4x lower latency vs. own baselines	Own results; unreplicated preprint

VI. RESEARCH GAPS AND OPEN CHALLENGES

Each gap in the synopsis is examined below using three headings: established findings, partially addressed problems, and open questions. Security is included as a sixth theme because it is one of the synopsis sources. The assessment is restricted by the reference sources retained, and where a gap would require sources outside it, this is indicated.

A. Multi-Hop Reasoning

Multi-hop questions draw from many docs at times which may be in a certain order. In the past we saw mostly single step open domain question answering but present in FiD are a mix of passages at once in parallel which is not in a chain. Also a paper which presented a modular set up used iterative and adaptive retrieval and Self-RAG determines when to get new info. Chen et al. look at models which combine info from more than one doc and report that they still do which is poor. Thus the issue that RAG has issues with multi-hop reasoning does play out but this review doesn't go in depth into the new methods and multi-hop benchmarks which are out of the scope here. What stands out is the issue of robustness. If a early doc in the sequence is missing the whole thing breaks down. Also retrieval is usually looked at per query not the full

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)

reasoner chain and as you add in each round of retrieval the cost goes up. Sharma also brings up that structured reasoning across multi-hop info is very much an open issue.

B. Retrieval Relevance and Context Quality

Established: Answer quality is a function of the relevant of the presented evidence and also models' sensitivity to noise and evidence positions[1], [13], [15]. Partially addressed: adaptive retrieval[12] and hybrid retrieval[4] improve in some cases. Open: what is the best way to evaluate and improve context quality during generation and what is the best way to combine retrieved and parametric knowledge.

C. Hallucination Evaluation

Retrieval reduces hallucination in some settings, for example by producing more specific and factual generations than a parametric-only model [1] and by improving medical question answering [14], but not completely, as the legal study shows [11]. The synopsis states that there are no standardised benchmarks. A more precise reading of the literature is that benchmarks and evaluation practices exist but are fragmented. The benchmarks of information integration, counterfactual robustness, negative rejection, and noise robustness are a few examples [13], the domain benchmark in medicine [14], the preregistered legal evaluation [11], and the evaluation practices summarised in [6]. The hallucination taxonomy of [3] supplies shared vocabulary but not a metric. What is missing is a consensus on a small set of well-vetted metrics that have demonstrated a degree of correlation with human judgment and provide a way to compare across systems and domains. This challenge is partly addressed and partially open in its standardization.

D. Latency and Computational Efficiency

The cost in latency for a high number of passages in the FiD method has been recognised [5], as have the methods that add an inference step to the pipeline [12]. Using a smaller memory with the FD-RAG method has resulted in significant decreases in latency when used in an edge setting for more frequently asked questions, but the results come from a study and the author's own baseline and thus is not a high quality study. Not one of the studies that was retained included a comprehensive analysis that spans different architectures on the trade-off of quality versus latency. Therefore, a general comparison has yet to be done. Equation (7) is not meant to be more than a simple illustration.

E. Domain Adaptation and Specialised Knowledge

The cost in latency for a high number of passages in the FiD method has been recognised [5], as have the methods that add an inference step to the pipeline [12]. Using a smaller memory with the FD-RAG method has resulted in significant decreases in latency when used in an edge setting for more frequently asked questions, but the results come from a study and the author's own baseline and thus is not a high quality study. Not one of the studies that was retained included a comprehensive analysis that spans different architectures on the trade-off of quality versus latency. Therefore, a general comparison has yet to be done. Equation (7) is not meant to be more than a simple illustration.

F. Reliability and Security

RAG technology is opening models up to new ways to enter and exit. One of the many new attack vectors that researchers have started to discover are structural issues between the attacks, defenses and evaluation systems. It's safe to say there is no perfect security system right now, but a survey by Xu et al. provides good guidance. More security issues stem from RAG technology being deployed, especially in terms of

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)

data and user privacy, so researchers are proposing to move away from a central model, and instead explore federated systems.

All of this research and its proposals should be taken with a grain of salt as it is currently pre-print. More studies will hopefully occur in the near future and we will have a better idea on the security of these systems, specifically for systems using a poisoned knowledge-base.

TABLE V

RESEARCH GAPS, EXISTING PROGRESS, REMAINING LIMITATIONS, AND FUTURE DIRECTIONS

Gap	What is established or partly solved	What remains open	Future direction
Multi-hop reasoning	Single-step focus of early work [1], [4], [5]; iterative and adaptive patterns described [6], [12]; integration difficulty shown [13]	Dedicated multi-hop methods not in retained base; robustness to missing bridge evidence; chain-level evaluation; cost [7]	Retrieval planning; per-step verification; chain-aware metrics
Retrieval relevance and context quality	Noise and position effects documented [13], [15]; adaptive retrieval [12]; hybrid retrieval [4]	Joint optimisation with generation; knowledge conflicts	Better reranking and compression; calibrated evidence use
Hallucination evaluation	Benchmarks and practices [3], [6], [13], [14]; reduction shown in some settings [1], [14]	No consensus, human-validated metric set; persistent errors [11]	Unified, validated evaluation frameworks; citation faithfulness checks
Latency and efficiency	Cost of many passages known [5]; extra inference steps [12]; edge-oriented design [9]	No general quality-latency model across architectures	Efficient and adaptive architectures; standard reporting of latency
Domain adaptation	Medical benchmark [14]; legal evaluation [11]	Multilingual and low-resource validation; finance and education not evidenced here; domain safety	Domain and multilingual RAG; expert-in-the-loop evaluation
Reliability and security	Attack and defence taxonomy [10]; privacy motivates federated designs [9]	Defences without blind spots; adaptive evaluation of defences	Provenance, access control, adaptive evaluation of defences

VII. FUTURE RESEARCH DIRECTIONS

Improved retrieval and reranking. As evidenced in the past, models are sensitive to noise and order in their inputs [13, 15]. The combination of hybrid retrieval systems, incorporating both semantic and lexical signals, with downstream components (rerankers and compressors) to properly order the evidence as models prefer, makes sense given these past results [4].

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)

Previous research, such as in [3] and [12], discusses methods to encourage faithfulness in AI and ensure grounding. Such methods involve training models to provide the source of the answer they provide and devising methods to measure whether the provided source actually answers the question.

The methods to assess RAG models are not as consistent as one would hope. There is little standard in the way they are developed and even in the ways that different research studies address RAG. Research publications and developers are often the ones that create these issues, leaving users in a difficult spot when trying to interpret or compare different RAG methods. One solution would be to establish consistent standards across the field of RAG. This means that publications would need to report the methods and the evaluation results using the same metrics and have standard procedures for validation. Consistency across RAG studies also means reporting results on the dimensions of accuracy, faithfulness, and latency together.

Adaptive and efficient models. Studies in this field deal with how much context of the answer to the prompt can be extracted from external documents and then should be added to the context window of the model, and when it is optimal for the model to stop generation [5, 12, 9]. Privacy oriented models are another area that researchers are investigating and how these privacy requirements might impact the structure of RAG systems [9].

Retriever design for multiple-hop and chain-aware steps. Unlike earlier retrieval research, the retriever of the future should incorporate an automated retrieval verification step that accounts for planning and the capability to evaluate chains of reasoning.

Using domain-specific and multilingual RAG systems. Creating RAG data collections for specialized domains, like medicine and law, is better done with the input of domain experts [14, 11]. There also needs to be a lot more work done on RAG for languages other than English.

Security and Robustness. Further research needs to explore RAG protection against malicious attacks that may occur during access to the data, along with a number of threats which may be generated by highly intelligent and advanced users. [10].

VIII. CONCLUSION

This work focuses on providing a grounded analysis of Retrieval-Augmented Generation (RAG) by systematically analyzing 15 existing research papers. Based on these sources, I show that incorporating retrieved evidence into generation is a quick way to improve the system's ability to remain grounded to the truth and update the model without retraining. The early contributions by others in the field have highlighted the gaps in retrieval, training, and combining of generated text. Later contributions have developed more adaptive, modular, and context aware retrieval [12], [8], [6]. One shortcoming that is NOT addressed is whether RAG helps with hallucinations. Research has shown that RAG actually has the tendency to introduce more hallucinations in certain scenarios [11]. The performance of the system relies on how much and the quality of the information that is provided to it, regardless of whether or not it has the potential to introduce more information than necessary.

I offer a study that analyzes RAG from multiple angles. It describes the framework of RAG by unifying previous work and breaks down different architectural styles, retrieval systems, and existing use cases of the system. The gaps listed in the summary are classified and assigned a grade. I explain in detail what each grade means. For instance, some points in the original RAG paper that seemed addressed have been evaluated and now fall into the 'open' category. The points on DPR and evaluation of RAG in the summary also needed refining and clarification.

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)

This review has some limitations. One of these is that it doesn't have a registered protocol, so there's a risk that some important information has been left out. There are also results that are not directly comparable because they come from different sources and four of the sources are still in the preprint phase. The speed at which new information in this area is generated makes the evidence outdated. Because of the focus of this review, some topics that should be covered like multi-hop retrieval and the application of graph-based indexing have been excluded. Overall, this review provides no new evidence and suggests some more research in the form of validation of the current methods and developing more efficient methods for information retrieval and evaluation in different contexts.

REFERENCES

[1] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 9459-9474.
[2] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 119, 2020, pp. 3929-3938.
[3] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, 2025, doi: 10.1145/3703155.
[4] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.
[5] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2021, pp. 874-880, doi: 10.18653/v1/2021.eacl-main.74.
[6] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," arXiv:2312.10997, 2023 (preprint).
[7] C. Sharma, "Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers," arXiv:2506.00054, 2025 (preprint, under review).
[8] O. Ram *et al.*, "In-context retrieval-augmented language models," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 1316-1331, 2023, doi: 10.1162/tacl_a_00605.
[9] T. Gao, K. Yang, and Y. Li, "FD-RAG: Federated dual-system retrieval-augmented generation," arXiv:2605.27432, 2026 (preprint).
[10] Y. Xu *et al.*, "Securing retrieval-augmented generation: A taxonomy of attacks, defenses, and future directions," arXiv:2604.08304, 2026 (preprint).
[11] V. Magesh *et al.*, "Hallucination-free? Assessing the reliability of leading AI legal research tools," *J. Empirical Legal Stud.*, vol. 22, no. 2, pp. 216-242, 2025, doi: 10.1111/jels.12413.
[12] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.
[13] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 16, 2024, pp. 17754-17762, doi: 10.1609/aaai.v38i16.29728.
[14] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings Assoc. Comput. Linguistics: ACL 2024*, 2024, pp. 6233-6251, doi: 10.18653/v1/2024.findings-acl.372.
[15] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 157-173, 2024, doi: 10.1162/tacl_a_00638.