

هديل السقاف

RAG6_Hadeel.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:406763860

Submission Date

Aug 12, 2026, 12:30 AM GMT+3

Download Date

Aug 14, 2026, 8:20 PM GMT+3

File Name

RAG6_Hadeel.docx

File Size

234.2 KB

28 Pages

6,415 Words

40,832 Characters

34% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely revised by AI.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.



11 AI-generated 34%

Likely generated or likely generated and revised by AI.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely human generated as AI generated and likely AI generated as human generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Comparative Evaluation of Retrieval Methods for Arabic Retrieval-Augmented Generation

Prepared By

Name: Hadeel Alsakkaf

ID No.: 261002581

Name: Shahad 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

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. The retrieval component is a key factor in the performance of a RAG system, as the quality of the retrieved documents directly influences the generated responses. Common retrieval strategies include sparse retrieval (keyword-based), dense retrieval (semantic-based), and hybrid retrieval, which combines both approaches, afterwards, reranking can be applied to further improvement of the retrieved answer. Although these retrieval methods have been widely studied, comparative evaluations for Arabic RAG systems remain limited. This is mainly due to the unique characteristics of the Arabic language, such as its rich morphology, lexical variations, and optional diacritics, which can affect retrieval performance. Therefore, this study aims to compare sparse, dense, and hybrid retrieval and reranked strategies using Midocean University admission and academic documents to identify the most effective approach for Arabic information retrieval. The findings are expected to provide practical guidance for developing more accurate and reliable Arabic RAG systems for university question-answering and other domain-specific applications.

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

List of Publications

Table of Contents

Abstract.....	1
List of Publications	2
List of Abbreviations	4
List of Figures.....	5
List of Tables	5
Chapter 1	6
Introduction.....	6
1.1 Overview	6
1.2 Problem Definition	7
1.3 Thesis Objectives	8
1.4 Research Questions	8
1.5 Thesis Contribution.....	9
1.6 Resources Needed	9
Chapter 2	11
Background and Related Work.....	11
2.1 Background	11
2.1.1 Retrieval-Augmented Generation (RAG)	11
2.1.2 Retrieval Methods	12
2.1.3 Evaluation Metrics	14
2.1.4 Arabic Language Challenges.....	16
2.2 Related work.....	17
2.2.1 Arabic RAG Studies.....	17
2.2.2 Retrieval Methods Studies.....	18
2.2.3 Ranking Models.....	20
Table 2.3 Comparative Analysis of Previous Studies on RAG.....	21
Reference	24

List of Abbreviations

RAG	Retrieval-Augmented Generation
LLM	Large Language Model
LLMs	Large Language Models
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

List of Figures

Figure 1. Overview of the three phases of a RAG system: Indexing, Retrieval, and Generation.
..... **11**

List of Tables

Table 2.3 Comparative Analysis of Previous Studies on RAG **21**

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 into 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.

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. RAG has achieved strong performance in many English-language applications. However, thus far its strength has yet to be established in Arabic. Arabic has rich linguistic characteristics, rich morphology, diverse dialects, and limited availability of high-quality language resources compared to English. Arabic language shows a unique challenge for document retrieval and response generation. Further, there is a lack of comprehensive studies that compare retrieval methods within Arabic RAG systems and evaluate their impact on retrieval quality and answer accuracy. 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 corpus being a realistic and divers selection of formal Arabic documents containing policies, regulations, and procedural information, making them well suited for assessing retrieval performance in practical information-seeking scenarios.

The importance of this study lies in improving the quality of information retrieval in Arabic by comparing and evaluating the best performance of different retrieval methods and reranking within the RAG system. It also supports and develops the university's smart electronic system for university student inquiries by improving

and facilitating access to information found in documents such as admissions guidelines and policies. Furthermore, it provides a scientific guide for researchers and developers to select the most appropriate approach when building RAG systems in Arabic, based on the results and metrics.

1.2 Problem Definition

Arabic is a linguistically rich and highly complex language, characterized by intricate word formation, extensive morphology, diverse dialects, and sophisticated syntactic structures (Alayba et al., 2025). Consequently, applying RAG retrieval methods to Arabic differs significantly from retrieval in English due to these linguistic characteristics. Among the most widely used Natural Language Processing (NLP) preprocessing techniques for Arabic are normalization, stemming, lemmatization, and word segmentation.

Stemming reduces words to their root forms, whereas lemmatization converts inflected words into their canonical dictionary forms (lemmas). These normalization techniques play an important role in 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, reducing words to their roots can remove important lexical distinctions. Since sparse retrieval relies primarily on exact lexical matching rather than semantic understanding, excessive stemming may cause 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 effectiveness 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, such as rich morphology, optional diacritics, and stemming strategies, influence the effectiveness of retrieval methods. Furthermore, no previous study has conducted a comprehensive comparison of Sparse, Dense, Hybrid, and reranked retrieval methods within an Arabic RAG system designed for university documents. This study addresses these

gaps by providing a systematic evaluation of the four retrieval strategies using Arabic academic documents and analysing their performance under the linguistic characteristics of Arabic.

1.3 Thesis Objectives

This study aims to address this 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, MRR, and 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.

1.4 Research Questions

The review aims to focus on Arabic-retrieval therefore in this study, the research questions are as follows:

RQ1: How can retrieval methods (Sparse, Dense, Hybrid) differ in retrieval effectiveness (Recall@k, MRR, NDCG@k) for Arabic advising documents?

RQ2: Does applying reranking to the best-performing method from RQ1 improve retrieval effectiveness compared with non-reranked retrieval?

RQ3: Does reranking improve generated-answer quality compared to the initial (non-reranked) retrieval output?

1.5 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. In the area of retrieval component evaluation, Arabic RAG research work has mainly been done separately while comparative evaluation has been rarely treated. (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 method 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.

1.6 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.

The primary data resource is a collection of Arabic academic documents obtained from MidOcean University, including admission guidelines, academic regulations, policies, procedures, course information, and student services documents. These documents form the knowledge base for the RAG system. A set of Arabic questions related to university academic advising will be constructed to evaluate retrieval and answer-generation performance, together with ground-truth relevance judgments identifying the relevant document or chunk for each question; these judgments will

be annotator-created and supervisor-validated to support the reliability of the evaluation.

For retrieval, sparse retrieval will be implemented using BM25. Dense retrieval will use the BGE-M3 embedding model, selected on the basis of its reported performance on Arabic retrieval-augmented generation tasks (Alsubhi et al., 2025). Hybrid retrieval will combine BM25 and BGE-M3 through Reciprocal Rank Fusion (RRF), which merges the ranked outputs of both methods without requiring manually tuned fusion weights. Reranking will be performed using bge-reranker-v2-m3, consistent with its demonstrated effectiveness in prior Arabic RAG evaluation (Alsubhi et al., 2025). A vector database (e.g., FAISS, Chroma, or Qdrant) will store and retrieve document embeddings for the dense and hybrid retrieval components. For answer generation, Gemini 3.5 Flash will serve as the Arabic-capable LLM that produces responses grounded in the retrieved document context.

Retrieval performance will be evaluated using Recall@k, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG@k), computed through standard evaluation libraries or custom Python scripts. Answer quality will be assessed through human evaluation conducted by Arabic-speaking evaluators, who will judge the relevance, correctness, and faithfulness of generated answers relative to the retrieved source documents.

The system will be implemented in Python, using a standard development environment (Google Colab), optionally supported by a RAG framework such as LangChain or LlamaIndex to integrate the retrieval and generation components. Embedding generation, indexing, and reranking will be run on personal computing hardware. Given the computational demands of the BGE-M3 and bge-reranker-v2-m3 models, particularly for reranking, which scores each query-document pair individually rather than through a single vector comparison, computational feasibility will first be validated on a representative subset of the document collection before full-scale indexing and evaluation. Where local GPU resources prove insufficient, free-tier cloud compute (e.g., Google Colab) may be used to supplement personal hardware for the embedding and reranking stages. Gemini 3.5 Flash, accessed via API, does not impose local computational demands.

Chapter 2

Background and Related Work

2.1 Background

2.1.1 Retrieval-Augmented Generation (RAG)

RAG is an artificial intelligence framework that combines Information Retrieval (IR) with the optimization of (LLMs) to generate responses based on the model parametric knowledge, it focuses on a specific domain and limited answers learned from pretraining, RAG has upgraded the FAQs applications and leveraged the quality of this field (Klesel & Wittmann, 2025).

Building a RAG system consist of four stages. Collecting and preprocessing by cleaning and chunking through the needed document, then the retrieval module identifies the most relevant documents for a given user query. 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.

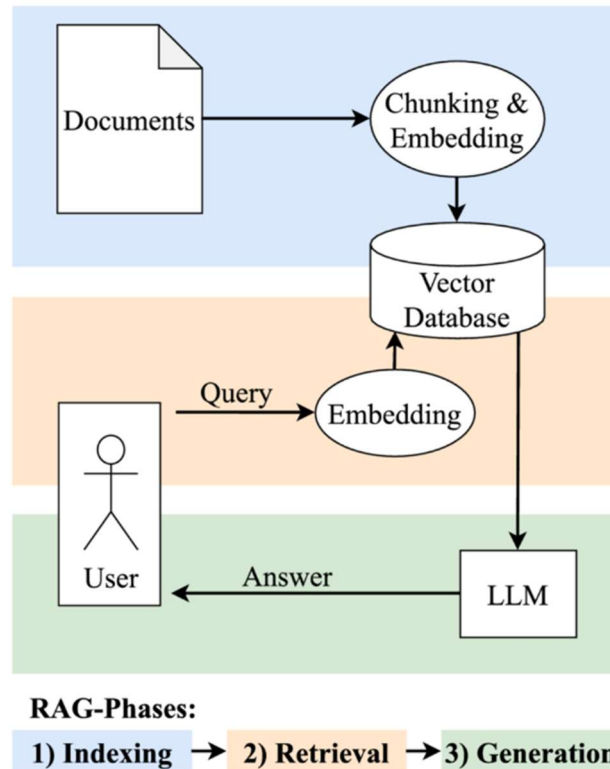


Figure 1. Overview of the three phases of a RAG system: *Indexing, Retrieval, and Generation.*
 Adapted from Knollmeyer et al. (2026)

2.1.2 Retrieval Methods

Retrieval methods is 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 strength and weakness 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 mostly 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 would suffer from the vocabulary mismatch problem (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. Similarity is typically computed using 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 sparse retrieval captures exact keyword matches, whereas dense retrieval captures semantic similarity (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 by a lot 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 accordingly 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 information retrieval methods that measure retrieval performance include metrics 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 information retrieval system where a relevant document is ranked as the one that first occurs after the first result. While Recall@k tells us if only if 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. The second challenge is diacritics. Diacritics are 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

2.2.1 Arabic RAG Studies

Several studies have evaluated components of Arabic RAG systems, though most focus on a single component — embedding models, LLM choice, or chunking — rather than systematically comparing retrieval strategies.

El-Beltagy and Abdallah (2024) compared semantic embedding models and LLMs for Arabic question answering using the Ar_EduText and ARCD datasets, finding that GPT-3.5 Turbo performed best on Ar_EduText while Mistral performed best on ARCD, with Llama 3 competitive on both. Abdelazim et al. (2023) evaluated ten multilingual embedding models for Arabic semantic retrieval on ARCD using Recall@k, reporting that Microsoft E5-Large achieved the strongest performance, exceeding 90% Recall@10. Both studies establish that embedding model choice materially affects Arabic retrieval and generation quality, but neither compares sparse retrieval or hybrid combinations against dense retrieval — their scope is limited to evaluating dense/embedding-based approaches against each other.

Alsubhi et al. (2025) conducted a more comprehensive evaluation, comparing chunking strategies, embedding models, reranking techniques, and LLMs on ARCD. Their results showed BGE-M3 and Multilingual-E5-large achieving the strongest embedding scores, with reranking (bge-reranker-v2-m3) improving retrieval quality and answer faithfulness, and Aya-8B achieving the best generation performance. This study is the closest in scope to the present research, as it evaluates reranking alongside retrieval components rather than in isolation. However, it does not include a sparse retrieval baseline, meaning it cannot establish how BM25-style lexical retrieval compares to the dense and reranked methods it evaluates — the comparison is confined to variations within embedding-based retrieval.

Allothman et al. (2026) developed an Arabic RAG pipeline combining multilingual-e5-large retrieval with NER-guided retrieval and an agentic query router, reporting an Acc@1 of 58.7% and MRR of 0.715 for the retriever, with Gemini-2-Flash achieving 0.83 answer relevance and 0.94 faithfulness. The authors note that evaluation relied primarily on RAGAS and a single prompt

template, which limited their ability to assess alternative retrieval and reasoning strategies. This study demonstrates that architectural additions (query routing, NER-guidance) can improve Arabic RAG performance, but it does not isolate the effect of the underlying retrieval strategy itself, since dense retrieval was the only retrieval method tested.

Collectively, these studies establish that embedding model selection, reranking, and architectural additions each 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 omits 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 Retrieval Methods Studies

Beyond Arabic-specific research, a broader body of work has compared sparse, dense, and hybrid retrieval strategies, primarily on English and multilingual benchmarks. This literature establishes the general trade-offs between these methods but does not evaluate them on Arabic text specifically.

Sparse versus dense retrieval. Arabzadeh et al. (2021) compared sparse and dense retrieval directly, finding 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 MS MARCO and Natural Questions. Together, these studies establish the semantic-versus-efficiency trade-off that motivates hybrid approaches, but both are evaluated exclusively on English benchmarks, leaving open whether the same trade-off holds for a morphologically rich language such as Arabic.

Hybrid retrieval. Several studies show hybrid retrieval consistently outperforming either sparse or dense retrieval alone. Lee et al. (2023) proposed a Ratio of Complementarity (RoC) objective to increase the complementarity between sparse and dense retrievers, outperforming traditional hybrid methods on MS MARCO, Natural Questions, and TREC Robust04. Sawarkar et al. (2024) proposed Blended RAG, combining sparse and dense retrieval with hybrid querying, and reported improved retrieval performance on Natural Questions, TREC-COVID, and SQuAD. Arivazhagan et al. (2023) combined sparse and dense retrieval through Hybrid Hierarchical Retrieval (HHR), reporting a 4.69% improvement in Recall@100 over a dense-only baseline across Natural Questions, Web Questions, and TriviaQA. These results consistently support hybrid retrieval as a stronger general-purpose approach, but again, none of these studies evaluate Arabic text, where lexical mismatch (the central weakness hybrid retrieval addresses) is expected to be more severe due to Arabic's morphological richness (see Section 2.1.4).

Domain-specific hybrid applications. Hybrid retrieval has also been evaluated in specialized domains, showing that its advantage depends on document and query characteristics. Akarsu et al. (2026) compared ten retrieval strategies spanning Sparse, Dense, Hybrid fusion, cross-encoder reranking, query expansion, and adaptive retrieval on a financial text-and-table benchmark, finding that hybrid retrieval combined with neural reranking achieved the best overall performance (Recall@5 = 0.816, MRR@3 = 0.605), outperforming single-stage methods. Bal and Puhan (2026) compared dense retrieval, hybrid BM25+dense retrieval, cross-encoder reranking, multi-query expansion, and MMR for biomedical RAG, finding hybrid retrieval provided 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.

2.2.3 Ranking Models

Moreira et al. (2024) evaluated several reranking models, including MiniLM, BGE, and Jina, on BEIR benchmark datasets such as Natural Questions, HotpotQA, and 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. Alike, 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 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

study	Methodology	Key Findings	Weaknesses/ Limitations
Exploring Retrieval Augmented Generation in Arabic	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
Optimizing RAG Pipelines for Arabic: A Systematic Analysis of Core Components	compare chunking strategies, multilingual and Arabic-specific embedding models, rerankers, and large language models.	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
Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)	evaluation of 10 multilingual embedding models using the ARCD dataset	Microsoft E5-Large achieved the best retrieval performance,	Limited dataset

		obtaining Recall@15 of 93%	
ARAG: An Agent-Based Hybrid Semantic-Lexical Retrieval-Augmented Generation Pipeline for Arabic Texts	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
Answering Students' Questions on Course Forums Using Multiple Chain-of-Thought Reasoning and Finetuning RAG-Enabled LLM	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.
From BM25 to Corrective RAG: Benchmarking Retrieval Strategies for Text-and-Table Documents	comprehensive comparison of retrieval techniques used in Retrieval-Augmented Generation (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.
Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection	The Author propose a classifier to choose the best retrieval model (sparse vs. dense vs. hybrid) for individual queries. Using standard Information Retrieval benchmark datasets. The adaptive method was	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

	compared with standalone dense and sparse retrieval methods based on retrieval effectiveness and computational efficiency.		semantic similarities when exact keywords are absent
RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering	interduced 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.
A Hybrid Retrieval and Reranking Framework for Evidence-Grounded Retrieval-Augmented Generation	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
Transformer Tafsir at QIAS 2025 Shared Task: Hybrid Retrieval-Augmented Generation for Islamic Knowledge Question Answering	Hybrid RAG using BM25 + Arabic dense retrieval + MiniLM cross-encoder reranking, followed by Fanar/Mistral generation.	improved evidence result accuracy 80% (Fanar) achived.	complexity of Islamic reasoning; these documents require a deep understanding
On Complementarity Objectives for Hybrid Retrieval	proposing a two-level orthogonality designed to improve RoC (Ratio of Complementarity) by encouraging them to representation diversity between sparse and dense retrievers	the model outperformed the traditional methods in all three benchmarks compared with the naive hybrid baseline improved MRR@10 by 0.78 on MS MARCO, MAP by 1.50 on NQ, and MAP by 2.21 on TREC Robust04	designing an objective that encourages the two retrieval methods to contribute distinct yet complementary evidence without sacrificing retrieval accuracy.
Benchmarking Retrieval Strategies for Biomedical Retrieval-Augmented Generation: A Controlled Empirical Study	Comparison five retrieval strategies Dense Retrieval, Hybrid BM25 + Dense, Cross-Encoder Reranking, Multi-Query Expansion,	Reranking (Cross-Encoder) performed best (composite = 0.827; contextual precision =	Data complexity needs a high-quality retrieval

	and MMR using biomedical benchmark datasets for Biomedical (RAG) and choose which methods provide the most accurate document retrieval	0.852), outperforming Dense Retrieval (0.822).	
Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers	develop a hybrid retrieval framework (Blended RAG), which combines dense and sparse retrieval techniques with hybrid querying	superiority of the hybrid lervative (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
Hybrid Hierarchical Retrieval for Open-Domain Question Answering	hierarchical retrieval framework that combines sparse, dense, and hybrid retrieval methods	(HHR) method outperformed the (DHR) method by 4.69%.	Limited dataset
Enhancing Q&A Text Retrieval with Ranking Models: Benchmarking, fine-tuning and deploying Rerankers for RAG	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
Efficient Title Reranker for Fast and Improved Knowledge-Intense NLP	enhancing Efficiency Title Reranker. through Broadcasting Query Encoder whith KILT Benchmark.	title reranking achieves a 20x-40x speedup rather than the vanilla passage reranke	May retrieve incorrect Documents and limited to reranking titles
mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval	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>

Ahmad, Muhammad & Ballout, Mohamad & 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., Almazrua, 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, J., Alahmadi, M. D., Alhusayni, A., Aldailami, I., Hamdine, I., Shabana, A., Iskandar, Y., & Khayyat, S. (2025). *Optimizing RAG pipelines for Arabic: A systematic analysis of core components*. arXiv. <https://doi.org/10.48550/arXiv.2506.06339>

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>

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. [10.18653/v1/2024.eacl-demo.16](https://doi.org/10.18653/v1/2024.eacl-demo.16)

- Irany, F. A., & Akwafuo, S. (2026). *A hybrid retrieval and reranking framework for evidence-grounded retrieval-augmented generation*. *arXiv*. <https://arxiv.org/abs/2605.01664>
- 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>
- Mahboub, A., Za'tar, M. E., Al-Rfooh, B., Estaitia, Y., Jaljuli, A., & Hakouz, A. (2024). *Evaluation of semantic search and its role in retrieved-augmented-generation (RAG) for Arabic language*. *arXiv*. <https://doi.org/10.48550/arXiv.2403.18350>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- Meftun Akarsu, 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>
- 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>
- Negar Arabzadeh, Xinyi Yan, and Charles L. A. Clarke. 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)*. Association for Computing Machinery, New York, NY, USA, 2862–2866. <https://doi.org/10.1145/3459637.3482159>
- 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., Shoeybi, 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>