SPARSE RETRIEVAL, DENSE RETRIEVAL, AND RETRIEVAL-AUGMENTED GENERATION: A LITERATURE REVIEW

A PREPRINT

Alireza Delavari
M.Sc. in AI & Robotics, Amirkabir University of Technology
alirezadelavari@aut.ac.ir

September 7, 2025

ABSTRACT

Information retrieval is a central component of knowledge-intensive natural language processing. This report reviews three seminal approaches that represent milestones in the evolution of retrieval methods: *BM25*, a probabilistic sparse retriever; *Dense Passage Retrieval (DPR)*, a neural embedding-based retriever; and *Retrieval-Augmented Generation (RAG)*, which integrates retrieval with generative models. For each method, we examine the authors' objectives, the limitations of prior approaches, their novel contributions, and how they evaluated success. We then provide a comparative analysis that highlights the strengths, weaknesses, and broader impact of these approaches. The synthesis discusses their applications, risks, and integration, showing how sparse and dense retrieval complement each other within the RAG framework. Taken together, these methods illustrate the trajectory from lexical matching to semantic retrieval to retrieval-augmented generation, which now forms the backbone of modern large language models and generalist AI agents.

Keywords Information Retrieval · BM25 · Dense Passage Retrieval · Retrieval-Augmented Generation

1 Introduction

Information retrieval (IR) has long been a cornerstone of natural language processing and artificial intelligence. At its core, IR addresses the problem of efficiently finding relevant information from large document collections, a challenge that becomes especially crucial in the era of large-scale knowledge bases and language models. In recent years, the interplay between retrieval systems and generative models has gained significant importance for building more factual, interpretable, and scalable AI agents.

This report focuses on three seminal approaches that represent key milestones in the evolution of retrieval methods: *Sparse Retrieval* through BM25 [Robertson and Zaragoza, 2009], *Dense Retrieval* through Dense Passage Retrieval (DPR) [Karpukhin et al., 2020], and *Retrieval-Augmented Generation* (RAG) [Lewis et al., 2020]. Together, these methods chart the trajectory from classical probabilistic term-matching, to neural semantic representation learning, and finally to hybrid architectures that integrate retrieval directly into large-scale generative models.

Sparse retrieval, exemplified by BM25, relies on probabilistic term weighting to rank documents by lexical overlap. While computationally efficient and interpretable, such models struggle with semantic matching. Dense retrieval, introduced by DPR, leverages neural embeddings to capture semantic similarity beyond exact token overlap, enabling better generalization to paraphrases and rephrasings. Finally, RAG integrates retrieval into end-to-end generative modeling, allowing language models to ground their outputs in retrieved evidence while remaining adaptable across tasks.

The rest of this report is structured as follows. Section 2 reviews the origins and contributions of BM25 in sparse retrieval. Section 3 discusses DPR as a representative dense retrieval method, highlighting its neural architecture and

empirical impact. Section 4 introduces Retrieval-Augmented Generation, which unifies dense retrieval with generative transformers. In Section 5, we provide a comparative analysis of the three approaches. Section 6 synthesizes insights, discussing their impact, limitations, and integration in modern AI pipelines. We conclude in Section 7 with final reflections on the broader implications of retrieval for knowledge-intensive NLP.

2 Sparse Retrieval: BM25

2.1 Objective

BM25 [Robertson and Zaragoza, 2009] is the de-facto sparse retrieval baseline, grounded in the probabilistic relevance framework introduced by Robertson and Spärck Jones [1976]. BM25 operationalizes a probabilistic relevance framework for ranking documents, introducing term-frequency saturation and document-length normalization in a practical scoring function [Robertson and Zaragoza, 2009].

2.2 Prior Approaches and Limitations

Before BM25, the dominant paradigm for retrieval was based on *vector space models* and simple term-weighting heuristics, most notably Term Frequency–Inverse Document Frequency (TF–IDF). While TF–IDF successfully prioritized rare and frequent terms differently, it suffered from important shortcomings:

- **Linear term growth:** TF–IDF treated term frequency as linearly proportional to importance, often overvaluing repeated words.
- **Lack of length normalization:** Longer documents tended to rank higher regardless of relevance, simply because they contained more words.
- **Absence of probabilistic grounding:** The weights had no clear probabilistic interpretation, making it difficult to reason about their theoretical soundness.

$$\text{score}(q,d) = \sum_{t \in q} \underbrace{\log \frac{N - n_t + 0.5}{n_t + 0.5}}_{\text{IDF}(t)} \cdot \frac{(k_1+1)\, f(t,d)}{f(t,d) + k_1 \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \tag{1}$$

Figure 1: BM25 scoring function with TF saturation (k_1) and length normalization (b) [Robertson and Zaragoza, 2009]. $f(t,d)$ is term frequency, $|d|$ document length, avgdl corpus average length, n_t document frequency, N corpus size.

Historical grounding (RSJ 1976). Earlier probabilistic relevance work by Robertson and Spärck Jones [1976] formalized *relevance weighting of search terms*: term statistics in relevant vs. non-relevant sets induce discriminative weights. Modern BM25 can be viewed as a practical instantiation within this probabilistic relevance framework, adding term-frequency saturation and document-length normalization to address TFIDFs linearity and length bias. This connects the standard BM25 scoring function used today to the original 1976 relevance-weighting formulation.

2.3 Novel Contributions

The BM25 formulation introduced several key innovations that addressed these issues:

- **Probabilistic Relevance Model:** BM25 grounded term weighting in probability theory, framing retrieval as estimating the odds of a document being relevant to a given query.
- **Term Frequency Saturation:** Instead of growing linearly, the contribution of a term increases with frequency but plateaus, reflecting diminishing returns of repeated occurrences.
- **Document Length Normalization:** BM25 penalizes overly long documents by normalizing term contributions relative to document length, reducing length bias.
- **Parameterization:** The introduction of tunable parameters k_1 (term frequency scaling) and b (length normalization) provided flexibility to adapt retrieval to different datasets.

These innovations made BM25 robust, interpretable, and computationally efficient, helping it remain a de facto standard in information retrieval.

2.4 Evaluation and Results

Empirical evidence summarized in Robertson and Zaragoza [2009] highlights BM25's strong accuracy–efficiency trade-offs and its sustained role as a widely used baseline. In practice, BM25 often complements dense retrievers within hybrid systems.

3 Dense Retrieval: DPR

3.1 Objective

Karpukhin et al. [Karpukhin et al., 2020] introduced Dense Passage Retrieval (DPR) to address the semantic limitations of sparse retrieval models such as BM25. The objective of DPR was to build a retrieval system that could capture semantic similarity beyond lexical overlap, enabling effective retrieval even when queries and relevant passages use different words or phrasings. The authors specifically targeted the problem of open-domain question answering, where retrieving semantically relevant passages is critical for downstream readers and generative models.

3.2 Prior Approaches and Limitations

Prior to DPR, open-domain systems typically paired a sparse retriever (e.g., BM25) with a neural reader: effective but limited by lexical mismatch at retrieval time. Early neural retrievers explored dense embeddings and retrieval-augmented pretraining but were comparatively heavy and complex.

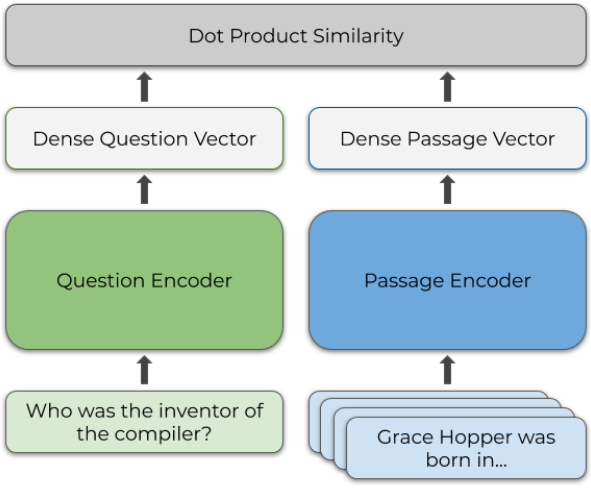

Figure 2: Dual-encoder architecture of Dense Passage Retrieval (DPR). A question encoder and a passage encoder independently produce dense vectors, which are compared using dot product similarity for retrieval.

3.3 Novel Contributions

DPR introduced a streamlined and effective dense retrieval approach:

- **Bi-encoder architecture:** Separate BERT-based encoders for questions and passages produce low-dimensional embeddings. Relevance is scored by the inner product of these embeddings.
- **Contrastive training:** The model is trained on question–positive passage pairs, contrasted against negative passages (random, BM25-selected, or in-batch negatives), optimizing embeddings for semantic separation.
- **Efficient retrieval with FAISS:** DPR leverages FAISS for Maximum Inner Product Search (MIPS), enabling scalable retrieval over millions of passages in sub-linear time.
- **Practicality:** DPR relies on supervised QA datasets (e.g., Natural Questions, TriviaQA) rather than complex specialized pretraining [Karpukhin et al., 2020].

These contributions demonstrated that dense retrieval could be both conceptually simple and empirically powerful, making it a practical alternative to sparse methods.

3.4 Evaluation and Results

DPR was evaluated on several open-domain QA benchmarks, including Natural Questions (NQ), TriviaQA, WebQuestions, and CuratedTREC. The evaluation followed a two-stage pipeline: (1) DPR retrieved top-k passages, and (2) a neural reader extracted answers from the retrieved contexts.

Key findings include:

- DPR outperformed BM25 by large margins in top-20 and top-100 retrieval accuracy (e.g., 78.4% vs. 59.1% top-20 accuracy on NQ).
- Higher retrieval accuracy translated directly into better end-to-end QA performance, achieving state-of-the-art results on multiple datasets at the time.
- DPR was sample-efficient: even with as few as 1,000 training examples, it already surpassed BM25 in retrieval accuracy.

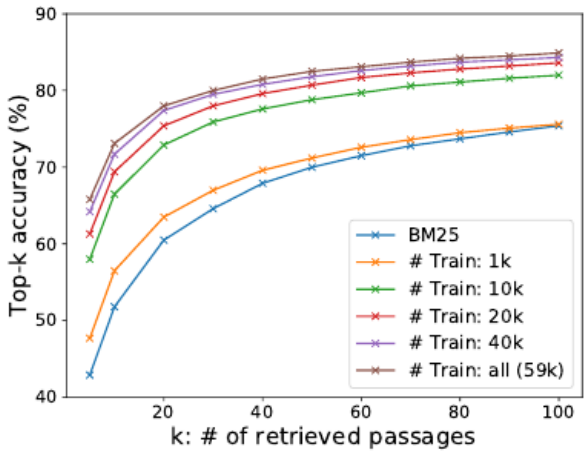

Figure 3: Comparative performance of DPR vs. BM25 across QA benchmarks.

The strong empirical performance of DPR established dense retrieval as a scalable, semantically informed alternative to sparse methods. It also laid the groundwork for hybrid approaches and retrieval-augmented generation, where dense retrievers provide contextual grounding for generative language models.

4 Retrieval-Augmented Generation (RAG)

4.1 Objective

Lewis et al. [Lewis et al., 2020] introduced Retrieval-Augmented Generation (RAG) to unify dense retrieval with generative language models in a single end-to-end framework. The core objective was to improve performance on knowledge-intensive tasks—such as open-domain question answering, fact verification, and slot filling—by allowing a generative model to condition its outputs not only on parametric memory (stored in model weights) but also on non-parametric memory (retrieved passages from a large corpus). This hybrid paradigm was designed to enhance factual accuracy, interpretability, and adaptability across domains.

4.2 Prior Approaches and Limitations

Before RAG, two separate paradigms were common:

- **Standalone language models:** Models such as GPT-2 relied solely on internalized parameters for knowledge. While powerful, they often produced outdated or hallucinated facts because their memory was fixed at training time.

- **Retrieve-then-read pipelines:** Earlier QA systems combined a retriever (e.g., BM25, DPR) with a reader in a multi-stage pipeline. While effective, this approach lacked tight integration: the retriever was not optimized jointly with the generator, leading to inefficiencies and missed opportunities for end-to-end learning.

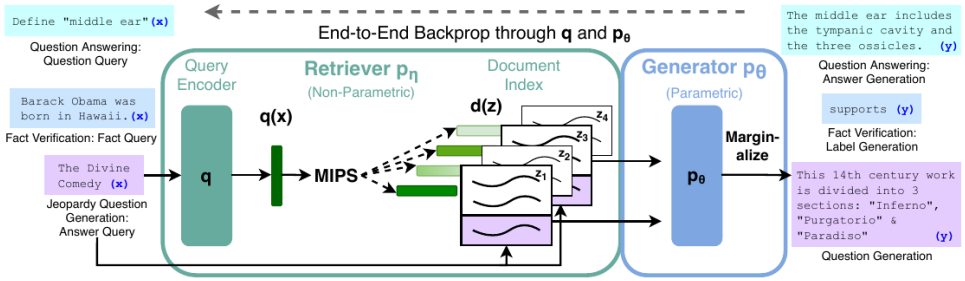

Figure 4: RAG framework combining a retriever with a generative model.

4.3 Novel Contributions

RAG introduced several innovations that made retrieval-augmented generation practical:

- **Retriever–Generator Integration:** RAG couples a retriever (typically DPR) with a generator (BART) in an end-to-end architecture. For each query, the retriever returns top-k passages, which the generator then conditions on during decoding.
- **Probabilistic Fusion:** RAG formalizes generation as a marginal likelihood over retrieved documents, effectively averaging the probability of generating an answer across evidence sources.
- **End-to-End Training:** The query encoder and generator are trained jointly while the document index is kept fixed (the document encoder is frozen when building the index) for efficiency [Lewis et al., 2020].
- **Flexible Modes:** Two variants were proposed: *RAG-Sequence* (conditioning the entire sequence on one passage) and *RAG-Token* (allowing token-level marginalization across passages).

4.4 Evaluation and Results

RAG was evaluated on multiple knowledge-intensive NLP benchmarks, including Natural Questions, WebQuestions, TriviaQA, and CuratedTREC. The evaluation compared RAG against standalone generators (BART, GPT-2) and traditional retrieve-then-read pipelines.

Key results:

- RAG consistently outperformed both parametric-only generators and pipeline-based baselines, achieving new state-of-the-art results at the time.
- RAG-Token, in particular, achieved higher accuracy than RAG-Sequence, benefiting from fine-grained conditioning across retrieved documents.
- Retrieval grounding reduced hallucinations and improved factual correctness, although failure cases still occurred when retrieval returned misleading or irrelevant passages.

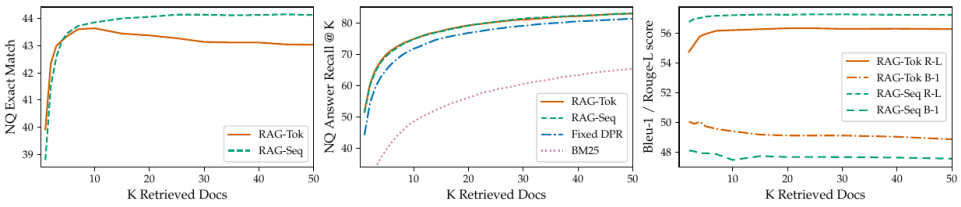

Figure 5: Performance of RAG compared to baseline models on QA benchmarks.

The success of RAG demonstrated the power of retrieval-augmented language models. By combining sparse or dense retrievers with generative architectures, RAG created a new paradigm that has since been adopted and extended in many large-scale systems, including modern LLMs with retrieval plugins. It also set the stage for hybrid retrieval approaches and motivated research into more reliable retrievers for grounding generative models.

5 Comparative Analysis

The three approaches discussed—BM25, DPR, and RAG—represent different stages in the evolution of information retrieval and its integration into modern language models. BM25 exemplifies the power and efficiency of sparse lexical retrieval, DPR demonstrates the scalability and semantic depth of dense embeddings, and RAG integrates retrieval with generation to enable knowledge-intensive NLP applications. Table 1 summarizes their core differences.

Table 1: Comparison of BM25, DPR, and RAG across key dimensions.

Aspect	**BM25 (Sparse)**	**DPR (Dense)**	**RAG (Retrieval + Generation)**
Core Idea	Probabilistic term weighting	Bi-encoder embeddings with contrastive training	End-to-end integration of retriever and generator
Representation	High-dimensional sparse vectors	Low-dimensional dense vectors	Hybrid: retrieved evidence + sequence model
Strengths	Efficient, interpretable, strong on rare terms	Semantic matching, robust to paraphrases, scalable with FAISS	Produces grounded outputs, improves factual accuracy, adaptable across tasks
Weaknesses	Ignores semantics, struggles with synonyms/paraphrases	Requires supervised data, resource-intensive, domain shift risk	Relies on retriever quality, can still hallucinate if evidence is weak
Evaluation Benchmarks	Classic IR and QA datasets	Natural Questions, TriviaQA, WebQuestions, CuratedTREC	Same QA datasets plus knowledge-intensive tasks
Impact	Standard baseline in IR for decades	Established dense retrieval as practical and effective	Blueprint for retrieval-augmented LLMs and modern agent pipelines

In summary, BM25 provides a lightweight and interpretable baseline that remains valuable in domains with technical vocabularies or limited training data. DPR offers a powerful alternative by learning semantic embeddings that enable retrieval based on meaning rather than surface form. Finally, RAG extends these ideas by coupling retrieval directly to a generative model, allowing end-to-end systems to answer knowledge-intensive queries with higher factual grounding.

These methods are not mutually exclusive: hybrid retrievers often combine BM25 and DPR for complementary strengths, and RAG can flexibly incorporate either sparse or dense retrieval modules. This interplay highlights that the field has moved from isolated improvements in retrieval toward fully integrated frameworks for knowledge-enhanced generation.

6 Synthesis and Discussion

6.1 Impact and Applications

The progression from BM25 to DPR and ultimately to RAG has significantly shaped the way modern AI systems access and use knowledge. BM25 remains widely deployed in search engines, digital libraries, and as a baseline in academic benchmarks due to its efficiency and interpretability [Robertson and Zaragoza, 2009]. DPR demonstrated that dense retrieval could scale to millions of passages, making semantic search practical and powering open-domain question answering systems [Karpukhin et al., 2020]. RAG pushed this evolution further by integrating retrieval into the generative process, laying the foundation for retrieval-augmented large language models [Lewis et al., 2020].

These technologies have had impact far beyond academic benchmarks. Practical applications include:

- **Search engines and enterprise search:** Improved relevance ranking and semantic matching of queries to documents.
- **Conversational agents:** RAG-style architectures enable chatbots to provide factual, grounded answers rather than relying solely on parametric memory.
- **Knowledge-intensive tasks:** Fact verification, slot filling, and domain-specific QA (e.g., biomedical and legal).
- **Hybrid retrieval systems:** Production systems often combine sparse and dense retrievers to balance efficiency, robustness, and semantic coverage.

6.2 Risks and Failure Modes

Despite their successes, these approaches introduce new risks and limitations:

- **Sparse retrieval risks (BM25):** Failure to capture semantics, vulnerability to lexical mismatch, and poor handling of paraphrases.
- **Dense retrieval risks (DPR):** Dependence on large labeled datasets, computational cost for training and indexing, and potential brittleness when applied to out-of-domain data.
- **RAG risks:** Retrieval errors can propagate into generation, amplifying hallucinations or introducing misleading information. Additionally, integrating retrievers with generators raises challenges in interpretability, latency, and resource efficiency.

Ethical concerns also arise: retrieval models may propagate biases present in training corpora, and retrieval-augmented generation may give users misplaced confidence in factually incorrect outputs. Mitigation strategies, such as retrieval diversification, fact verification modules, and hybrid retrievers, are active areas of research.

6.3 Integration in RAG Framework

The three technologies fit together naturally within the RAG paradigm. Sparse retrieval provides a fast, interpretable backbone that excels with rare terms and domain-specific vocabulary. Dense retrieval complements this by capturing semantic similarity and enabling paraphrase robustness. RAG then acts as a unifying architecture that leverages either or both retrievers to supply evidence for generation.

In practice, hybrid retrievers—which combine BM25 scores with dense embeddings—often outperform either method alone. This synergy reflects the complementary nature of lexical and semantic matching. When coupled with a generative model, such as BART or T5, the result is an AI system that is both factually grounded and semantically flexible.

Overall, the integration of sparse, dense, and retrieval-augmented generation illustrates the trajectory of modern IR: from lexical matching, to semantic embedding, to full integration with generative models. This trajectory has set the stage for current and future research on retrieval-augmented large language models.

7 Conclusion

This report has traced the evolution of retrieval methods from probabilistic sparse models to dense neural retrieval and, finally, to retrieval-augmented generation. BM25 established a principled and efficient baseline for sparse retrieval, introducing probabilistic relevance weighting and remaining influential for decades as both a practical tool and a benchmark [Robertson and Zaragoza, 2009]. Dense Passage Retrieval (DPR) marked a paradigm shift by demonstrating that neural embeddings, trained with contrastive objectives, could outperform sparse methods in open-domain question answering, enabling retrieval based on semantic similarity rather than lexical overlap [Karpukhin et al., 2020]. Building upon both, Retrieval-Augmented Generation (RAG) integrated retrieval into generative architectures, allowing models to combine parametric memory with non-parametric evidence for improved factual grounding and scalability [Lewis et al., 2020].

Together, these approaches illustrate a trajectory from *isolated retrieval improvements* to *integrated retrieval-generation frameworks*. They highlight how progress in information retrieval has become deeply interwoven with the development of large language models and AI agents. The comparative analysis underscores that BM25, DPR, and RAG are not replacements but complements: sparse retrieval remains valuable for rare terms and efficiency, dense

retrieval excels at semantic generalization, and RAG provides the scaffolding to unite them for knowledge-intensive tasks.

Looking forward, hybrid retrievers, adaptive retrieval mechanisms, and tighter integration with large-scale generative models are promising avenues for further research. The risks identified—including domain shift, computational cost, and hallucination propagation—remain open challenges. Nevertheless, the trajectory from BM25 to RAG has already transformed how modern AI systems interact with knowledge, laying the groundwork for generalist AI agents that can combine reasoning, memory, and language generation in a unified framework.

References

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi:10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.550. URL https://arxiv.org/abs/2004.04906.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kulshreshtha, Veselin Stoyanov, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Stephen E. Robertson and Karen Spärck Jones. Relevance weighting of search terms. *Journal of Documentation*, 32 (1):2–11, 1976. doi:10.1108/eb026594.