

Comparative Evaluation of Retrieval Methods for Arabic Retrieval-Augmented Generation

Prepared By

Name: Hadeel Alsakkaf

ID No.: 261002581

Name: Shahd Dairi

ID No.: 261002707

Submitted in fulfilment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

KM

Supervisor Name: Dr. Yasser Abdel-Hamid

Collage of Informatics,

Midocean University, Comoros

Acknowledgements

Thanks to Dr. Yasser Abdel-Hamid for his guidance, support and constructive feedback. His expertise, encouragement, and thoughtful recommendations have greatly contributed to the development of this work.

Abstract

Retrieval-Augmented Generation (RAG) has become an effective approach in Natural Language Processing (NLP) for improving the accuracy and reliability of Large Language Models (LLMs) by incorporating external knowledge during response generation. Retrieval methods are a fundamental key factor in the act of RAG systems; the quality of retrieved documents heavily influences the generated response. Popular retrieval strategies include sparse retrieval (keyword-based), dense retrieval (semantic-based), and hybrid retrieval, which combines both approaches; eventually reranking is applied to further improve the retrieved answer. While these retrieval methods have been studied, comparative evaluations for Arabic RAG systems are still limited. This is primarily due to the unique characteristics of the Arabic language, being its rich morphology, lexical variations, and optional diacritics, which can affect retrieval performance. Thus, this study aims to evaluate a comparison of sparse, dense, and hybrid retrieval and reranking strategies. This study is conducted by using academic documents to identify the most effective approach for Arabic information retrieval. The expected results are to provide a rational guidance for developing more accurate and reliable Arabic RAG systems for domain-specific applications.

Keywords: Retrieval-Augmented Generation (RAG), Arabic Retrieval, Sparse Retrieval, Dense Retrieval, Hybrid Retrieval, Ranking Retrieval.

List of Publications

No publication list is to be reported at this time, since research has produced no publications yet.

Table of Contents

Acknowledgements	1
Abstract	2
List of Publications	4
List of Abbreviations	5
List of Tables	6
Chapter 1: Introduction	6
1.1 Overview	7
1.2 Problem Definition.....	7
1.3 Research Gap & Motivation	8
1.4 Thesis Objectives	8
1.5 Research Questions	9
1.6 Thesis Contribution.....	9
1.7 Scope/Delimitation	10
1.8 Resources Needed	11
1.9 Thesis Organization	12
Chapter 2: Background and Related Work	13
2.1 Background	13
2.1.1 Retrieval-Augmented Generation (RAG)	13
2.1.2 Retrieval Methods	14
2.1.3 Evaluation Metrics	15
2.1.4 Arabic Language Challenges.....	17
2.2 Related Work	18
2.2.1 Arabic RAG Studies	18
2.2.2 RAG Application in Educational Contexts	20
2.2.3 Retrieval Methods Studies	20
2.2.4 Ranking Models	22
Table 2.3 Comparative Analysis of Previous Studies on RAG	23
Reference	26

List of Abbreviations

Abbreviation	Expansion
RAG	Retrieval-Augmented Generation.
LLM	Large Language Model.
IR	Information Retrieval.
NLP	Natural Language Processing.
MRR	Mean Reciprocal Rank.
NDCG	Normalized Discounted Cumulative Gain.
TF-IDF	Term Frequency–Inverse Document Frequency.
CRAG	Corrective Retrieval-Augmented Generation.
QA	Question Answering.
RoC	Ratio of Complementarity.
ARCD	Arabic Reading Comprehension Dataset.
MS MARCO	Microsoft Machine Reading Comprehension.
HHR.	Hybrid Hierarchical Retrieval.
MiniLM	Mini Language Model.
BGE	BAAI General Embedding.
BEIR	Benchmarking Information Retrieval.
ETR	Efficient Title Reranker.
FiQA .	financial Question Answering.
HotpotQA	Hotpot Question Answering.
mGTE	Multilingual General Text Embeddings.
GPU	Graphics Processing Unit.
FAISS	Facebook AI Similarity Search.
RRF	Reciprocal Rank Fusion.
MAP	Mean Average Precision.
MMR	Maximal Marginal Relevance.
NQ	Natural Question.
TREC	Text Retrieval Conference.
SQuAD	Stanford Question Answering Dataset.
API	Application Programming Interface.
KILT	Knowledge Intensive Language Tasks.

List of Figures

Figure 1 Overview of the three phases of a RAG system: Indexing, Retrieval, and Generation. Adapted from Knollmeyer et al. (2026)	13
--	----

List of Tables

Table 1.9 Thesis Organization Explanation	12
Table 2.3 Comparative Analysis of Previous Studies on RAG	23

Chapter 1: Introduction

1.1 Overview

Retrieval-Augmented Generation (RAG) is the approach of enhancing Large Language Models (LLMs) answer accuracy and relevance. RAG can reduce hallucinations and outdated knowledge (Oche et al., 2025). The retrieval strategy is an important step towards getting the precise answers. Despite advancements in research in the field of RAG, specialized studies in Arabic are fewer compared to those in English, particularly in comparative methodologies across different retrieval methods. Furthermore, the complexity of the Arabic language and its grammatical rules may exacerbate these challenges.

1.2 Problem Definition

Digital information has rapidly grown in recent years, and there is a need for intelligent systems that can accurately retrieve and generate relevant information from large collections of documents. Many academic institutions, corporations, and healthcare organizations are increasingly adopting RAG to handle complex information queries. However, deploying such a system raises a practical question that current research does not clearly answer: which retrieval strategy should the system use to reliably surface the correct document for a given Arabic-language query?

It's harder to answer this question for Arabic than English due to how Arabic text is formulated. Arabic is linguistically rich and a highly complex language, characterized by intricate word formation, extensive morphology, and diverse dialects. Eventually, applying RAG retrieval methods to Arabic is completely different from retrieval in English because of these linguistic characteristics. Among the commonly used Natural Language Processing (NLP) preprocessing techniques for Arabic are normalization, stemming, and lemmatization (Alayba et al., 2025).

Stemming reduces words to their root forms, whereas lemmatization converts inflected words into their canonical dictionary forms (lemmas), these normalization techniques play a significant role for both sparse and dense retrieval systems by reducing lexical variation and improving text consistency. However, stemming may

negatively affect retrieval quality in Arabic, particularly for sparse retrieval methods such as BM25, because reducing words to their roots can remove important lexical distinctions. Therefore, sparse retrieval relies primarily on exact lexical matching rather than semantic understanding, extreme stemming may result in semantically different words sharing the same roots to be treated as identical, leading to less accurate retrieval results.

Recently a lot of studies have demonstrated the strength of sparse, dense, and hybrid retrieval methods, through evaluating these approaches independently or on English benchmark datasets. In addition, existing Arabic RAG research rarely examines how Arabic-specific linguistic characteristics affect retrieval components. Furthermore, no previous study has conducted a comprehensive comparison of sparse, dense, hybrid, and reranked retrieval methods within an Arabic RAG system.

1.3 Research Gap & Motivation

Arabic is a complex language, being linguistically very rich with complex morphology; however, due to its many dialects and the limited availability of good quality resources compared to English, it poses many challenges to information retrieval (IR) and response generation. Also, the literature lacks comprehensive evaluation studies of retrieval methods and reranking in Arabic RAG systems and their impacts on retrieval quality, answer accuracy, and faithfulness. This study addresses the gap in Arabic language studies in retrieval methods by conducting a comparative evaluation of retrieval methods for Arabic RAG. The study will be conducted using Midocean University documents as the evaluation documents as a realistic and divers selection of formal Arabic documents including policies, regulations, and procedural information, making them well suited for assessing retrieval performance in practical information-seeking scenarios.

The study is important to improve the quality of IR in Arabic by comparing and evaluating the best performance of different retrieval methods and reranking within the RAG system. Moreover, the study provides a scientific adviser for further researchers to select the most appropriate approach when building RAG systems in Arabic, based on the results.

1.4 Thesis Objectives

This study aims to address the gap by comparing and evaluating the performance of sparse, dense, hybrid, and reranked retrieval methods from Arabic academic documents in RAG; therefore, this study pursues the following four objectives:

1. To design and implement sparse, dense, and hybrid retrieval pipelines for an Arabic RAG system using academic and administrative documents.
2. To evaluate and compare the retrieval effectiveness of the sparse, dense, and hybrid methods using Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@k).
3. To apply a reranking stage to the best-performing method identified in Objective 2, and to evaluate whether reranking improves retrieval effectiveness relative to that method alone.
4. To evaluate whether reranking improves the quality of the answers generated by the RAG system, compared with answers generated from non-reranked retrieval output.

1.5 Research Questions

The research aims to focus on Arabic-retrieval therefore in this study, the study addresses the following questions:

1. Which retrieval method (sparse, dense or hybrid) achieves the highest retrieval effectiveness for Arabic university advising documents, as measured by Recall@k, MRR, NDCG@k for Arabic advising documents?
2. Does applying reranking to the best-performing method from the first question improve retrieval effectiveness compared with non-reranked retrieval?
3. Does reranking improve generated-answer quality compared to the initial (non-reranked) retrieval output?

1.6 Thesis Contribution

The work of this paper brings the following contributions to Arabic RAG field:

1. It offers a systematic comparative study of sparse, dense, and hybrid retrieval methods under a consistent experimental setup, using an Arabic document collection from a university. When it comes to evaluating retrieval components specifically, most existing Arabic RAG research examines individual methods in isolation, and few studies attempt a direct, side-by-side comparison (Abdelazim et al., 2023; El-Beltagy & Abdallah, 2024).
2. It measures the added benefit of re-ranking for retrieval performance and answer quality, and it isolates the impact of re-ranking from that of the initial base retrieval technique, something which was rarely addressed in previous Arabic RAG evaluations.
3. It presents a real-life case where selection of retrieval methods directly influences a level on which a query is handled, using Arabic institutional documents as an example. Thus, the research not only lends valuable support for Midocean University's own national digital library development project, but also provides ground, if not a blueprint, for other comparable studies on Arabic document collections.

1.7 Scope/Delimitation

This study is limited in scope to the following:

1. **Document domain and language:** This study evaluates the collection of Arabic academic documents from a single institution (MidOcean University), which include admission guidelines, policies, academic regulations, procedures, course information, and student services documents. The study is limited to Modern Standard Arabic is the limited study scope as it appears in formal institutional writing, and dialectal Arabic, informal text, and documents in other languages fall outside of its scope. Results do not claim to generalize beyond institutional document collections of a similar type.
2. **Retrieval and generation components:** The study compares the three retrieval methods: sparse retrieval (BM25), dense retrieval (BGE-M3), and hybrid retrieval (BM25 and BGE-M3 using RRF). For reranking, using (bge-reranker-v2-m3) applied to the best retrieval best-performing retrieval method

as defined in Thesis Objectives. The study does not evaluate different embedding models, rerankers, or fusion models other than those specified, nor does it compare multiple LLMs for answer generation; Gemini 2.5 Flash is used as the only generation model, consistent with the study's focus on retrieval methods comparison rather than generation-model comparison.

3. **Preprocessing:** Preprocessing techniques for Arabic include stemming, lemmatization, and diacritic handling, will not be treated as independent experimental variables. These linguistic aspects are analysed as characteristics that may affect the Arabic retrieval performance, yet the study does not conduct a separate experiment to disconnect their individual effects.
4. **Evaluation:** Retrieval effectiveness is tested using Recall@k, MRR, and NDCG@K, and answer quality is checked through human evaluation of relevance, correctness, and faithfulness, as defined in the resources needed section. This study does not evaluate system latency, computational cost, or scalability, and it does not evaluate actual students.
5. **Task setting:** The study is using a collection of Arabic academic documents for single-turn question answering. Multi-turn conversational RAG, and cross-lingual retrieval (e.g., English queries over Arabic documents) are outside the scope of this study.

1.8 Resources Needed

This study requires a combination of Arabic document resources, retrieval and generation components, and evaluation resources to implement and assess the retrieval strategies defined in the Research Objectives.

1. Dataset: Collection of Arabic University Advising Documents obtained from MidOcean University, including admission guidelines, academic regulations, policies, procedures, course information, and student services documents.

2. Retrieval methods: Sparse retrieval will be implemented using BM25. Dense retrieval will use the BGE-M3 embedding model, selected based on its reported performance on Arabic RAG tasks (Alsubhi et al., 2025). Hybrid retrieval will use BM25 and BGE-M3 through Reciprocal Rank Fusion (RRF).

3. Reranking: Reranking will use bge-reranker-v2-m3, consistent with its demonstrated effectiveness in prior Arabic RAG evaluation (Alsubhi et al., 2025).

4. Generation LLM: Gemini 2.5 Flash will serve as the Arabic-capable LLM and accessed through its application programming interface (API) to eliminate the need to host the model locally.

5. Retrieval Metrics: Will be evaluated using Recall@k, MRR, and NDCG@k, computed through standard evaluation by custom Python scripts.

6. Implementation Environment: Will be implemented in Python using Google Colab for embedding generation, indexing, retrieval, and reranking, with GPU resources used when required.

7. Answer Evaluation: Will be assessed through human evaluation conducted by Arabic-speaking evaluators.

1.9 Thesis Organization

Table 1.9 Thesis Organization Explanation

Chapter	Title	Contents
1	Introduction	Reviews the problem definition, research gap, objectives, contribution, Scope and resources needed.
2	Background and Related Work	Explain the foundations of RAG, sparse, dense, hybrid, and reranked retrieval, evaluation metrics, and Arabic-specific linguistic challenges, followed by a synthesis of related work.
3	Methodology	Describes the proposed methodology and system architecture.
4	System Implementation	Describes the project Implementation into a functional and operational software system.
5	Evaluation and Results	Reports the retrieval-effectiveness results and addresses the finding of each research question.
6	Discussion and Conclusion	Interprets the results, discusses the study's limitations, and presents its contributions and directions for future work.

Chapter 2: Background and Related Work

2.1 Background

2.1.1 Retrieval-Augmented Generation (RAG)

RAG is an artificial intelligence framework that combines IR with the optimization of LLMs to generate responses grounded on external, retrieved knowledge, rather than relying solely on parametric knowledge learned during pretraining, RAG has upgraded the FAQs applications and leveraged the quality of this field (Klesel & Wittmann, 2025).

Building a RAG system typically consists of four main stages. Collecting and preprocessing by cleaning and chunking the needed document, then the retrieval module identifies the most relevant documents for a given user query (Lewis et al., 2020; Gao et al., 2024). Retrieved documents are combined with the user's query and provided as contextual input to the LLM. Finally, the language model generates a response grounded in the retrieved evidence. The quality of RAG depends heavily on the retrieval stage. Selecting an appropriate retrieval strategy is critical factor in improving RAG performance (Huang & Huang, 2024; Gao et al., 2024).

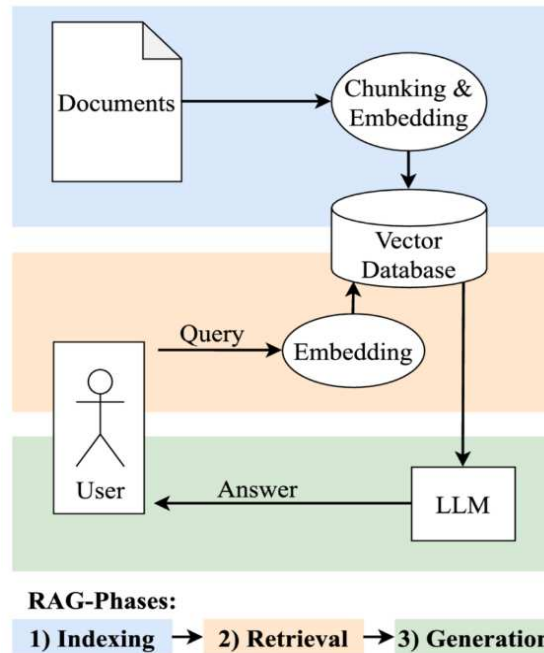


Figure 1 Overview of the three phases of a RAG system: Indexing, Retrieval, and Generation. Adapted from Knollmeyer et al. (2026)

2.1.2 Retrieval Methods

Retrieval methods are a crucial step of building RAG systems; it determines which document are provided to the LLM for answer generation. Many retrieval strategies have been developed through the years, each offering different strengths and weaknesses depending on the characteristics of the corpus, query type, and application domain.

2.1.2.1 Sparse Retrieval

Sparse is a lexical-based retrieval method that represents documents and queries using sparse, high-dimensional representations known as the bag-of-words model. Sparse is a retrieval method that ranks based on exact term matching. The most well-known sparse retrieval algorithms are TF-IDF and BM25, where BM25 is the most commonly used ranking function (Robertson & Zaragoza, 2009). Sparse retrieval has many advantages, mainly its computational efficiency, interpretability, and ability to scale to very large document collections. Since it relies on exact lexical matching, it performs well when users apply the same terminology found in the documents. However, if different words or expressions are used in the user's query,

sparse retrieval struggles, since it cannot match a query term to a document that expresses the same idea with different wording (Arabzadeh et al., 2021), this limitation becomes more pronounced in morphologically rich languages such as Arabic because multiple word forms may represent the same semantic meaning.

2.1.2.2 Dense Retrieval

Dense retrieval represents documents and queries as low-dimensional dense vectors generated by neural embedding models. Unlike sparse retrieval, which relies on exact keyword matching, dense retrieval searches for semantically similar vectors in a continuous embedding space. To measure how close a query and a document are in this space, dense methods usually rely on a vector-similarity metric such as cosine similarity or the dot product (Karpukhin et al., 2020). Regardless of using different vocabulary or paraphrased expressions, dense retrieval can retrieve relevant documents. Generally, dense retrieval achieves better semantic matching and is more robust for complex natural language queries. Yet, dense retrieval requires high-quality embedding models, vector indexing techniques, and substantially greater computational resources for embedding generation and similarity search (Gao et al., 2024).

2.1.2.3 Hybrid Retrieval

Hybrid retrieval is the combination of sparse and dense retrieval methods to get the advantages of both lexical matching and semantic retrieval. Mostly, documents are retrieved using both approaches and then merged using fusion algorithms or weighted ranking strategies. This combination improves retrieval robustness because one method (sparse) catches queries that share exact wording with a document, while the other (dense) catches queries that mean the same thing but use different words (Gao et al., 2024). Hybrid retrieval has consistently demonstrated higher retrieval performance than using either sparse or dense retrieval alone, particularly for heterogeneous document collections and diverse query types. Nevertheless, combining two retrieval pipelines increases implementation complexity, computational cost, and system latency.

2.1.2.4 Reranked Retrieval

Reranking refers to a second-stage retrieval method that comes after the first retrieval stage. Instead of pulling more documents, reranking changes the order of the documents already retrieved, top to bottom, that a human reader might be interested in most. Cross-encoders are one example of the more advanced models used in cross-document retrieval. They can take in both query and document as single examples and then, from a machine learning perspective, produce a more reliable measure of relevance (Nogueira & Cho, 2019). The main point of the paper is that reranking can improve ranking results significantly and get most useful documents to appear first.

Reranking is generally a process that evaluates documents one by one against the query and re-ranks them according to the level of relevance, which increases the computation and thus the time it takes to reply is also increased compared with simple retrieval.

2.1.3 Evaluation Metrics

In RAG models, the evaluation stage plays a significant role in determining the level of success of various search techniques. Retrieval performance is determined mainly by the number of relevant documents found and the position at which they were ranked. Standard IR methods that measure and quantify retrieval quality retrieval performance include metrics three widely used measures in this study like Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) (Manning et al., 2008).

2.1.3.1 Recall@k

Recall@k measures the proportion of relevant documents that are retrieved within the top k results for a given query. It is particularly important in RAG systems because the language model can only use the information contained in the retrieved context. Therefore, failing to retrieve a relevant document may prevent the generation model from producing a correct answer.

$$\text{Recall@k} = \frac{\text{Number of relevant documents retrieved in top } k}{\text{Total number of relevant documents}}$$

A higher Recall@k indicates that the retrieval method is more successful in finding relevant information. In this study, Recall@k can be used to compare the ability of sparse, dense, hybrid, and reranked retrieval methods to retrieve relevant university documents.

2.1.3.2 Mean Reciprocal Rank (MRR)

Mean reciprocal rank (MRR) is an evaluation method used in IR system where a relevant document is ranked as the one that first occurs after the first result. While Recall@k tells us whether relevant documents were retrieved, MRR is about how quickly the first relevant document appears in the ranking. As a result, a retrieval system's effectiveness can be judged by the means of the MRR for those cases when the top rank is supposed to be the most relevant document.

$$MRR = \frac{1}{Q} \sum_{i=1}^Q \frac{1}{rank_i}$$

where Q is the set of queries and $rank_i$ is the position of i query's first relevant document. Thus, a large value of Mean reciprocal rank means that the retrieval has returned the first document relevant to the query more frequently closer to the top of the results list. (Manning et al., 2008).

2.1.3.3 Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain (NDCG) serves as a performance measure in ranked retrieval that takes into account both the relevance of the retrieved documents and their position in the ranking. It awards high scores to documents that are not only relevant, but have been fetched at an early stage. This is one of the major advantages of NDCG over Recall@k it can be used to assess whether or not the reranking stage is effective at pushing heavily relevant documents toward the top of the ranking list.

NDCG is essentially derived from Discounted Cumulative Gain (DCG) which is expressed mathematically as:

$$DCG@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

NDCG is then calculated by normalizing DCG against the ideal ranking:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

where rel_i is the relevancy score of the document located at position i , and the IDCG symbolizes theoretically highest DCG that can be achieved in any given scenario. The NDCG can be anything from zero to one and in general it is better when higher values are achieved that is an indicator of a better-ranking model (Manning et al., 2008).

2.1.4 Arabic Language Challenges

Arabic is one of the most spoken languages worldwide, with over 400 million speakers. In recent years, many studies and innovations have been conducted and applied to the Arabic language. However, there is a gap in the studies regarding Arabic RAG retrieval; conducting this work for the Arabic language poses many challenges. One of the challenges is complex morphology; for many words, a root can be represented by three base consonants that carry the core meaning. This can affect the understanding of the retrieval method. Another challenge is diacritics, the small marks placed above or below letters to represent short vowels and specific phonetic rules. Nowadays, the use of diacritics is less common as words are understood through the context of the sentence (Alayba, 2025).

2.2 Related Work

Recent systematic reviews demonstrate the growing research interest in RAG. Brown et al. (2025) found that across the 2020–2025 period, the number of published articles rose steadily each year from 2020 to 2023, surpassing multiple application

domains. Similarly, Li et al. (2025) conducted in their study that collected 51 studies on the topic of using RAG in the field of education. Despite this expanding body of work, application of Arabic in RAG growing research interest but still underexplored (El-Beltagy & Abdallah, 2024).

This study review utilized approximately 17 related works; it is classification in terms of subject matter. Including Four studies that discuss the Arabic language in the field of RAG, one study taking about RAG for Educational Question Answering, nine studies specializing in retrieval methods, and three other studies for reranking.

2.2.1 Arabic RAG Studies

Many studies have evaluated the Arabic RAG system components; nevertheless, most studies focus on a single element (embedding models, LLM choice, or chunking) instead of comparing retrieval strategies.

El-Beltagy and Abdallah (2024) used Ar_EduText and Arabic Reading Comprehension Dataset (ARCD) to compare a semantic embedding model and LLMs for Arabic answering, GPT-3.5 Turbo results show outperformance on Ar_EduText, whilst Mistral performed best on ARCD, with Llama 3 competitive on both. Likewise, Abdelazim et al. (2023) evaluated ten multilingual embedding models for Arabic semantic retrieval on ARCD using Recall@k, achieving the best performance on Microsoft E5-Large, exceeding 90% Recall@10. The two studies found that embedding model choice substantially affects the retrieval and generation quality on Arabic; however, these studies do not compare sparse retrieval or hybrid retrieval against dense retrieval; their scope is focused on evaluating dense/embedding-based approaches against each other.

On a more comprehensive evaluation, Alsubhi et al. (2025) conducted a study comparing chunking strategies, embedding models, reranking techniques, and LLMs on ARCD. The results showed BGE-M3 and Multilingual-E5-large achieving the strongest embedding scores, along with reranking (bge-reranker-v3-m3) improving retrieval quality and achieving the best generation performance in Aya-8B. For the present research, this is the nearest study in scope, as it evaluates reranking with retrieval components rather than isolation. Yet, it does not add a sparse retrieval baseline.

Alothman et al. (2026) developed an Arabic RAG pipeline combining multilingual-e5-large retrieval with NER-guided retrieval and an agentic query router, showing an Acc@1 of 58.7% and MRR of 0.717 for the retriever, along with Gemini-2-Flash achieving 0.83 answer relevance and 0.94 faithfulness. The authors note that evaluation depends mainly on RAGAS (Es et al., 2024) and a single prompt template which limited their ability to assess different retrieval and reasoning methods. This study demonstrates that architectural additions (query routing, NER-guidance) can enhance Arabic RAG performance, however it does not separate the effect of the underlying retrieval strategy itself, since dense retrieval was the only retrieval method used.

Generally, these studies establish that embedding model selection, reranking, and architectural additions independently affect Arabic RAG performance. However, none provides a controlled comparison across sparse, dense, hybrid, and reranked retrieval within a single, consistent experimental setup: El-Beltagy and Abdallah (2024) and Abdelazim et al. (2023) compare only within dense retrieval; Alsubhi et al. (2025) evaluates reranking but discards a sparse baseline; and Alothman et al. (2026) evaluates only dense retrieval within a more complex pipeline. This leaves open the specific question this study addresses: how sparse, dense, and hybrid retrieval compare directly against one another, and what reranking adds once applied to the strongest of these, within a single Arabic institutional document collection.

2.2.2 RAG Application in Educational Contexts

Wang and Singh (2025) developed a system to answer course-related student queries by combining open-source LLMs with RAG and fine-tuning for a domain specific question-answering system. Their methodology involved building a knowledge base from course materials, collecting relevant content through RAG to support user queries. This combination of fine-tuning and RAG heavily enhanced answer accuracy, with F1 increasing from 26.8 without fine-tuning to 62.2 with fine-tuning. However, the study does not compare alternative retrieval strategies to determine which most effectively supports answer generation, and the authors note that course content is not purely text-based, both of which limit the extent to which its findings generalize to retrieval-method selection.

2.2.3 Retrieval Methods Studies

In addition to Arabic-specific research, a broader body of research has compared sparse, dense, and hybrid retrieval strategies, focusing on English and multilingual benchmarks and establishing the general trade-offs between them, but does not evaluate them on Arabic text specifically.

2.2.3.1 Sparse versus dense retrieval.

Arabzadeh et al. (2021) compared two retrievals (sparse and dense) directly, the results showed that dense retrieval generally performed better for semantically complex queries, while sparse retrieval was more effective for keyword-specific queries and required fewer computational resources. Qu et al. (2021) focused specifically on improving dense retrieval through the RocketQA framework, reporting higher performance than prior dense approaches on Microsoft Machine Reading Comprehension (MS MARCO) and Natural Questions (NQ). These studies establish the semantic-versus-efficiency trade-off that motivates hybrid approaches, but both of studies evaluated only on English benchmarks, this raises the question the same trade-off applies for an Arabic, morphologically rich language.

2.2.3.2 Hybrid retrieval.

The results of several studies have shown the superiority of hybrid retrieval over sparse and dense retrieval alone.

Lee et al. (2023) proposed to apply a Ratio of Complementarity (RoC) objective to increase the complementarity between sparse and dense retrievers. This work outperforms hybrid methods on MS MARCO, NQs, and Text Retrieval Conference (TREC) Robust04. Sawarkar et al. (2024) proposed another work Blended RAG, combining sparse and dense retrieval with hybrid querying, and reported improved retrieval performance on NQs, TREC-COVID, and Stanford question answering dataset (SQuAD). Arivazhagan et al. (2023) took a different route, layering sparse and dense retrieval into a Hybrid Hierarchical Retrieval (HHR) pipeline that lifted Recall@100 by 4.69% relative to a dense-only setup across NQs, Web Questions, and TriviaQA. These results consistently show hybrid retrieval as a stronger general-purpose approach. However, none of these studies evaluate Arabic languages where

the lexical mismatch (central weakness of hybrid retrieval addresses) is expected to be more pronounced due to Arabic's morphological richness.

2.2.3.3 Domain-specific hybrid applications.

Hybrid retrieval has also been evaluated in specialized domains, showing that its advantage depends on document and query characteristics. In a financial dataset containing a mix of text and tables, Akarsu et al. (2026) evaluated ten different retrieval strategies against each other. These strategies spanning sparse, dense, hybrid fusion, cross-encoder reranking, query expansion, and adaptive retrieval on a financial text-and-table benchmark after compared, found that pairing hybrid retrieval with neural reranking beat every single-stage method achieved best performance ($\text{Recall}@5 = 0.816$, $\text{MRR}@3 = 0.605$). In biomedical RAG Bal and Puhon (2026) compared dense retrieval, hybrid BM25 plus dense retrieval, cross-encoder reranking, multi-query expansion, and maximal marginal relevance (MMR) for biomedical RAG. Finding hybrid retrieval provides the best balance of accuracy and robustness. Irany and Akwafuo (2026) combined hybrid retrieval with neural reranking and claim-level grounding evaluation, achieving 100% grounding accuracy across a 25-query pilot evaluation. Abu Ahmad et al. (2025) developed a three-stage hybrid RAG pipeline for Islamic knowledge tasks, combining BM25 sparse retrieval, dense embedding retrieval, and cross-encoder reranking, reporting accuracy improvements of up to 25% over non-hybrid baselines depending on task and model configuration. Notably, this is the only study in this group applying sparse, dense, and reranked retrieval jointly to Arabic-language text; however, the components are combined into one fixed pipeline rather than compared against each other individually, so it does not establish which stage contributes most to performance, nor how each method would perform in isolation on this Arabic corpus.

Collectively, this literature demonstrates that retrieval strategy meaningfully affects RAG performance, that the best-performing strategy depends on dataset, domain, and query characteristics, and that hybrid retrieval frequently — though not universally — outperforms single-method approaches. However, this evidence base is drawn almost entirely from English or domain-specific (financial, biomedical) collections rather than Arabic text. Where Arabic text does appear (Abu Ahmad et al., 2025), the comparison is not systematic: sparse, dense, and reranking are

combined as a fixed pipeline, not benchmarked individually against each other on the same Arabic corpus. This leaves open whether the sparse-dense-hybrid trade-offs established here generalize to Arabic, and specifically to Arabic university administrative documents — the question this study investigates directly.

2.2.4 Ranking Models

Moreira et al. (2024) evaluated several reranking models, including Mini Language Model (MiniLM), BAAI General Embedding (BGE), and Jina, on Benchmarking IR (BEIR) benchmark datasets such as NQs, Hotpot Question Answering (HotpotQA), and Financial Question Answering (FiQA). Their results showed that the proposed NV-RerankQAMistral-4B-v3 achieved a 14% improvement in accuracy compared with other reranking pipelines, demonstrating the effectiveness of advanced reranking models in improving retrieval performance. However, the authors highlighted the additional computational requirements and their potential impact on indexing and serving latency. Likewise, Chen et al. (2023) proposed an Efficient Title Reranker (ETR) to reduce the computational cost of reranking while maintaining retrieval effectiveness. The proposed approach achieved 20–40 times faster reranking than traditional methods while maintaining comparable accuracy. However, the method was limited to reranking document titles rather than the full document content, which may affect its ability to assess the relevance of detailed textual information.

Then as well, Zhang et al. (2024) proposed Multilingual General Text Embeddings (mGTE), a multilingual framework that combines dense retrieval with cross-encoder reranking for multilingual and long-text retrieval. Evaluated across multilingual benchmarks covering more than 100 languages, mGTE outperformed existing multilingual models such as BGE-M3 in retrieval and reranking performance. The study identified the high computational cost of training and inference for long-context models as a major challenge.

These studies show that reranking can substantially improve retrieval effectiveness by reordering the initially retrieved candidates according to their relevance. However, improved ranking performance often comes with additional computational cost and latency. Lastly, existing studies have mainly focused on English or multilingual benchmark datasets, while limited research has examined

reranking within Arabic RAG systems using university documents. Therefore, the current study includes reranking as one of the retrieval strategies to determine whether the additional ranking stage can improve retrieval performance and answer quality for Arabic academic documents.

Table 2.3 Comparative Analysis of Previous Studies on RAG

Authors	Methodology	Key Findings	Weaknesses/ Limitations
Arabic RAG Studies			
Abdelazim et al. (2023)	Evaluation of 10 multilingual embedding models using the ARCD dataset	Microsoft E5-Large achieved the best retrieval performance, obtaining Recall@15 of 93%	Limited dataset
El-Beltagy & Abdallah (2024)	Comparative evaluation of (RAG) different semantic embedding models and (LLMs) in answering Arabic questions using the Ar_EduText and ARCD datasets	E5-Large achieved the best retrieval performance, with MRR scores of 0.934 on Ar_EduText and 0.790 on ARCD	Improved prompting strategies to enhance answer quality
Alsubhi et al (2025)	Compare chunking strategies, multilingual and Arabic-specific embedding models, rerankers, and LLMs.	BGE-M3 achieved the best overall embedding score 70.99, followed by Multilingual-E5-large 70.31. Reranking with bge-reranker-v2-m3 increased score from 70.99 to 74.15, with the largest gain on ARCD 80.29 to 86.31.	Limited dataset

Alothman et al. (2026)	Data preparation, retrieval, and generation—coordinated by an intelligent agentic query router	Multilingual-e5-large achieved 58.7% Acc@1 and 0.715 MRR. Gemini-2-Flash achieved 0.83 answer relevance, 0.84 context utilization, and 0.94 faithfulness.	-Some models sporadically ignored the required JSON schema. - Evaluation Bias in RAGAS may inadvertently encode Arabic-specific
RAG Application in Educational Contexts			
Wang & Singh (2025)	Constructing a local knowledge base containing course materials, retrieving relevant documents through a RAG to support user queries, and integrating multi-chain-of-thought reasoning to reduce hallucination in LLM outputs.	Fine-tuned RAG achieved F1 = 62.2, compared with 26.8 without fine-tuning it shows the combination of fine-tuned LLMs with RAG and reasoning mechanisms significantly improves answer accuracy.	They found course content is not just text-based and there is not compare different retrieval strategies to decide which is suppurative retrieval methods.
Retrieval Methods Studies – Sparse vs. Dense			
Arabzadeh et al. (2021)	The Author propose a classifier to choose the best retrieval model (sparse vs. dense vs. hybrid) for individual queries. Using standard IR benchmark datasets. The adaptive method was compared with standalone dense and sparse retrieval methods based on retrieval effectiveness and computational efficiency.	The proposed classifier achieved Recall@1000 of 0.95 for sparse–dense selection and 0.96 for sparse–hybrid selection	Dense retrieval requires significant computational resources and embedding generation, and sparse retrieval may fail to capture semantic similarities when exact keywords are absent
Qu et al. (2021)	Introduced an optimized training approach, RocketQA, to improve dense passage retrieval.	Recall@100 achieved the best performance = 88.5 on NQ.	limited negative samples during training, noisy false-negative passages, and the lack of sufficient labeled data.
Retrieval Methods Studies – Hybrid Retrieval			
Lee et al. (2023)	Proposing a two-level orthogonality designed to improve RoC (Ratio of Complementarity) by encouraging them to representation diversity between sparse and dense retrievers	MRR@10 by 0.78 points (from 37.85 to 38.64) on MS MARCO, and mean average precision (MAP) by 1.50 and 2.21 points on NQ and TREC Robust04, respectively	Designing an objective that encourages the two retrieval methods to contribute distinct yet complementary evidence without

			sacrificing retrieval accuracy.
Arivazhagan et al. (2023)	Hierarchical retrieval framework that combines sparse, dense, and hybrid retrieval methods	(HHR) method outperformed the (DHR) method by 4.69%.	Limited dataset
Sawarkar et al. (2024)	Develop a hybrid retrieval framework (Blended RAG), which combines dense and sparse retrieval techniques with hybrid querying	Superiority of the hybrid Blended RAG achieved NDCG@10 of 0.67 on NQ (+5.8%) and 0.87 on TREC-COVID (+8.2%); on SQuAD, it achieved F1 = 68.4% and Top-5 retrieval accuracy = 94.89%.	Not investigate Reranking
Retrieval Methods Studies – Domain Hybrid Applications			
Abu Ahmad et al. (2025)	Hybrid RAG using BM25 + Arabic dense retrieval + MiniLM cross-encoder reranking, followed by Fanar/Mistral generation.	Improved evidence result accuracy 80% (Fanar) achieved.	Complexity of Islamic reasoning; these documents require a deep understanding
Akarsu et al (2026)	Comprehensive comparison of retrieval techniques used in RAG systems for financial question-answering tasks involving documents that contain both textual content and tables.	BM25 outperformed dense retrieval. Hybrid with neural reranking achieved Recall@5 = 0.816 and MRR@3 = 0.605, outperforming single-stage methods.	Retrieval in mixed text-and-table documents, where textual descriptions provide contextual information and tables contain precise numerical values.
Bal, Devi & Puan, Subhashree (2026)	Comparison five retrieval strategies Dense Retrieval, Hybrid BM25 + Dense, Cross-Encoder Reranking, Multi-Query Expansion, and MMR using biomedical benchmark datasets for Biomedical (RAG) and choose which methods provide the most accurate document retrieval	Reranking (Cross-Encoder) performed best (composite = 0.827; contextual precision = 0.852), outperforming Dense Retrieval (0.822).	Data complexity needs a high-quality retrieval
Irany & Akwafuo (2026)	Framework for retrieval and reranking with answer generation and grounding evaluation to ensure the quality of biomedical and healthcare information	Evaluation against the retrieved evidence with 100% accuracy Hybrid retrieval with Cohere reranking achieved 100% grounding accuracy only 8.8% of retrieved chunks retained their original rank after reranking	Small dataset

Ranking Models			
Chen et al. (2023)	Enhancing efficiency title reranker. through Broadcasting Query Encoder whith knowledge intensive language tasks (KILT)Benchmark.	Title reranking achieves a 20x-40x speedup rather than the vanilla passage reranke	May retrieve incorrect Documents and limited to reranking titles
Moreira et al (2024)	Compare between several reranking models on the BEIR benchmark datasets using retrieval metrics such as (NDCG@10, MRR, and Recall@k)	NV-RerankQAMistral-4B-v3 model, achieving a significant accuracy increase of 14% compared to other reranking pipelines	Bad impacts indexing and latency
Zhang et al. (2024)	A multilingual embedding model for dense retrieval and a cross-encoder reranking model.	mGTE-reranker model achieved superior performance in multilingual retrieval and reranking 78.7 and 68.5, outperforming existing multilingual models such as BGE-M3	High computational cost

Reference

Abdelazim, H., Tharwat, M., & Mohamed, A. (2023). *Semantic embeddings for Arabic retrieval augmented generation (ARAG)*. *International Journal of Advanced Computer Science and Applications*, 14(11), 1328. <https://doi.org/10.14569/ijacsa.2023.01411135>

Abu Ahmad, Muhammad & Ballout, Mohamad & Abu Ahmad, Raia & Bruni, Elia. (2025). *Transformer Tafsir at QIAS 2025 Shared Task: Hybrid Retrieval-Augmented Generation for Islamic Knowledge Question Answering*. <https://doi.org/10.48550/arXiv.2509.23793>

Alayba, A. M. (2025). *Arabic Natural Language Processing (NLP): A comprehensive review of challenges, techniques, and emerging trends*. *Computers*, 14(11), 497. <https://doi.org/10.3390/computers14110497>

Alothman, M., Altammami, A., Alhoshan, M., Almazrui, A., Al-Rasheed, R., Almatham, R., Altamimi, A., Elnashar, O., Amin, M., Almuqhim, B., Alosaimy, A., & Alfaifi, A. (2026). *ARAG: An agent-based hybrid semantic-lexical retrieval-augmented generation pipeline for Arabic texts*. *Procedia Computer Science* 275, 682–691. <https://doi.org/10.1016/j.procs.2026.01.079>

Alsubhi, A., Aljuaid, H., Alharthi, A., Alotaibi, N., & Alzahrani, M. (2025). *Optimizing RAG pipelines for Arabic: A systematic analysis of core components*. *arXiv*. <https://doi.org/10.48550/arXiv.2506.06339>

Akarsu, M. M., Karaman, R. K., & Mierbach, C. (2026). *From BM25 to corrective RAG: Benchmarking retrieval strategies for text-and-table documents*. *arXiv*. <https://doi.org/10.48550/arXiv.2604.01733>

- Arabzadeh, N., Yan, X., & Clarke, C. L. A. (2021). *Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection*. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)* (pp. 2862–2866). Association for Computing Machinery. <https://doi.org/10.1145/3459637.3482159>
- Arivazhagan, M. G., Liu, L., Qi, P., Chen, X., Wang, W. Y., & Huang, Z. (2023). *Hybrid hierarchical retrieval for open-domain question answering*. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10680–10689). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.679>
- Bal, D., & Puhan, S. (2026). *Benchmarking retrieval strategies for biomedical retrieval-augmented generation: A controlled empirical study*. <https://doi.org/10.48550/arXiv.2605.02520>
- Brown, A., Roman, M., & Devereux, B. (2025). A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *Big Data and Cognitive Computing*, 9(12), 320. <https://doi.org/10.3390/bdcc9120320>
- Chen, W., Zhang, Y., Lin, J., & Van Durme, B. (2023). *Efficient title reranking for fast and improved knowledge-intensive NLP*. arXiv. <https://doi.org/10.48550/arXiv.2312.12430>
- El-Beltagy, S. R., & Abdallah, M. A. (2024). *Exploring retrieval augmented generation in Arabic*. *Procedia Computer Science*, 244, 296–307. <https://doi.org/10.1016/j.procs.2024.10.203>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). *RAGAs: Automated Evaluation of Retrieval Augmented Generation*. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1434–1451. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Wang, Y., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. *ACM Computing Surveys*. <https://doi.org/10.48550/arXiv.2312.10997>
- Huang, Y., & Huang, J. (2024). *A survey on retrieval-augmented text generation for large language models*. arXiv. <https://arxiv.org/abs/2404.10981>
- Irany, F. A., & Akwafuo, S. (2026). *A hybrid retrieval and reranking framework for evidence-grounded retrieval-augmented generation*. <https://doi.org/10.48550/arXiv.2605.01664>
- 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 2020*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>

- Klesel, M., & Wittmann, H. F. (2025). *Retrieval-Augmented Generation (RAG)*. *Business & Information Systems Engineering*, 67(4), 551–561. <https://doi.org/10.1007/s12599-025-00945-3>
- Knollmeyer, S., Caymazer, O., Koval, L. *et al.* Evaluating Retrieval Augmented Generation: A Comprehensive Review of Evaluation Dimensions, Question Types, and Application. *SN COMPUT. SCI.* 7, 576 (2026). <https://doi.org/10.1007/s42979-026-05134-x>
- Lee, D., Hwang, S.-W., Lee, K., Choi, S., & Park, S. (2023). *On complementarity objectives for hybrid retrieval*. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13357–13368). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.746>
- Lewis, P., Perez, E., Piktus, A., Petroni, V., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://arxiv.org/abs/2005.11401>
- Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8, Article 100417. <https://doi.org/10.1016/j.caeai.2025.100417>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- Moreira, G. D. S. P., Ak, R., Schifferer, B., Xu, M., Osmulski, R., & Oldridge, E. (2024). *Enhancing Q&A text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for RAG*. arXiv. <https://doi.org/10.48550/arXiv.2409.07691>
- Nogueira, R., & Cho, K. (2019). *Passage re-ranking with BERT* (arXiv:1901.04085). ArXiv. <https://doi.org/10.48550/arXiv.1901.04085>
- Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). *A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions*. arXiv. <https://doi.org/10.48550/arXiv.2507.18910>
- Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., & Wang, H. (2021). *RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5835–5847). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.466>

- Robertson S, Zaragoza H (2009), "The Probabilistic Relevance Framework: BM25 and Beyond". *Foundations and Trends in Information Retrieval*, Vol. 4 No. 1-2 pp. 1–174, <https://doi.org/10.1561/15000000019>
- Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). *Blended RAG: Improving retrieval-augmented generation accuracy with semantic search and hybrid query-based retrievers*. arXiv. <https://doi.org/10.48550/arXiv.2404.07220>
- Wang, N., & Singh, S. (2025). *Answering students' questions on course forums using multiple chain-of-thought reasoning and finetuning RAG-enabled LLM*. arXiv. <https://doi.org/10.48550/arXiv.2511.09831>
- Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., Shoneybi, M., & Catanzaro, B. (2024). *RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs*. *Advances in Neural Information Processing Systems*, 37. <https://doi.org/10.52202/079017-3850>
- Zhang, X., Wang, Y., Chen, Z., Liu, X., Zhang, Y., Li, Z., ... & Xiao, J. (2024). *mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval*. arXiv. <https://doi.org/10.48550/arXiv.2407.19669>

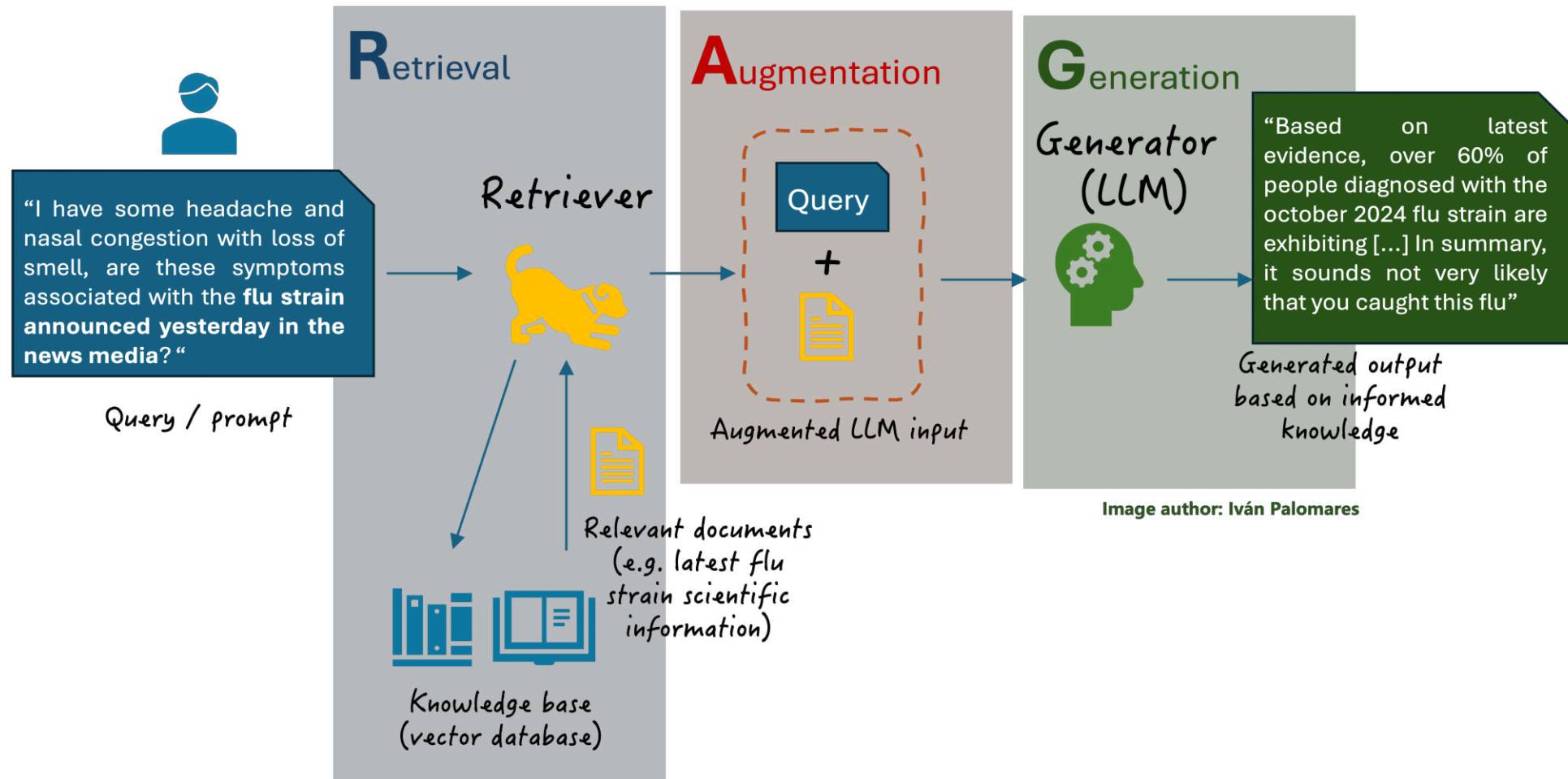
Comparative Evaluation of Retrieval Methods for Arabic Retrieval-Augmented Generation

BY: HADEEL ALSAKKAF & SHAHAD DAIRI

What is Retrieval-Augmented Generation (RAG)?

RAG is an artificial intelligence framework that combines information retrieval with the optimization of large language models to generate responses based on the model parametric knowledge

What is Retrieval-Augmented Generation (RAG)?



Problem Definition

- Organizations across different domains are increasingly adopting RAG systems to retrieve and generate information from large document collections.
- The effectiveness of RAG depends heavily on selecting an appropriate retrieval strategy, and there is no clear answer regarding which strategy performs best for Arabic queries.
- Arabic presents additional challenges due to its rich morphology, inflection, and lexical variation, which can affect retrieval performance.
- Existing research has limited comprehensive comparisons of sparse, dense, hybrid, and reranked retrieval approaches for Arabic RAG systems.

What are the retrieval methods in RAG?

Retrieval methods in RAG fetch relevant text segments for a user query to ground the LLM response. The core methods include:

Sparse Retrieval

- keyword matching like BM25

Dense Retrieval

- semantic vector similarity

Hybrid search

- Combining both

Thesis Objectives

- 1 • Build Sparse, Dense, and Hybrid retrieval pipelines for Arabic academic and administrative documents.
- 2 • Compare their effectiveness using Recall@k, MRR, and NDCG@k.
- 3 • Apply reranking to the best-performing method and test whether it improves retrieval further.
- 4 • Evaluate whether reranking improves the quality of the final generated answers.

Thesis Contribution

1

The first controlled, side-by-side comparison of Sparse, Dense, Hybrid, and Reranked retrieval for Arabic RAG.

2

measures the added benefit of re-ranking for retrieval performance

3

A practical, evidence-based blueprint for building Arabic RAG systems, tested on real university documents.

Research Questions

RQ1

Which retrieval method (sparse, dense or hybrid) achieves the highest retrieval effectiveness for Arabic university advising document, as measured by Recall@k, MRR, NDCG@k for Arabic advising documents?

RQ2

Does applying reranking to the best-performing method from the first question improve retrieval effectiveness compared with non-reranked retrieval?

RQ3

Does reranking improve generated-answer quality compared to the initial (non-reranked) retrieval output?

How Do We Measure Success?

Recall@k

Measures whether the relevant passage appears anywhere in the top k retrieved results

MRR

Measures how close to the top the first relevant result appears — higher is better.

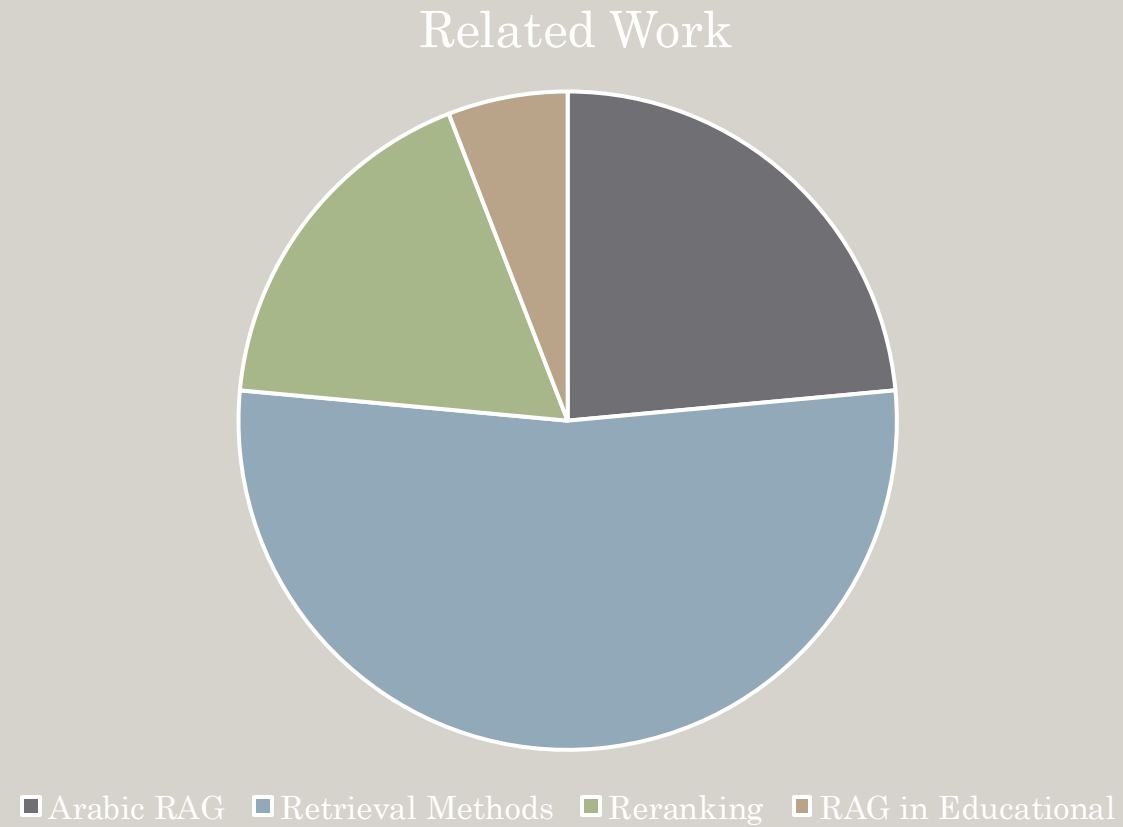
NDCG@k

Rewards relevant results that also appear early in the ranking, not just present.

Resources Needed

Component	Final Choice
Sparse Retrieval	BM25
Dense Retrieval	BGE-M3
Hybrid Retrieval	BM25 + BGE-M3 + RRF
Reranker	bge-reranker-v2-m3
Generation LLM	Gemini 2.5 Flash
Retrieval Metrics	Recall@k, MRR, NDCG@k
Answer Evaluation	Human evaluation by Arabic-speaking evaluators
Ground Truth	Annotator-created and supervisor-validated relevance judgments

Related Work



THANK YOU