

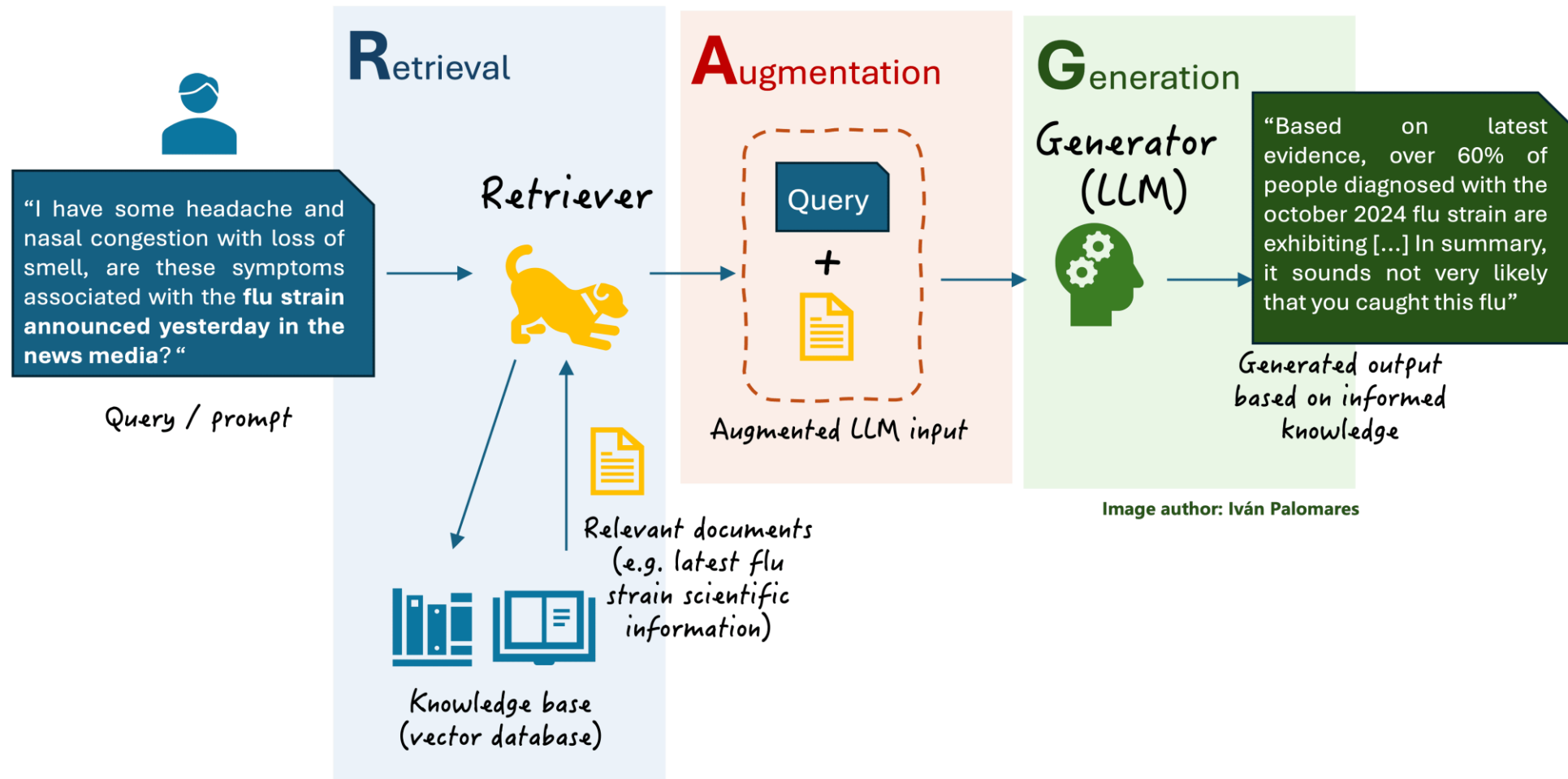
Comparative Evaluation of Retrieval Methods for Arabic Retrieval-Augmented Generation

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.

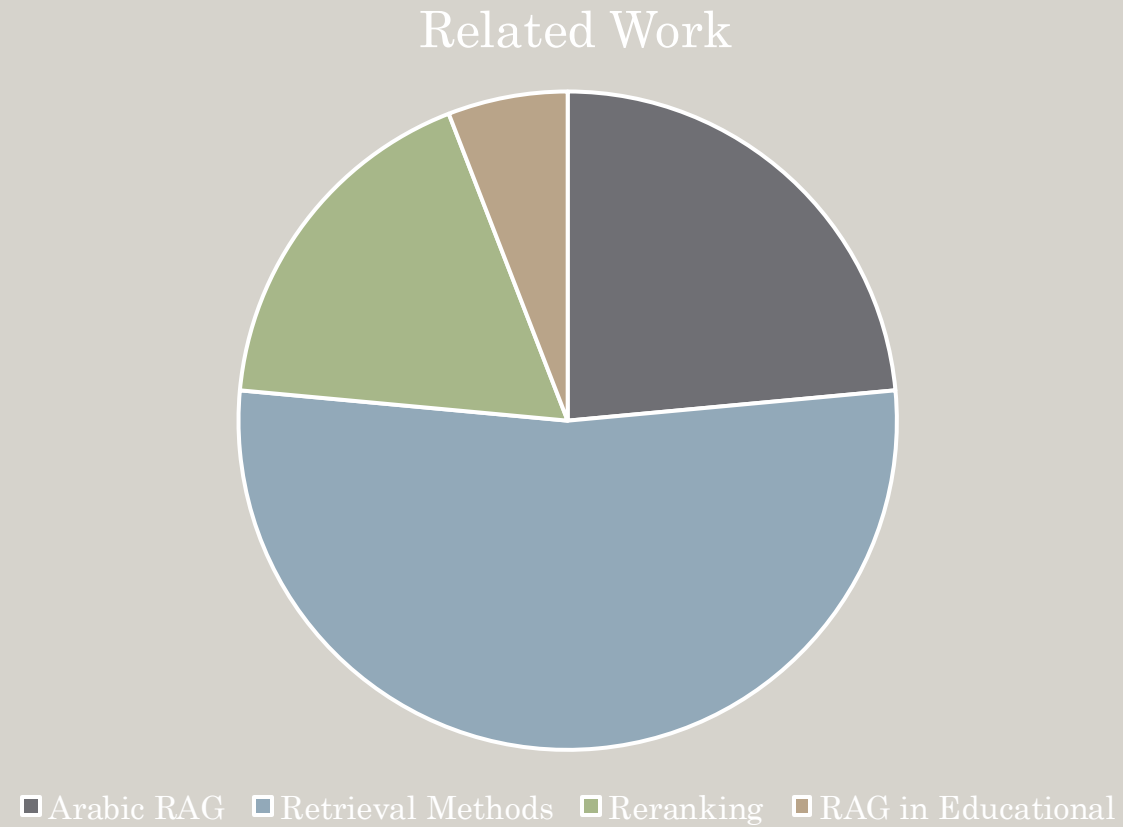
NDCG@k

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