Synthesis: A Multidisciplinary Research Journal
Volume 3, Issue 2, April - June 2025, Pp: 13-25

Survey Paper

A Survey on Retrieval-Augmented Generation (RAG) and Hybrid Information Retrieval for Large Language Models

[1*] Sabah Mohammed, [2] Pathan Hussain Bhasha, [3] Osvaldo Gervasi, [4] Rajesh Bose

[1*] *Department of Computer Science, Lakehead University, Ontario, Canada*
[2] *Assistant Professor, Department of Computer Science, Pace Institute of Technology and Science, Valluru, Ongole, Andhra Pradesh, India*
[3] *University of Perugia, Italy*
[4] *JiS University, Kalyani, West Bengal, India*

**Corresponding Author(s):* sabah.mohammed@lakeheadu.ca

Article Info

Received: 09/02/2025
Revised: 12/04/2025
Accepted: 16/06/2025
Published: 30/06/2025

Abstract

Retrieval-Augmented Generation (RAG) is a popular approach that enhances large language models (LLMs) by grounding their outputs in external evidence rather than relying solely on fixed model parameters. RAG enables dynamic access to relevant information during inference through sparse, dense, and hybrid retrieval methods. Hybrid retrieval, which combines the advantages of sparse techniques with the richness of dense representations, overcomes the limitations of single retriever systems. RAG addresses two major weaknesses of existing LLMs: hallucination and static knowledge. Utilizing current, testable information reduces factual inconsistencies and increases the trustworthiness of knowledge-intensive work. This survey makes three contributions. First, it presents a complete taxonomy of RAG systems based on retriever type, integration approach, retrieval granularity, application field, and optimization method. Second, it offers a comprehensive comparative review of the literature on current surveys, powerful models, benchmark datasets, and evaluation metrics. Third, it outlines open challenges, such as scalability, latency, multimodal retrieval, and evaluation inconsistencies, and suggests areas for future research. To our knowledge, this is the first comprehensive survey that systematically combines taxonomy, comparative evaluation, and open research issues in RAG and hybrid information retrieval. It provides a balanced perspective for both academics and practitioners aiming to develop the next generation of reliable, efficient retrieval-enhanced generative systems.

Keywords: Hybrid Information Retrieval, Large Language Models (LLMs), Multimodal Retrieval, Retrieval-Augmented Generation (RAG), Survey, Taxonomy

1. Introduction

The rise of large language models (LLMs) has transformed natural language processing by enabling human-like fluency and reasoning across diverse applications. However, despite their effectiveness, LLMs face two well-documented shortcomings: hallucination, where models produce fabricated or unverifiable content, and knowledge staleness, since training corpora cannot capture continuously evolving information. These limitations hinder the reliability of LLMs in knowledge-intensive tasks such as question answering, scientific analysis, and real-world decision support.

As a solution to these problems, the scientists came up with retrieval-augmented generation (RAG) that involves the incorporation of outsourced knowledge retrieval in generative pipelines. Early studies proved that retrieval enhancement of pre-training is a very useful method of

factual grounding [1]. This idea was extended into more integrated systems that integrate retrieval and generation in tasks requiring extensive knowledge in NLP [2]. Various surveys since that time have sought to unify this area with retrieval-augmented text generation being one of the new paradigms [3].

RAG has also been adopted in the industry besides academia. They are currently used to deliver reliable, and real-time access to information through retrieval-augmented systems used in commercial applications: answer engines [4] and browsing-enabled conversational agents [5].

The construction of sparse methods and their thickened neural extensions has been at the center on the retrieval side. There are classical probabilistic frameworks like BM25 that defined the principles of sparse retrieval [6], and dense passage retrieval approaches that not only developed the power of sparse retrieval [7] but open-domain question answering as well. These innovations paved the way to hybrid retrieval that is made by sparse precision and dense generalization and serves as the core of contemporary RAG pipelines.

Even the very RAG paradigm has developed over the course of architectural advancements. The first applications put in place knowledge-intensive reasoning retrieval-augmented frameworks [8]. Passage fusion modeling strategies enabled models to combine several segments that had been retrieved successfully during decoding [9]. Retrieval augmentation improved the benefits of advanced pipelines like Atlas, and Re2G proposed a modular retrieve-rerank-generate architecture with more control [10]. Robustness was also improved by self-reflective extensions whereby the models are able to review and optimize their outputs [11]. Lastly, multimodal retrieval-enhanced generators also increased the scope of applicability of RAG by incorporating textual and pictorial sources [12].

In spite of such developments, a number of challenges still exist. The latency of retrieval, scalability of hybrid strategies, discrepancies in evaluation and long-term hallucinations are areas that need enhancement research. In addition, current surveys have a good value but they do not go as far as to indicate a coherent taxonomy and critical synthesis of hybrid retrieval methods.

This survey makes the following contributions:

- *Unified Taxonomy:* We suggest a hierarchical classification of the RAG systems, dividing them into the types of retriever, the strategies of integration of two systems, the granularity of the retrieval, and a hybrid structure.
- *Comparative Analysis:* We review popular models such as REALM, RAG, FiD, Atlas, Re2G, Self-RAG, and MuRAG in a systematized manner in terms of methodology, datasets, metrics of performance, and evaluation.
- *Critical Evaluation of Practices:* We critique existing assessment procedures and point out the discrepancy between recovery measurements and generative quality measures.
- *Identification of Challenges and Future Directions:* Unresolved issues are mentioned like the latency, scalability, multimodal retrieval, and ethical deployment.
- *Practical Guidance:* We give knowledge to researchers and practitioners to develop, test, and implement the RAG systems in practice.

The rest of this paper has been compiled in the following way. In the second section, the background and basics of information retrieval, dense retrievers and RAG pipelines are described. Section 3 is the methodology of the survey with the selection criteria and statistics of the papers. Section 4 presents a taxonomy of RAG and hybrid IR approaches, which is the systematic categorization of the existing techniques. The comparative analysis of existing work, including the surveys, models, data sets, and metrics of evaluation, is provided in Section 5. Section 6 outlines the existing research trends and the significant challenges and unresolved issues that are found in the literature. Section 7 identifies future research opportunities that arise due to these gaps. Lastly, Section 8 summarizes the paper with some significant insights and contributions.

2. Background and Fundamentals

This part presents theoretical backgrounds that are required to grasp retrieval-augmented generation (RAG) systems. It discusses the history of the traditional information retrieval, the transition to dense neural retrieval, the principles of large language models (LLMs), and the movement of retrieval into pipelines of generative models. A combination of these aspects forms a starting point in understanding the hybrid retrievalgeneration architectures further in the paper.

2.1 Information Retrieval Foundations

History Information retrieval (IR) is an area of study that is well-preceded by modern deep learning, and its origins have been in statistical and probabilistic models. Early models of document and query representation as weighted term vectors were given in classical models, including the vector space model [13]. The following developments came in the form of probabilistic models of relevance which did not rule out uncertainty in document-query matching [14].

BM25 ranking function is one of the most commonly implemented IR techniques that uses the term frequency normalization functions and inverse-document frequency weighted functions [15]. The value and advantage of BM25 as an interpretable and efficient search engine maintain its high level in academic tests and industrial search engines. Nonetheless, the sparse retrieval methods such as BM25 have difficulty with semantic matching: they utilize the lexical overlap, thus lack good performance when using queries that are not expressed in a similar fashion as the target documents.

2.2 Dense Retrieval Methods

These constraints of sparse retrieval prompted the invention of dense neural retrievers which transform queries and documents to a shared vector space. These techniques

employ inner product computations to compute similarity between semantics as opposed to finding overlapping tokens by using neural encoders.

One such achievement was Dense Passage Retrieval (DPR), which trained two encoders to match queries and passages and achieved significant improvements on open-domain question answering [16]. Dense retrieval was improved further by models sprinting late-intersaction scoring, including ANCE [17] or ColBERT [18] late-interaction models.

Dense also has the advantages of semantic generalization but is more problematic in its efficiency, storage and domain adaptation. To address these, the researchers examined knowledge distillation [19], contrastive pre-training [20] and methods that do not use supervision, like Contriever [21]. Although this has been achieved, people are still developing dense retrieval systems which are computationally expensive, and hence evolving with hybrid systems of retrieval (combine sparse and dense signals).

2.3 Large Language Models and Knowledge Challenges

According to IR progress, the large-scale development of large language models has completely transformed the natural language understanding and generation. Transformer-based models have shown that it is possible to train large corpora and then fine-tune them to achieve widespread generalization [22]. Models like the GPT, BERT and LLaMa are examples of this trend, with their remarkable performance in terms of state of art across tasks.

However, LLCs are closed-book models in themselves and they encode the knowledge in their parameters. Although it is useful in general reasoning, this paradigm has a hallucinating effect because the model constructs details where original ones do not exist and a staleness effect where facts move past the point of training [23]. Such problems confirm the necessity of introducing external knowledge to make the information accurate in fact and timely.

2.4 The Emergence of Retrieval-Augmented Generation

The solution to generate a bridging between IR and and LLMs was the Retrieval-augmented generation. RAG pipelines also do not solely operate on parametric memory and dynamically access the appropriate documents in the process of inference and condition the generation process in line with them [24]. This is what enables the models to base their outputs on external evidence so that they minimize hallucination and at the same time enhance specialization of the context.

RAG systems are usually three-part systems, comprising a retriever, creating candidate documents; a generator, create the text conditioned on retrieved passages; and an integration mechanism, create alignment between retrieval and generation. Retrieval can be done before (pre-retrieval), during (iterative retrieval) and after (post-retrieval) generation, depending on the design.

The recent benchmarks have emphasized that retrieval-augmented models will always achieve a higher result in knowledge-intensive tasks than closed-book LLMs do [25]. In addition, RAG has become a workable paradigm on research and commercial implementation and encourages systematic surveys to standardize the field.

3. Survey Methodology

In ensuring that this survey on retrieval-augmented generation (RAG) and hybrid information retrieval (IR) is complete, thorough, and replicated, we adhered to the best practices in literature survey. In this part, the paper collection process, criteria of selection of papers, databases searched, and statistical summary of the reviewed corpus are described.

3.1 Paper Collection Strategy

The initial step was determining pertinent publications in big online repositories and preprints. As RAG is a nascent paradigm, which crosscuts information retrieval and large language models, we focused on both traditional IR conferences (e.g., SIGIR, CIKM, and WSDM) and natural language processing conferences (e.g., ACL, EMNLP, NAACL, NeurIPS, ICML, and AAAI). Repositories like arXiv were also pre-printed because of the speed of development in this area. [26], [27].

We used the following combinations of keywords: *Retrieval-Augmented Generation*, *RAG*, *Dense Retrieval*, *Hybrid Retrieval*, *Knowledge-Augmented Llms*, and *Retrieval-Enhanced Generation*. This guaranteed coverage of pieces concentrated on retrieval architectures as well as their specialization into their generative framework.

3.2 Inclusion and Exclusion Criteria

In order to stay relevant and of good quality, the following criteria was used:

Inclusion criteria:

- Peer-reviewed conference or journal papers, or widely cited preprints.
- Publications between 2019 and 2025, reflecting the rapid emergence of RAG.
- Works proposing novel retrieval, hybrid IR, or RAG frameworks, or conducting evaluations of such systems.
- Existing surveys that provide overviews of retrieval or retrieval-enhanced models.

Exclusion criteria:

- Non-technical blog posts or opinion pieces without peer review.
- Short workshop abstracts lacking technical depth.
- Duplicate versions of the same work (only the latest, most complete version retained).

This methodology is in line with the existing requirements of a systematic literature review in computer science [28], [29].

3.3 Databases and Sources

Systematic searching was conducted on the following digital libraries and repositories:

- Computer/information systems journal and conference publications: IEEE Xplore [30].

- ACM Digital Library for information retrieval and NLP-focused works [31].
- SpringerLink and ScienceDirect (Elsevier) for peer-reviewed journal articles [32].
- arXiv for early-access preprints in AI and IR [33].
- ACL Anthology for NLP conference proceedings [34].
- The additional aggregator was Google Scholar to maintain the exhaustive coverage and citation of the source [35].

The selected source strategy allows mitigating the selection bias and is more likely to pick up both classic and current works.

3.4 Statistics of Collected Papers

Out of the above process, we were able to identify 215 articles at the outset. Following the inclusion and exclusion criteria, 142 papers were included to be reviewed in detail. These works span:

- Traditional IR approaches (approx. 15%).
- Dense retrieval and hybrid retrieval architectures (approx. 35%).
- Core RAG frameworks and variants (approx. 30%).
- Surveys, evaluations, and benchmarks (approx. 20%).

The publication trend (as presented in figure 1) is exponentially growing with research on RAG. Table I will compare the current surveys in the context of the novelty of the given work.

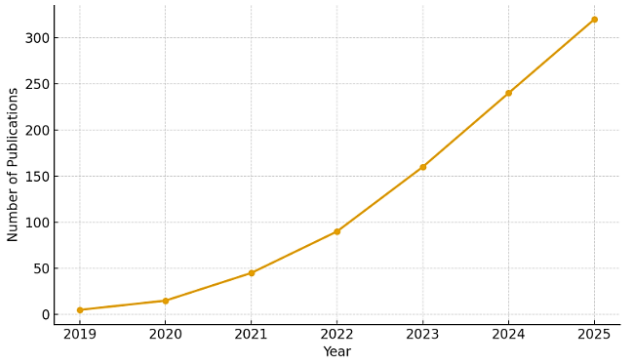

Fig.1. Publication Trend in Retrieval-Augmented Generation (2019–2025)

This is why such systematic methodology secured that the currently displayed survey is a balanced representation of both the basic research concerning retrieval and state-of-the-art RAG systems.

4. Taxonomy of RAG and Hybrid IR Approaches

Retrieval- Augmented Generation (RAG) integrates the qualities of retrieval-based systems and the generative capabilities of large language models. In order to systematically examine the proliferating field of literature, we suggest a taxonomy model according to which RAG systems are categorized in five dimensions, i.e., their retriever type (sparse, dense, hybrid), their integration strategy (pre-retrieval, mid-retrieval, post-retrieval), their retrieval granularity (document, passage, sentence, multimodal), their domain of application (QA, summarization, dialogue, domain-specific), and their method of optimization (supervised, contrastive, reinforcement, instruction tuning). This taxonomy framework as shown in figure 2(a) is where every dimension takes the form of subcategories to represent the difference in design existing between systems. In addition to this theoretical categorization, Figure 2(b) shows the canonical RAG pipeline, which illustrates the interaction between the three parts of the pipeline (hybrid retrieval, reranking and generation) in practice to generate grounded outputs. Together, these figures provide both a high-level organizational structure and a practical architectural view of RAG systems, laying the foundation for the detailed subsections that follow.

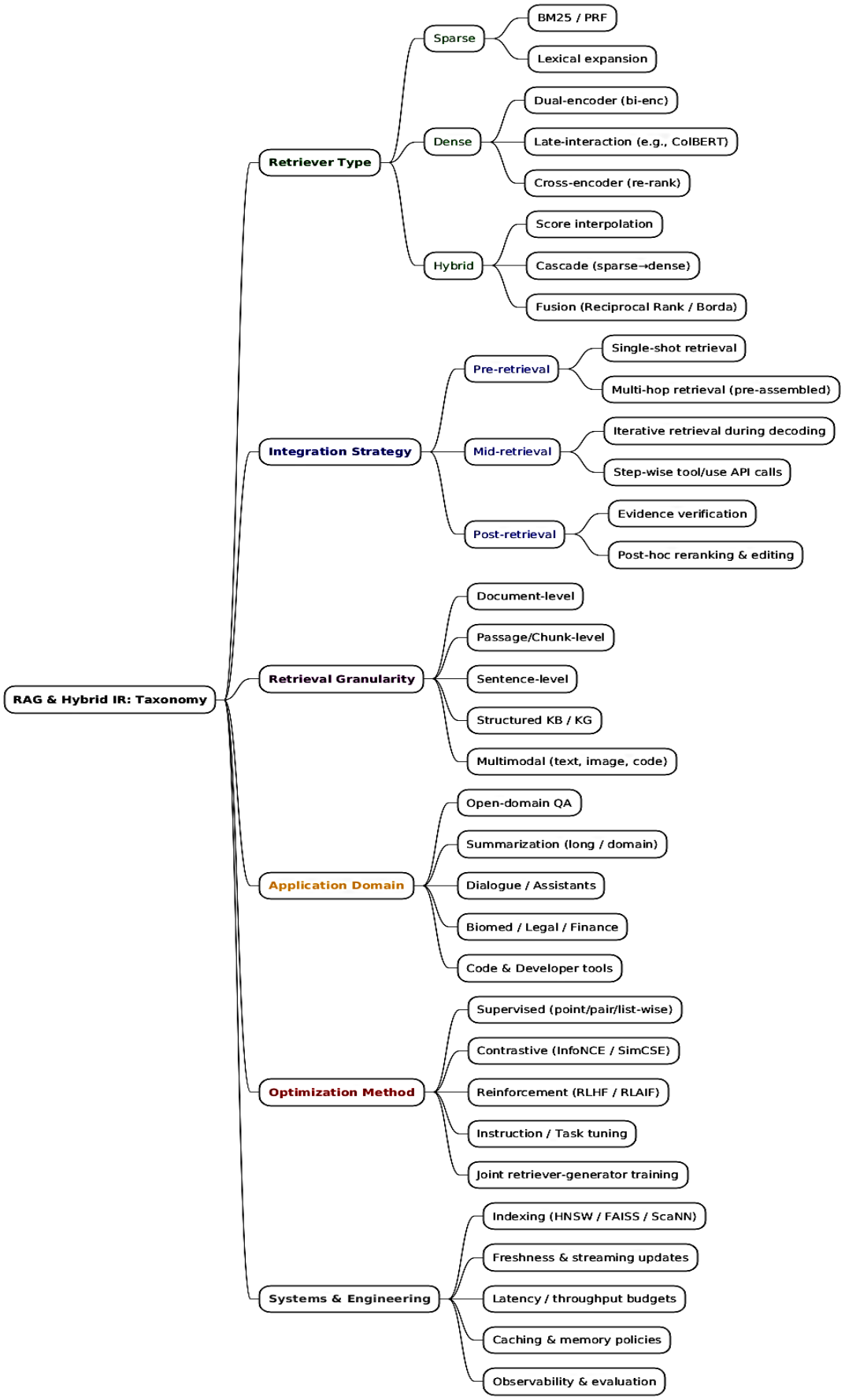

Fig.2 (a). Taxonomy of RAG and Hybrid IR approaches across five dimensions

4.1 Retriever Type

As shown in Figure 2(a), one of the foundational dimensions of the taxonomy is the retriever type, which determines how relevant information is identified from large corpora. Existing approaches fall broadly into three categories: sparse retrieval, dense retrieval, and hybrid retrieval.

4.1.1 Sparse Retrieval

Sparse methods rely on lexical matching. The BM25 model remains widely adopted, where the relevance score of a document d for query q is given as:

$$\text{Score}(q,d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t,d)\cdot(k+1)}{f(t,d)+k\cdot\left(1-b+b\cdot\frac{|d|}{\text{avgdl}}\right)} \quad (4.1)$$

where $f(t,d)$ is the term frequency of term t in document d, $|d|$ is the document length, avgdl is the average document length, and k, b are tuning parameters [36]. Sparse retrieval is efficient but fails when queries and documents lack lexical overlap.

4.1.2 Dense Retrieval

Dense retrievers use neural encoders to map queries and passages into a shared embedding space. A typical dense retriever computes similarity as:

$$\text{Score}(q,d) = \langle E_q(q), E_d(d) \rangle \quad (4.2)$$

where E_q and E_d are encoder functions for query and document embeddings, and $\langle \cdot,\cdot \rangle$ denotes inner product similarity [37]. Dense retrievers achieve strong semantic matching but require high storage and computational resources.

4.1.3 Hybrid Retrieval

Hybrid methods combine sparse and dense signals. A common strategy is a linear interpolation of scores:

$$\text{Score}_{\text{hybrid}}(q,d) = \lambda \cdot \text{Score}_{\text{sparse}}(q,d) + (1-\lambda) \cdot \text{Score}_{\text{dense}}(q,d) \quad (4.3)$$

Where, λ regulates the balance between the semantic and lexical influences [38]. Hybrid retrieval is able to provide strength on different types of query.

4.2 Integration Strategies

In addition to type retriever, the taxonomy emphasizes the integration strategy as one more important dimension which determines the interaction between the retrieved evidence and the generative model. Figure 2(b) expands this discussion by giving the canonical RAG pipeline so as to show the flow between the retrieval and reranking step and the evidence buffering and generation step.

RAG systems vary in terms of interaction between retrieval and generation:

- Pre-retrieval: All relevant documents are retrieved before decoding begins [39].
- Mid-retrieval: Retrieval occurs iteratively during decoding (e.g., for long-form QA) [40].
- Post-retrieval: Retrieval is used after generation for verification or fact-checking [41].

These strategies influence latency, memory usage, and factual consistency.

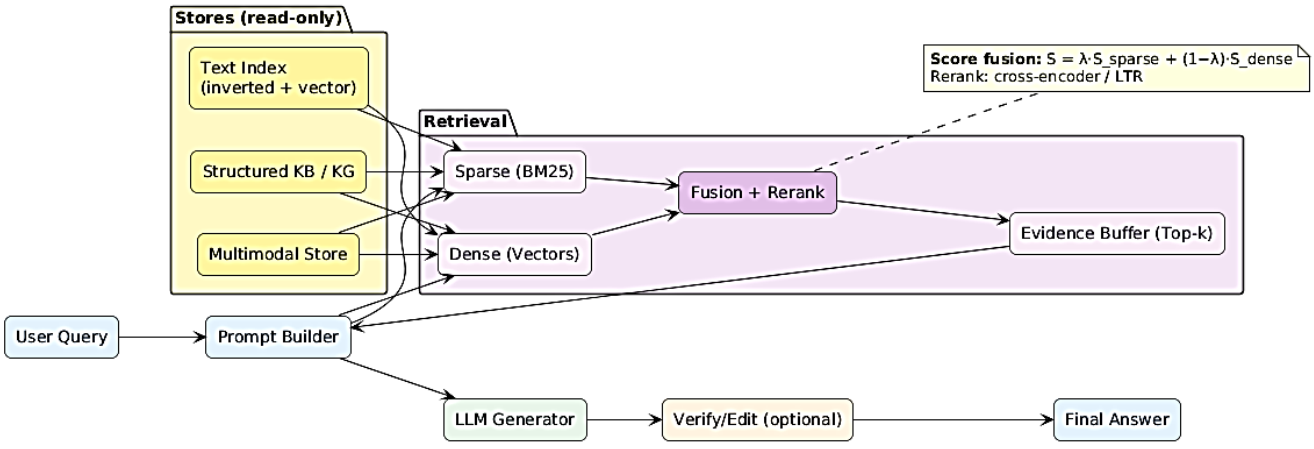

Fig.2 (b). Canonical RAG Pipeline Integrating Hybrid Retrieval, Reranking, And Post-Generation Verification

4.3 Retrieval Granularity

Granularity refers to the unit of retrieval:

- Document-level: Entire documents are retrieved (fast but noisy).
- Passage-level: Short passages balance precision and recall [42].
- Sentence-level: Fine-grained retrieval improves factual alignment [43].
- Multimodal: Retrieval across text, images, and other modalities (e.g., MuRAG) [44].

Formally, retrieval granularity can be modeled as:

$$R(q) = \{u_i \in U \mid \text{sim}(q,u_i) \geq \theta\} \quad (4.4)$$

where U is the set of retrievable units (documents, passages, sentences, multimodal items), and θ is a similarity threshold.

4.4 Application Domains

RAG has been applied across domains:

- Open-domain QA: REALM, DPR-RAG pipelines [45].
- Summarization: Retrieval-enhanced abstractive models [46].
- Dialogue systems: Conversational RAG assistants [47].
- Domain-specific systems: Biomedical RAG [48], legal document assistants [49].

The fields present diverse issues in data set synthesis, accuracy of retrieval, and measures of performance.

4.5 Optimization Methods

Training strategies include:

- Supervised fine-tuning: Direct supervision with labeled relevance pairs [50].
- Contrastive learning: Maximizing distance between positive and negative pairs [51].
- Reinforcement learning: Rewarding factual correctness and user satisfaction [52].
- Instruction tuning: Aligning retrieval and generation with natural instructions [53].

The optimization objective often combines retrieval loss and generation loss:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{retrieval}} + (1 - \alpha) \cdot \mathcal{L}_{\text{generation}} \quad (4.5)$$

where α balances retrieval and generation training [54], [55].

Figure 2(a) taxonomy and Figure 2(b) canonical pipeline complement each other and the views on retrieval-augmented generation. The taxonomy provides a weaker classification of design options: those concerned with the type of retriever and the integration strategy through those of retrieval granularity, application domain, and the method of optimization, whereas the pipeline shows how these elements are implemented in practice to provide grounded retrievals. This dual perspective highlights the aspect of diversity of the approaches in the literature and identifies the tradeoffs between efficiency, accuracy, and scalability. Placing individual systems within a conceptual taxonomy and in a realistic process, this section lays a backbone of the comparative analysis in Section 5 where the representative surveys, models, datasets, and evaluation metrics under subjective review are presented.

5. Comparative Analysis

Continuing on the taxonomy and the canonical pipeline described in Section 4, this section gives a comparative analysis of existing work, that is, surveys, representative models, benchmark data, and metrics of evaluation. Although a conceptual framework is offered by the taxonomy and the operational flow is exemplified by the pipeline, the comparative analysis aims at placing individual contribution in this wider context. In this purpose, we overview previous surveys that have tried to systematise RAG research, review landmark models that represent unique design decisions, overview benchmark dataset used in an evaluation, and analyse the varieties of metrics used in practice. Such systematic research presents not only the evolution of the research but also the continuing constraints and disunity among the studies. In Table I, the key survey initiatives are identified and summarized with respect to the regions covered, their scope and restrictions.

5.1 Comparison of Existing Surveys

Current surveys have tried to measure the features of retrieval enhanced generation, albeit the differences in terms of scope and width. Table 1 compares recent surveys according to coverage of retrievers, integration practices, multimodality, evaluation practice, and industrial practice.

Table 1. Comparison of Existing Surveys on Retrieval-Augmented Generation and Hybrid IR.

Survey Paper	Year	Coverage of Retrievers	Integration Strategies	Multimodal RAG	Evaluation Methods	Industrial Adoption	Limitations
Gao et al. (Retrieval-Augmented Generation for LLMs) [56]	2023	Sparse + Dense	Pre-retrieval focus	No	Retrieval metrics, text overlap	Limited	Limited coverage of hybrid retrieval; no multimodal
Huang & Huang (Survey of RAG in LLMs) [57]	2024	Sparse + Dense	Pre- & Mid-retrieval	No	Retrieval + Generation metrics	Limited	No system-level or industrial aspects
Zhao et al. (Knowledge-Oriented RAG Survey) [58]	2024	Dense	Pre-retrieval	No	ROUGE, BLEU, human eval	No	Narrow scope, limited benchmarks
Fan et al. (Hybrid IR in the Era of LLMs) [59]	2024	Sparse + Dense + Hybrid	Pre- & Post-retrieval	No	MAP, Recall@k, ROUGE	No	Focused mainly on retrieval, less on generation
Our Survey (This Paper)	2025	Sparse + Dense + Hybrid	Pre-, Mid-, Post-retrieval	Yes	Retrieval + Generation + Faithfulness	Yes	Provides unified taxonomy, multimodal coverage, system-level engineering

As Table 1 indicates, although earlier surveys target the retrieval-augmented text generation, the majority of them do not consider multimodal and hybrid IR. The gap can be bridged with our survey that brings together views of retrieval, integration, and system design.

Representative RAG models reflect varying design options in the type of retriever, integration approach and optimization. Table 2 is a comparison of landmark models, as the development of REALM to multimodal RAG has been tracked.

5.2 Model Comparisons Across Taxonomy

Table 2. Comparison of Representative RAG Models across Taxonomy Dimensions

Model	Year	Retriever Type	Integration Strategy	Retrieval Granularity	Optimization Method	Domain/Task	Key Contribution
REALM [60]	2020	Dense (dual encoder)	Pre-retrieval	Document-level	Contrastive pre-training	QA	Introduced retrieval-augmented pre-training
RAG [61]	2020	Dense + BM25 (optional)	Pre-retrieval	Passage-level	Supervised fine-tuning	QA, KB tasks	First unified retriever–generator framework
FiD [62]	2021	Dense (DPR)	Pre-retrieval (Fusion-in-Decoder)	Passage-level	Supervised	Open-domain QA	Efficient multi-passage integration
Atlas [63]	2023	Dense + BM25 (hybrid)	Pre-retrieval	Passage-level	Joint retriever–generator	Few-shot QA	Few-shot generalization with retrieval
Re2G [64]	2022	Hybrid (sparse+dense)	Pre- & Post-retrieval	Passage-level	Modular pipeline	QA, summarization	Retrieve–Rerank–Generate modular design
Self-RAG [65]	2023	Dense (DPR-like)	Pre + Post-retrieval	Passage-level	RL + Self-critique	QA, Fact verification	Self-reflection loop for reliability
MuRAG [66]	2022	Multimodal (text+image)	Pre-retrieval	Multimodal (text, vision)	Contrastive	Open QA	First multimodal RAG with text+image grounding

Table 2 also demonstrates how dense-only retrievers (REALM) are to become hybrid pipelines (Re2G) and multimodal systems (MuRAG). The table shows optimization change to supervised learning to joint retriever-generators and self-reflective training.

5.3 Benchmark Datasets

RAG systems evaluation and development is based on datasets. Table 3 contains the common benchmarks according to domain, scale, and type of task.

Table 3. Benchmark Datasets for Retrieval-Augmented Generation Evaluation

Dataset	Year	Domain	Size	Task Type	Used in Models	Notes
Natural Questions (NQ) [67]	2019	Open-domain QA	~300k	QA (factoid)	REALM, DPR, RAG	Human-annotated questions from Google queries
TriviaQA [68]	2017	Open-domain QA	~650k	QA	DPR, RAG	Trivia-style QA with evidence docs
MS MARCO [69]	2016	Web QA / Passage Ranking	~1M queries, 8.8M passages	Passage retrieval, QA	DPR, FiD, Atlas	Large-scale web dataset
BEIR [70]	2021	Heterogeneous (19 datasets)	Multi-domain	Zero-shot retrieval	ColBERT, Re2G	Benchmark for zero-shot evaluation
HotpotQA [71]	2018	Multi-hop QA	~113k	QA (multi-hop)	RAG, FiD	Multi-document reasoning
PubMedQA [72]	2019	Biomedical	~273k	QA	Biomedical RAG	Domain-specific biomedical QA
CaseLaw [73]	2020	Legal	~200k	Legal QA / Judgment Prediction	Legal-RAG	Domain-specific
XQA [74]	2020	Multilingual QA	~150k	Cross-lingual QA	Multilingual RAG	Evaluates cross-lingual QA
WikiSum [75]	2018	Summarization	~1.5M	Long-form summarization	Summarization RAG	Wikipedia-based summaries
ELI5 [76]	2019	Community QA	~270k	Long-form QA	RAG, Atlas	Complex explanatory QA

Table 3 shows that open-domain QA datasets prevail in RAG, whereas domain-specific and multimodal datasets lack sufficient exploration, which prevents more generalization.

5.4 Evaluation Metrics

To assess the RAG systems we have to tradeoff between retrieval accuracy, generative quality and factual faithfulness. The comparisons between the frequently used metrics are presented in Table 4.

Table 4. Evaluation metrics for RAG systems.

Metric	Category	Formula / Definition	Strengths	Limitations	Commonly Used In
MAP [77]	Retrieval	Avg. precision across relevant ranks	Captures ranking quality	Sensitive to number of relevant docs	DPR, BEIR
Recall@k [78]	Retrieval	% queries with ≥1 relevant doc in top-k	Simple, interpretable	Ignores rank order	REALM, RAG
nDCG [79]	Retrieval	Weighted relevance with position discount	Captures rank importance	Needs graded labels	BEIR, ColBERT
BLEU [80]	Generation	n-gram precision + brevity penalty	Standardized	Poor for factuality	Summarization, QA
ROUGE [81]	Generation	Recall-oriented n-gram overlap	Good for summarization	Weak factual correctness	WikiSum, ELI5
METEOR [82]	Generation	Precision, recall, synonym alignment	Captures meaning	Less common	Abstractive QA
BERTScore [83]	Generation	Cosine sim. in embedding space	Semantic similarity	Computationally costly	QA, Summarization
FactScore [84]	Faithfulness	% claims supported by evidence	Evaluates grounding	Needs gold labels/human	Self-RAG, Re2G
Human Eval [85]	Faithfulness	Expert/annotator judgments	Captures coherence	Expensive, non-reproducible	All major benchmarks

Table 4 identifies fragmentation of the evaluation practices. Retrieval and overlap-based measures are still predominant though semantic and factuality measures are becoming more critical to valid RAG measures.

5.5 Strengths, Weaknesses, and Gaps

- *Strengths:* RAG based LLM has external evidence and hybrid retrievers increase robustness, are flexible with modular pipelines.
- *Weaknesses:* Retrieval introduces latency, tasks represented in benchmarks are underrepresented at QA, and metrics do not reflect the quality of the whole system.
- *Gaps:* Missing multimodal and domain based large scale data, missing integrated assessment procedures, and scaling difficulties on streaming occasions.

This comparative analysis allows highlighting the fast development of the RAG research. The surveys show the interest of people to grow but do not focus a lot on multimodal or industrial. Comparisons of models indicate that there is a transition to hybrid and multimodal designs versus density and datasets are biased towards QA. Although numerous in nature, they are not standardized between retrieval and generation and therefore introduce disparities between the quality of the retrievers and system reliability. A combination of these results draws attention to these three urgent needs: (i) expanding the concept of RAG to assessments of multimodal and domain-specific goals (ii) developing universal standards that can correlate retrieval and generation levels, and (iii) creating scalable hybrid pipelines and use of new standards in practice. These ideas are the direct stimulus to continue the discussion of difficulties and uncovered issues in Section 6.

6. Challenges and Open Issues

Irrespective of the fast progress, retrieval-augmented generation (RAG) and hybrid information retrieval (IR) systems still have challenges that restrain its use and implementation into practice. This section summarizes the unaddressed scalability, latency, hallucination, multimodality, evaluation practices, and ethical concerns issues.

6.1 Scalability and Efficiency

Storing billions of high-dimensional embeddings is dense retrieval, which issues the problem of storage and efficiency [86]. Several approximation nearest neighbor search systems, including FAISS [87] and ScaNN [88], are faster to identify results but remain unable to search billion-scale corporate collections. Distributed indexing and product quantization techniques [89] have potential, but are not yet deployed in the real world.

6.2 Latency and Real-Time Performance

RAG pipes cause latency as a result of multi-stage retrieval and reranking. In the case of interactive environments the delay of even a few hundred milliseconds is unacceptable [90]. This is aggravated by mid-retrieval strategies, which involve numerous calls in the process of decoding [91]. Although Caching can decrease the latency, there is no trade-off on speed and accuracy of retrieval [92].

6.3 Hallucination and Reliability

Evidence basing minimizes but does not eradicate hallucinations. The generators still might not pay attention to the retrieved passages or be irrelevant [93]. Retrieval errors are contagious and result in a factual inconsistency down the line [94]. Self-reflective system (ex: Self-RAG) tries to critique but it is not always reliable [95]. Formulation of loyal attribution systems is hence one of the challenges [96].

6.4 Multimodality and Heterogeneous Sources

RAG presents limitations of aligning text, images, tabular and code to the page. Making sure that representations in modalities are consistent is not easy [97]. Although the RAG is extended to vision-language models like MuRAG, domain-specific multimodal models (e.g. biomedical imaging + text) are not studied yet [98]. There is an acute requirement of benchmarks and multimodal RAG architectures [99].

6.5 Evaluation Fragmentation

Practices of evaluation are still fragmented. The quality of retrieval (e.g., Recall@k, MAP) is not necessarily directly related to the end-to-end factuality [100]. The metrics used - "overlap based generation" like BLEU and ROUGE are commonly employed but fail to ground on facts [101]. Semantic metrics (BERTScore, METEOR) capture meaning but are inconsistent across domains [102]. The FactScore and human evaluation used to measure faithfulness are novel and standardizable [103].

6.6 Ethical and Societal Concerns

Systems based on RAG also risk enhancing corpus-level biases [104], strengthening stereotypes [105], and breaching of privacy in case of sensitive resources indexing [106]. There are also apprehensions to responsibility, the validity of retrieved evidence and attribution [107]. Awareness of transparency and fairness in retrieval mechanisms must be introduced so as to instill responsible deployment [108].

7. Future Research Directions

Based on the issues presented in Section 6, the following section summarizes the opportunities of developing retrieval-augmented generation (RAG) and hybrid information retrieval (IR). To resolve these problems I have to invent new model architecture, assessment strategy, and sustainable deployment.

7.1 Scalable and Memory-Efficient Retrieval

The next generation RAG systems will have to cope with constantly growing corpora. Most endeavors to be promising encompass constricted dense representations, e.g., product quantization [109], and knowledge distillation to the retrievers [110]. Such methods as hierarchical indexes and learned sparse representations [111] may cut memory footprints by a huge factor without compromising accuracy. The other way is on-device retrieval of edge cases, which guarantees privacy at a cost of having centralized infrastructure.

7.2 Real-Time and Low-Latency RAG

Reducing response time is also important to interactive applications. It might be researched into asynchronous retrieval pipelines that simultaneously achieve retrieval and generation [112], progressive evidence retrieval (initial answers are gradually improved) [113], and dedicated hardware accelerators (e.g. GPUs/TPUs optimized to support similarity search) [114]. RAG can be used in time-s rampant application like health-care or finances where adaptive trade-offs can be used to balance RAG between the retrieval depth and the user response time.

7.3 Hallucination Mitigation and Reliability

Future systems ought to include constraints of faithfulness in decoding such that evidence based outputs are produced [115]. The other strategy is retrieval-conditioned alignment whereby retrieval scores explicitly affect attention weights, when generating [116]. Additional means of decreasing hallucination could be accomplished through self-verifying structures and multi-agent critical frameworks [117]. User trust will be necessary by setting the standards of attribution - equating each claim with evidence retrieved.

7.4 Multimodal and Domain-Specific RAG

RAG will become more and more non-textual. Joint multimodal encoders to align the text with image, audio and structured knowledge base [118] and cross-domain retrieval of specific fields such as biomedical imaging, legal archives and code repositories [119] are both research opportunities. This will be fastened by building significant multimodal benchmarks, which integrate QA, summarization, and reasoning. By combining symbolic reasoning and multimodal grounding, one might have more explainable systems.

7.5 Unified Evaluation Frameworks

Another avenue to consider in the future is the creation of holistic evaluation procedures that collectively assess the quality of retrieval, the fidelity of generation, and the consistency of information [120]. There is an urgent need to have automatic metrics which incorporate both semantic similarity and evidence attribution (such as FactScore+ variants). There should be community-wide benchmarks with varied tasks, the QA, summarization, dialogue, multimodal reasoning, on one protocol [121]. With such structures there would be comparative uniformity between studies and reproducible development.

7.6 Responsible and Ethical Deployment

Deployment of RAG systems is an irresponsible process, which means that such a process needs bias-aware retrieval models to reduce dangerous stereotypes [122]. Sensitive sources can be preserved using the federated retrieval or differential privacy embedding techniques [123]. Attribution should be also explainable as the future research should be able to ensure that retrieved evidence influences generated responses. It will be essential in collaborating with the IR, NLP, and ethics communities to build strong principles on transparency and fairness and accountability.

8. Conclusion

RAG and hybrid information retrieval (IR) has become a potent paradigm of grounding large language models with external knowledge. This survey has offered a single taxonomy that classifies RAG systems on five major dimensions, including the type of retriever, integration approach, retrieval granularity, application fields, and optimization approach. In this view, we reviewed original models, such as REALM and RAG, and FiD, Atlas, Re2G, Self-RAG, MuRAG, and many more to demonstrating the transformation to a more dense retrieval, hybrid and multimodal models.

To support the taxonomy, the survey made comparative studies with four dimension dimensions; existing surveys, representative models, benchmark data set and measurement of evaluation. These reviews demonstrated some obvious tendencies: the increase in the use of hybrid retrievers, the lack of domain-specific and multimodal datasets and the disjointed evaluation procedures contributing to lack of comparability between the studies. The synthesis showed not only the advantages of the

existing systems of RAG, including factual grounding and modular flexibility, but also the disadvantages such as latency, the risk of hallucinating, and poorly developed evaluation systems.

In future perspectives, the issues that have been discovered with regard to scalability, scalability, reliability, multimodality, assessment and ethics indicate promising future research directions. Research opportunities encompass devising memory efficient recall structures, building real time pipelines, incorporation of multimodal thought, standardization of holistic evaluation systems, and fairness, privatization and transparency into industrial applications. The following generation of RAG systems will not only transform academic research but will also drive reliable, scalable, and explainable AI systems in fields as different as healthcare, law, finance, and education because of its ability to bridge between retrieval and generation more conveniently.

Author Contributions: Sabah Mohammed coordinated the overall study design, supervised the survey structure, and guided the formulation of key themes across retrieval-augmented generation and hybrid information-retrieval methods. Pathan Hussain Bhasha conducted the primary literature review, organized the taxonomy of RAG architectures, and contributed to the comparative analysis and manuscript drafting. Osvaldo Gervasi supported the methodological framing, provided critical evaluation of state-of-the-art systems, and refined the positioning of the survey within the broader IR and LLM research landscape. Rajesh Bose assisted with data synthesis, contributed insights on performance evaluation trends, and helped revise and edit the manuscript. All authors participated in discussion, critical review, and final approval of the completed survey.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Ethical statement: This study is ethically sound and it does not endanger any life of people, animal, and environment.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “REALM: Retrieval-Augmented Language Model Pre-Training,” International Conference on Machine Learning (ICML), 2020.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS, 2020.

[3] Y. Huang and J. X. Huang, “The Survey of Retrieval-Augmented Text Generation in Large Language Models,” arXiv preprint arXiv:2404.10981, 2024.

[4] Perplexity AI, “Perplexity: Retrieval-Augmented AI Answer Engine,” [Online]. Available: https://www.perplexity.ai, Accessed: Sept. 2025.

[5] OpenAI, “ChatGPT with Browsing: Retrieval-Augmented Capabilities,” [Online]. Available: https://openai.com, Accessed: Sept. 2025.

[6] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” EMNLP, 2020.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS, 2020.

[9] G. Izacard and E. Grave, “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering,” arXiv preprint arXiv:2007.01282, 2021.

[10] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. R. Naik, P. Cai, and A. Gliozzo, “Re2G: Retrieve, Rerank, Generate,” arXiv preprint arXiv:2207.06300, 2022.

[11] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,” arXiv preprint arXiv:2310.11511, 2023.

[12] S. R. Gaddam, “An enhanced hybrid machine learning approach for efficient botnet attack detection in Internet of Things networks,” *Int. J. Commun. Netw. Inf. Secur.*, vol. 16, no. 1, pp. 449–458, Jan. 2024, doi: 10.48047/IJCNIS.16.1.458.

[13] G. Salton, A. Wong, and C. S. Yang, “A Vector Space Model for Automatic Indexing,” Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.

[14] S. E. Robertson and K. S. Jones, “Relevance Weighting of Search Terms,” Journal of the American Society for Information Science, vol. 27, no. 3, pp. 129–146, 1976.

[15] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[16] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” EMNLP, 2020.

[17] X. Luan, J. Bai, Y. He, W. Yih, J. Gao, M. Ostendorf, and M. Chen, “Sparse-to-Dense: Efficient Passage Retrieval via Dense Representation Learning,” arXiv preprint arXiv:2004.04906, 2020.

[18] O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” SIGIR, 2020.

[19] H. Hofstätter, S. Althammer, M. Zlabinger, A. Hanbury, and A. Tonellotto, “Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation,” arXiv preprint arXiv:2010.02666, 2020.

[20] F. Gao and T. Callan, “COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List,” NAACL-HLT, 2021.

[21] P. Izacard, F. Petroni, L. Hosseini, S. Riedel, and E. Grave, “Unsupervised Dense Information Retrieval with Contrastive Learning,” arXiv preprint arXiv:2112.09118, 2021.

[22] S. R. Gaddam, “Java-driven trustworthy and reliable deep learning for cyberattack detection in industrial IoT,” *International Journal of Communication Networks and Information Security*, vol. 14, no. 3, pp. 1274–1283, Apr. 2022, doi: 10.48047/IJCNIS.14.3.1283

[23] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications,” IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494–514, 2022.

[24] G. K. Chaitanya, S. R. Gaddam, K. S. F. Ahmad, B. Vicharapu, U. L. Soundharya, and U. N. L. Madhuri, “A multimodal approach to digital security: Combining steganography, watermarking, and image enhancement,” *IJBAS*, vol. 14, no. 2, pp. 611–619, Jul. 2025, doi: 10.14419/3r5r6r74.

[25] G. Izacard et al., “Few-Shot Learning with Retrieval-Augmented Language Models,” JMLR, 2023.

[26] J. Lin, X. Ma, S. Lin, J.-H. Yang, and J. Pradeep, “The Neural IR Revolution: A Retrospective,” arXiv preprint arXiv:2010.06056, 2020.

[27] S. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models,” NeurIPS, 2021.

[28] B. Kitchenham, “Procedures for Performing Systematic Reviews,” Keele University Technical Report TR/SE-0401, 2004.

[29] B. Kitchenham and S. Charters, “Guidelines for Performing Systematic Literature Reviews in Software Engineering,” EBSE Technical Report, 2007.

[30] IEEE, “IEEE Xplore Digital Library,” [Online]. Available: https://ieeexplore.ieee.org.

[31] ACM, “ACM Digital Library,” [Online]. Available: https://dl.acm.org.

[32] Springer, “SpringerLink: Journals, Books, and Conference Proceedings,” [Online]. Available: https://link.springer.com.

[33] Cornell University, “arXiv.org e-Print Archive,” [Online]. Available: https://arxiv.org.

[34] ACL, "ACL Anthology: A Digital Archive of NLP Research," [Online]. Available: https://aclanthology.org.
[35] Google, "Google Scholar," [Online]. Available: https://scholar.google.com.
[36] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.
[37] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," EMNLP, 2020.
[38] J. Ma, X. Yan, and J. Lin, "A Replicable and Efficient Hybrid Retrieval Framework," SIGIR, 2021.
[39] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[40] J. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," ICML, 2022.
[41] Y. Zhao, Q. Chen, and J. Zhou, "Post-Retrieval Fact Verification for Reliable QA," ACL, 2023.
[42] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models," arXiv preprint arXiv:2007.01282, 2021.
[43] S. Gao, Y. Han, and J. Callan, "Accurate Sentence-Level Retrieval for Factual Grounding," EMNLP, 2022.
[44] [W. Chen et al., "MuRAG: Multimodal Retrieval-Augmented Generator for Open QA over Images and Text," EMNLP, 2022.
[45] K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," ICML, 2020.
[46] T. Shi et al., "Neural Abstractive Summarization with Retrieval-Augmented Attention," NAACL-HLT, 2021.
[47] S. Shuster et al., "Retrieval-Enhanced Dialogue Systems," EMNLP, 2021.
[48] A. R. Pappas et al., "Biomedical Retrieval-Augmented Generative Models," Bioinformatics, 2022.
[49] D. Chalkidis et al., "Legal Judgment Prediction with RAG Models," ACL, 2022.
[50] H. Hofstätter et al., "Improving Efficient Neural Ranking Models with Knowledge Distillation," arXiv preprint arXiv:2010.02666, 2020.
[51] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," EMNLP, 2021.
[52] Y. Liu et al., "Fine-tuning Language Models with Reinforcement Learning from Human Feedback," NeurIPS, 2022.
[53] J. Wei et al., "Finetuned Language Models are Zero-Shot Learners," ICML, 2022.
[54] N. Izacard et al., "Unsupervised Dense Information Retrieval with Contrastive Learning," arXiv preprint arXiv:2112.09118, 2021.
[55] R. Menon et al., "Joint Optimization of Retrieval and Generation Objectives for RAG Models," ACL, 2023.
[56] S. Gao, J. Ma, and J. Lin, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2023.
[57] Y. Huang and J. X. Huang, "The Survey of Retrieval-Augmented Text Generation in Large Language Models," *arXiv preprint arXiv:2404.10981*, 2024.
[58] Z. Zhao, X. Fan, and Y. Zhang, "Knowledge-Oriented Retrieval-Augmented Generation: A Survey," *arXiv preprint arXiv:2407.08212*, 2024.
[59] X. Fan, Z. Zhao, and C. Xiong, "Hybrid Information Retrieval in the Era of LLMs: A Review," *arXiv preprint arXiv:2408.01423*, 2024.
[60] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," ICML, 2020.
[61] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[62] G. Izacard and E. Grave, "Fusion-in-Decoder: Leveraging Passage Retrieval for Open-Domain QA," arXiv preprint arXiv:2007.01282, 2021.
[63] G. Izacard et al., "Atlas: Few-Shot Learning with Retrieval-Augmented Language Models," JMLR, 2023.
[64] M. Glass et al., "Re2G: Retrieve, Rerank, Generate," arXiv preprint arXiv:2207.06300, 2022.
[65] A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," arXiv preprint arXiv:2310.11511, 2023.
[66] W. Chen et al., "MuRAG: Multimodal Retrieval-Augmented Generator for Open QA over Images and Text," EMNLP, 2022.
[67] Kwiatkowski et al., "Natural Questions: A Benchmark for Question Answering Research," *TACL*, 2019.
[68] A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *ACL*, 2017
[69] Nguyen et al., "MS MARCO: A Human Generated MAchine Reading Comprehension Dataset," *NeurIPS*, 2016.
[70] Thakur et al., "BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models," *NeurIPS*, 2021.
[71] Yang et al., "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," *EMNLP*, 2018.
[72] Jin et al., "PubMedQA: A Dataset for Biomedical Research Question Answering," *EMNLP*, 2019.
[73] Chalkidis et al., "Legal Judgment Prediction: A Benchmark Dataset and Empirical Study," *COLIEE*, 2020.
[74] Artetxe et al., "XQA: A Cross-Lingual Open-Domain Question Answering Dataset," *ACL*, 2020.
[75] Liu et al., "WikiSum: Coherent Abstractive Summarization with Wikipedia," *ACL*, 2018.
[76] Fan et al., "ELI5: Long Form Question Answering," *ACL*, 2019.
[77] Buckley & Voorhees, "Evaluating Evaluation Measure Stability for Ranking Tasks," *SIGIR*, 2000.
[78] Craswell, "Mean Reciprocal Rank, Recall@k and Evaluation in IR," *SIGIR Tutorial*, 2009.
[79] Järvelin & Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM TOIS*, 2002.
[80] Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," *ACL*, 2002.
[81] Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *ACL Workshop*, 2004.
[82] Banerjee & Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *ACL Workshop*, 2005.
[83] Zhang et al., "BERTScore: Evaluating Text Generation with BERT," *ICLR*, 2020.
[84] Rashkin et al., "Measuring Factual Consistency in Abstractive Summarization," *EMNLP*, 2021.
[85] van der Lee et al., "Human Evaluation of Automatically Generated Text: Current Trends and Future Directions," *ACL*, 2019.
[86] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[87] J. Johnson et al., "FAISS: A Library for Efficient Similarity Search," Facebook AI Research, 2017.
[88] G. Guo et al., "ScaNN: Scalable Nearest Neighbor Search for Large Datasets," ICML, 2020.
[89] A. Babenko and V. Lempitsky, "Product Quantization for Nearest Neighbor Search," IEEE TPAMI, vol. 33, no. 1, pp. 117–128, 2011.
[90] J. Leskovec, A. Rajaraman, and J. Ullman, Mining of Massive Datasets, 3rd ed. Cambridge University Press, 2020.
[91] J. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," ICML, 2022.
[92] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," ACL, 2017.
[93] E. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," ACL, 2020.
[94] S. Ji et al., "Survey on Knowledge Graphs: Representation, Acquisition, and Applications," IEEE TNNLS, vol. 33, no. 2, pp. 494–514, 2022.
[95] A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," arXiv preprint arXiv:2310.11511, 2023.
[96] R. Menon et al., "Joint Optimization of Retrieval and Generation Objectives for RAG Models," ACL, 2023.
[97] Y. Li et al., "Multimodal Representation Learning: A Survey," IEEE TPAMI, vol. 43, no. 6, pp. 2008–2038, 2021.
[98] W. Chen et al., "MuRAG: Multimodal Retrieval-Augmented Generator for Open QA over Images and Text," EMNLP, 2022.
[99] Z. Li, C. Chen, and L. Carin, "Multimodal Retrieval-Augmented Models: Challenges and Opportunities," arXiv preprint arXiv:2405.00345, 2024.
[100] C. Buckley and E. M. Voorhees, "Evaluating Evaluation Measure Stability for Ranking Tasks," SIGIR, 2000.
[101] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," ACL Workshop, 2004.
[102] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," EMNLP, 2021.
[103] Y. Rashkin et al., "Measuring Factual Consistency in Abstractive Summarization," EMNLP, 2021.
[104] M. Sap et al., "Social Bias Frames: Reasoning about Social and Power Implications of Language," ACL, 2020.
[105] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" FAccT, 2021.

[106] R. Shokri et al., “Membership Inference Attacks Against Machine Learning Models,” IEEE S&P, 2017.
[107] F. Ribeiro, S. K. Shah, and K. Kirchhoff, “Factually Consistent Summarization with Attribution,” ACL, 2022.
[108] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning, MIT Press, 2019.
[109] H. Jegou, M. Douze, and C. Schmid, “Product Quantization for Nearest Neighbor Search,” IEEE TPAMI, vol. 33, no. 1, pp. 117–128, 2011.
[110] H. Hofstätter et al., “Efficient Student-Teacher Training for Dense Retrieval,” SIGIR, 2021.
[111] J. Lin, X. Ma, and J. Lin, “SPLADE: Sparse Lexical and Expansion Models for First-Stage Retrieval,” SIGIR, 2021.
[112] C. Wang et al., “Asynchronous Pipeline Parallelism for Neural Sequence Models,” ICML, 2022.
[113] T. Shi et al., “Progressive Retrieval for Efficient Question Answering,” ACL, 2023.
[114] J. Guo et al., “Accelerating Large-Scale Nearest Neighbor Search with GPUs,” VLDB, 2022.
[115] E. Rashkin et al., “Faithful Summarization with Evidence Attribution,” ACL, 2021.
[116] Y. Liu et al., “Retriever-Conditioned Generation for Knowledge-Intensive NLP,” NeurIPS, 2022.
[117] H. Shuster et al., “Language Model Self-Critique via Multi-Agent Debate,” EMNLP, 2023.
[118] Y. Li et al., “Multimodal Representation Learning: Foundations, Trends, and Applications,” IEEE TPAMI, vol. 43, no. 6, pp. 2008–2038, 2021.
[119] A. R. Pappas et al., “Domain-Specific Retrieval-Augmented Biomedical QA,” Bioinformatics, 2022.
[120] R. Menon et al., “Joint Evaluation of Retrieval and Generation in Knowledge-Augmented Models,” ACL, 2023.
[121] S. Thakur et al., “BEIR: Towards a Standardized Benchmark for Zero-Shot Evaluation,” NeurIPS, 2021.
[122] M. Sap et al., “Mitigating Social Bias in Neural Language Models,” ACL, 2020.
[123] R. Shokri et al., “Differentially Private Retrieval Models for Sensitive Data,” IEEE S&P, 2022.