# K. N. Toosi University of Technology

**Aidin Sahneh**

**Topic:**

Foundations of Retrieval-Augmented Generation

# Contents

# 1 Analysis of Core Papers

This section analyzes the three foundational papers on Sparse Retrieval, Dense Retrieval, and Retrieval-Augmented Generation (RAG).

## 1.1 Sparse Retrieval: Relevance Weighting [1]

- **Objective:** The author aimed to establish a theoretical framework for weighting search terms. The goal was to move beyond simple term counting and utilize "relevance information" to mathematically determine which words are most useful for distinguishing relevant documents from non-relevant ones.
- **Prior Art & Limits:** Previously, systems relied on rigid Boolean logic (unranked results) or simple statistical frequency methods (like early IDF). The limitation was that these methods were "request-independent"; they assigned the same weight to a term regardless of the specific query context or user feedback.
- **Innovation:** The innovation was the introduction of a **Probabilistic Model** based on a contingency table. This table analyzed the presence/absence of terms in relevant vs. non-relevant documents. It led to the derivation of weighting formula **F4**, which considers both the presence and absence of terms. This formula became the theoretical foundation for the famous BM25 algorithm.
- **Evaluation:** The method was evaluated using standard test collections (like Cranfield 1400). The "exam" was comparing Recall-Precision graphs against baselines (unweighted terms and simple IDF). The results proved that the new relevance-based weighting significantly outperformed previous methods.

## 1.2 Dense Retrieval: DPR [2]

- **Objective:** The authors wanted to prove that retrieval could be done effectively using only **Dense Representations** (embeddings), without relying on traditional keyword matching (BM25). They aimed to show this could be achieved with a simple architecture and modest training data.
- **Prior Art & Limits:** The standard was BM25, which suffers from the "lexical gap" (it cannot match synonyms like "villain" and "bad guy"). Previous dense attempts (like ORQA) existed but required complex and computationally expensive pre-training tasks (ICT) to work well.
- **Innovation:** The core innovation was the training strategy using a simple **Dual-Encoder** architecture. They used **"In-batch negatives"** (using other questions' answers in the same batch as negative examples) to train efficiently. They also incorporated "Hard Negatives" (wrong passages that looked right to BM25) to sharpen the model's discrimination.
- **Evaluation:** The method was evaluated on Open-Domain QA datasets (Natural Questions, TriviaQA, etc.). The primary metric was **Top-k Retrieval Accuracy**. DPR achieved 78.4% Top-20 accuracy on Natural Questions, significantly beating BM25 (59.1%). It also improved end-to-end QA performance when paired with a reader.

## 1.3 Retrieval-Augmented Generation: RAG [3]

- **Objective:** The goal was to fix major flaws in large language models (LLMs). LLMs store all their knowledge in their weights, which is called **"parametric memory."** This memory is hard to update, can be factually wrong ("hallucination"), and is not interpretable. The

objective was to combine this with a **"non-parametric memory"** (an external document index) that could be accessed on the fly.

- **Prior Art & Limits:** Previous approaches were either "Extractive" (could find answers but not write summaries) or purely "Generative" (like GPT/BART, which hallucinated facts and couldn't access new data).
- **Innovation:** RAG introduced a hybrid architecture where a Neural Retriever (DPR) and a Seq2Seq Generator (BART) are **jointly fine-tuned**. The model learns to retrieve documents that help generate the correct output. They introduced two variants: **RAG-Sequence** (one document for the whole answer) and **RAG-Token** (different documents for each word).
- **Evaluation:** RAG was tested on "Knowledge-Intensive Tasks" including Open-Domain QA, Fact Verification, and Question Generation. It set new state-of-the-art results on Open-Domain QA. Crucially, human evaluators found RAG's generation to be more **factual** and **specific** than the baseline BART model.

# 2 Synthesis and Discussion

## 2.1 Broader Impact ("Who Cares?")

The transition from keyword-based search to semantic dense retrieval, and finally to Retrieval-Augmented Generation, represents a paradigm shift in AI.

- **Accuracy:** It allows AI to answer questions based on verifiable external data rather than just memorized patterns, reducing "hallucinations."
- **Updatability:** As shown in the RAG paper, we can update the AI's "knowledge" simply by swapping the document index (e.g., updating Wikipedia dump) without retraining the massive neural network. This makes AI systems much more practical and maintainable for real-world applications.

## 2.2 Risks and Failure Modes

- **Retrieval Errors:** If the retriever fails to find the relevant document (or finds a misleading one), the generator will produce a wrong answer ("Garbage In, Garbage Out").
- **Bias Propagation:** The model is only as good as its index. If the source documents (e.g., Wikipedia) contain bias or misinformation, the RAG system will propagate it.
- **Complexity:** Dense retrieval requires heavy computation (GPUs) for indexing and inference compared to the lightweight CPU-based BM25.

## 2.3 Synthesis: How They Fit Together

These three technologies form the evolution of modern Information Retrieval:

- **Sparse Retrieval (BM25)** serves as the robust baseline. It is fast, requires no training, and is excellent for exact keyword matching (e.g., searching for a specific name or ID).
- **Dense Retrieval (DPR)** solves the semantic understanding problem. It acts as the "eyes" of the system, finding relevant context that doesn't necessarily share words with the query.
- **RAG** is the framework that connects these retrieval mechanisms to a "brain" (Generator).

In a practical RAG system, **Hybrid Retrieval** is often the best approach: using BM25 for precision (exact matches) and DPR for recall (semantic understanding), giving the Generator the best possible context to answer the user.

## References


[1] Jones, K. S. (1976). *Relevance Weighting of Search Terms.* Journal of Documentation.

[2] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). *Dense Passage Retrieval for Open-Domain Question Answering.* Proceedings of EMNLP.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* NeurIPS.