

Comparing BM25, Multilingual-E5, and Reciprocal Rank Fusion for Retrieval-Augmented Generation

M. Sulthan Mahir Fadha ^{1*}, Yunita Ardila ^{2*}, Noor Wahyudi ^{3*}

* Information Systems, Faculty of Science and Technology, UIN Sunan Ampel Surabaya

mahirfadha@gmail.com ¹, yunita.ardilla@uinsa.ac.id ², n.wahyudi@uinsa.ac.id ³

Article Info

Article history:

Received 2026-09-18

Revised 2026-10-14

Accepted 2026-10-17

Keyword:

BM25,
Hybrid Retrieval,
Multilingual-E5,
Retrieval-Augmented
Generation,
Semantic Retrieval.

ABSTRACT

Retrieval-Augmented Generation (RAG) incorporates externally retrieved information as contextual knowledge to improve the quality of responses generated by language models. However, a RAG system's effectiveness depends heavily on the retrieval method's ability to identify and rank relevant documents. This study compares BM25, Multilingual-E5, and a hybrid approach based on Reciprocal Rank Fusion (RRF) using a product and service catalog corpus consisting of 145 documents and 40 benchmark queries. The evaluation was conducted in two stages: retrieval performance was assessed using Hit@1, Hit@3, Hit@5, and MRR@5, while RAG answer quality was evaluated using Faithfulness, Answer Relevancy, and Answer Accuracy. The results show that BM25 achieved the highest Hit@1 score (0.775) and MRR@5 (0.8279), whereas Multilingual-E5 achieved the highest Hit@5 score (0.950). For answer quality, Multilingual-E5 obtained the highest Faithfulness score (0.9667), while the hybrid RRF approach achieved the highest Answer Relevancy score (0.8451); both achieved an Answer Accuracy score (0.7938). Statistical tests revealed no significant differences among the retrieval methods. Answer Relevancy and Answer Accuracy were descriptively lower when the target document was absent from the Top-5 context, whereas the target's exact position within ranks 1-5 showed no significant correlation with answer quality. A context-size robustness analysis further showed descriptive improvements in Answer Relevancy and Answer Accuracy from Top-1 to Top-3 and Top-5, although none of the nine context-size effects remained statistically significant after Holm correction. Under a criterion emphasizing Top-5 target coverage and downstream answer quality, Multilingual-E5 showed a favorable operational profile under the evaluated configuration.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

The development of Large Language Models (LLMs) has accelerated the adoption of generative artificial intelligence in various knowledge-based and question-answering systems. Although LLMs can generate fluent natural-language responses, they remain limited in accessing knowledge not adequately encoded in their model parameters, dynamically incorporating updated information, and providing the underlying knowledge sources used to generate their responses. Retrieval-Augmented Generation (RAG) was introduced as an approach that combines the generative capabilities of language models with external knowledge

obtained through a retrieval process, allowing retrieved information to be used as contextual input for answer generation [1]. This approach has also been increasingly applied in various contexts in Indonesia. Pujiono et al. implemented RAG and a vector database for a chatbot in a public service agency [2], Hadie et al. employed RAG to improve the accessibility of the Satu Data Indonesia portal in local government [3], while Putro et al. developed an Indonesian-language RAG chatbot to support customer complaint handling in the energy sector [4].

The effectiveness of a RAG system depends not only on the generative model's capabilities but also on the quality of the information provided by the retrieval component. The

model then uses the retrieved documents as context to generate responses. Therefore, the selection and configuration of retrieval methods can affect overall system performance. Wu et al. reviewed RAG as a configurable pipeline in which design choices across its stages can affect system performance and efficiency [5]. A systematic review of the RAG literature by Hajar et al. also showed that retrieval effectiveness, the integration between retrieval and generation, and evaluation mechanisms remain important aspects of RAG system development [6].

One classical approach in information retrieval is lexical retrieval, which relies on term matching between queries and documents. BM25 remains a strong baseline; the BEIR benchmark demonstrated its robust performance across various domains, while the performance of neural and dense retrieval methods depends on the characteristics of the data and task [7]. These characteristics make BM25 relevant for catalog corpora containing explicit attributes such as brands, capacities, service types, and technical specifications. In contrast to lexical retrieval, dense semantic retrieval represents queries and documents as vectors, enabling retrieval based on semantic similarity. E5 was developed through contrastive pre-training, while Multilingual-E5 extends this approach to multilingual data [8], making it potentially useful when queries and documents convey similar meanings without using identical terms. The complementary characteristics of lexical and dense retrieval have motivated hybrid approaches that combine both methods. Zhao et al. reviewed hybrid retrieval approaches that combine sparse and dense retrievers to exploit complementary lexical and semantic signals [9], although the effectiveness of such approaches still depends on the domain and query characteristics.

In addition to evaluating whether relevant documents are successfully retrieved, RAG evaluation should also consider the quality of the generated answers after the retrieved context is provided to the generative model. Chen et al. showed that LLMs in RAG systems still face challenges when dealing with noisy information, the need to reject irrelevant information, the integration of multiple pieces of information, and conflicting contexts [10]. RAGAs subsequently introduced an evaluation framework that separates several dimensions of RAG systems, including the ability of the retrieval component to provide relevant context, the ability of the generator to produce responses supported by that context, and the overall quality of the generated answers [11]. This distinction matters because improvements in retrieval performance do not necessarily lead to proportional improvements in answer quality, especially when the relevant document is already available among multiple contexts provided to the generator.

Several studies in Indonesia have demonstrated the application of RAG in public-service chatbots, government data portals, and customer-service systems [2], [3], [4]. However, these studies primarily focus on the implementation of RAG-based applications rather than on controlled

comparisons of retrieval strategies and their downstream effects on answer generation. Meanwhile, broader information-retrieval studies have compared lexical and dense retrieval across general or open-domain benchmarks [7], [12], and recent RAG literature has highlighted the potential of hybrid retrieval approaches [9]. Nevertheless, the behavior of lexical, dense semantic, and rank-fusion retrieval remains less understood in compact, domain-specific catalogs containing products and services with a mixture of explicit attributes, descriptive information needs, and highly similar entities. In particular, it remains unclear whether improvements in retrieval rank directly translate into better RAG answers or whether the availability of the relevant document within the retrieved context is more important than its exact ranking position.

To address these gaps, this study provides four main contributions. First, it conducts a controlled comparison of BM25, Multilingual-E5, and hybrid Reciprocal Rank Fusion (RRF) using the same corpus, benchmark queries, and relevance judgments, allowing differences among lexical, dense semantic, and hybrid retrieval to be examined under identical conditions. Second, the evaluation connects retrieval-level performance with downstream RAG answer quality by combining Hit@K and MRR@5 with Faithfulness, Answer Relevancy, and Answer Accuracy under an identical generation configuration. Third, the study analyzes retrieval behavior across query categories and catalog domains and investigates whether the presence and rank position of the relevant document within the retrieved contexts are associated with answer quality. Fourth, the study evaluates the robustness of downstream answer quality to context size by comparing Top-1, Top-3, and Top-5 retrieved contexts under otherwise identical generation and evaluation configurations.

Accordingly, this study addresses the following research questions: RQ1: How do BM25, Multilingual-E5, and hybrid RRF differ in retrieval performance across the evaluated catalog corpus, query categories, and domains? RQ2: How do differences among the three retrieval methods affect RAG answer quality when the generation configuration is held constant? RQ3: How are the presence and rank position of the relevant document within the retrieved contexts associated with downstream answer quality? RQ4: How sensitive is RAG answer quality to the number of retrieved contexts when Top-1, Top-3, and Top-5 contexts are used under otherwise identical settings? Based on these questions, the study aims to characterize the retrieval behavior of the three approaches and identify their operational implications for RAG systems based on product and service catalogs under the evaluated configuration.

II. METHODS

This study employed a comparative experimental approach to evaluate three retrieval methods: lexical retrieval using BM25, dense semantic retrieval using Multilingual-E5, and hybrid retrieval using Reciprocal Rank Fusion (RRF). All three methods were evaluated using the same corpus, queries,

and relevance judgments to ensure a controlled comparison. The evaluation covered two levels: retrieval performance and answer quality in a Retrieval-Augmented Generation (RAG) system.

In general, the main experiment began with the construction of the corpus from a PostgreSQL database, followed by the development of 40 queries and their corresponding relevance judgments, retrieval using the three methods, ranking evaluation, construction of the Top-5 context for each method, answer generation using the same model and configuration, answer-quality evaluation, and statistical and comparative analyses. Figure 1 illustrates the workflow of the main Top-5 experiment. The additional Top-1, Top-3, and Top-5 context-size robustness experiment is described separately in Section II-G.



Figure 1. Main experimental workflow for comparing retrieval methods under the Top-5 RAG configuration

Figure 1 shows that BM25 and Multilingual-E5 generate rankings independently, which are subsequently combined to construct the hybrid RRF ranking. Each retrieval method is evaluated at the ranking level and is also used to construct the Top-5 context for the main RAG experiment. The additional context-size robustness experiment uses the same rankings and generation configuration but varies the number of supplied contexts, as described in Section II-G.

A. Corpus Construction

The data were obtained from a PostgreSQL database containing the product and service catalogs of a partner

company operating in the AC and CCTV sectors. Only active entities were included, and the dataset was frozen before the experiments to ensure that all retrieval methods were evaluated against an identical corpus. The final corpus consisted of 145 documents, comprising 53 AC products, 71 AC services, and 21 CCTV services, as summarized in Table I. Each product or service was represented as a single document without chunking.

TABLE I
COMPOSITION OF THE RESEARCH CORPUS

Data Type	Number of Documents
AC Products	53
AC Services	71
CCTV Services	21
Total	145

The corpus was primarily written in Indonesian, with code-mixing involving English brand names, technical terms, and industry specifications. Document labels and attribute templates were written in Indonesian, while English terms such as inverter, indoor, outdoor, channel, WDR, and product-brand names were retained in their original forms. This characteristic reflects the natural structure of commercial product and service catalogs in the evaluated domain.

Two deterministic textual representations were constructed for each entity. The `index_text` representation was used by both BM25 and Multilingual-E5 during retrieval, while `context_text` was used as input to the language model during RAG answer generation. For AC products, both representations were identical because the documents were sufficiently compact. For AC services, `index_text` retained the basic service fields together with selected sections describing duration, benefits, reasons for service, and need indicators, while lengthy procedural sections were omitted. In contrast, `context_text` preserved the complete service details for answer generation. For CCTV services, `index_text` contained the basic fields and cleaned service details without call-to-action content, whereas `context_text` additionally retained package descriptions and complete package specifications.

Document construction included HTML removal, HTML entity decoding, whitespace normalization, and numeric-format normalization. Database identifiers were retained only as metadata. No synonym expansion, query expansion, LLM-based summarization, or semantic enrichment was applied to the retrieval representation. Therefore, BM25 and Multilingual-E5 operated on the same `index_text` for each corresponding document, ensuring that differences in retrieval performance were attributable to the retrieval method rather than to differences in document content.

Document length was measured using the official tokenizer of `intfloat/multilingual-e5-base` on the actual model input, consisting of the `passage:` prefix followed by `index_text`, including special tokens and with truncation disabled. Across the 145 documents, the token length ranged from 36 to 477 tokens, with a mean of 257.72 tokens, a median of 298 tokens, and a standard deviation of 121.43 tokens. No document

exceeded the 512-token limit, resulting in zero truncation loss.

Because the corpus contained only 145 documents, dense retrieval was performed by exhaustively comparing each query against all document embeddings rather than using an approximate nearest-neighbor index. The corpus is intentionally domain-specific and relatively small; therefore, the results should be interpreted as characterizing retrieval behavior under the evaluated AC and CCTV product-and-service catalog setting rather than as evidence of general performance across large-scale or open-domain corpora.

B. Query Development and Relevance Judgments

The benchmark consisted of 40 manually constructed and human-curated queries derived from information available in the frozen corpus. All queries were designed by the first author to represent different information-seeking behaviors that may occur in commercial product and service catalogs. No query was generated automatically using an LLM. To provide balanced coverage of retrieval characteristics and difficulty, the benchmark was divided into four categories—exact, semantic, multi-constraint, and fine-grained—with 10 queries in each category, as summarized in Table II.

TABLE II
DISTRIBUTION OF BENCHMARK QUERIES

Category	Number of Queries
Exact	10
Semantic	10
Multi-constraint	10
Fine-grained	10
Total	40

Exact queries contained explicit lexical cues, such as product or service names and directly stated attributes. Semantic queries expressed user needs or symptoms without necessarily mentioning the target entity or using the same terminology as the corresponding document. Multi-constraint queries combined multiple technical or functional requirements that had to be satisfied simultaneously, whereas fine-grained queries were designed to distinguish a single target from other catalog entities with highly similar characteristics. By domain, the benchmark consisted of 12 AC product queries, 16 AC service queries, and 12 CCTV queries.

During benchmark construction, each candidate query underwent manual ambiguity and hard-negative auditing. Candidate queries that could validly refer to more than one catalog entity were revised by adding or refining constraints until a unique target could be identified from information explicitly contained in the corpus. Hard-negative documents were also identified for each query when appropriate. These documents were lexically or semantically similar to the target but failed to satisfy at least one required condition.

Relevance judgments followed a binary relevance scheme with exactly one ground-truth relevant document for each query. The ground-truth target was manually assigned by the

first author by matching the query requirements against the canonical corpus. A document was labeled relevant only if it satisfied all mandatory constraints expressed in the query. Documents that matched only a subset of the constraints were labeled non-relevant, including designated hard negatives. Consequently, each query had a single relevant document, such that $|R_q| = 1$.

Before the benchmark was frozen, the query–target mappings and hard-negative assignments were manually rechecked against the canonical corpus. This process was intended to reduce ambiguity and prevent retrieval results from influencing the definition of ground truth. Because the benchmark contains exactly one relevant document per query, Recall@K is numerically equivalent to Hit@K; this limitation is considered in the retrieval-evaluation section.

C. Retrieval Methods

1. Lexical Retrieval Using BM25

BM25 is a probabilistic method in information retrieval that estimates relevance based on the occurrence of query terms in documents while considering term frequency and document-length normalization [13]. The BM25 score for a document and a query is calculated using Equation (1).

$$BM25(D, Q) = \sum_{q_i \in Q} IDF(q_i) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b \frac{|D|}{avgdl})} \quad (1)$$

In Equation (1), $f(q_i, D)$ denotes the frequency of query term q_i in document D , $IDF(q_i)$ represents the inverse document frequency of q_i , $|D|$ represents the document length, and $avgdl$ denotes the average document length in the corpus. The parameter k_1 controls term-frequency saturation, while b controls the degree of document-length normalization [13]. This study used BM25Okapi from rank-bm25 version 0.2.2 with $k_1 = 1.5$, $b = 0.75$, and $\epsilon = 0.25$. No hyperparameter tuning was performed based on the benchmark results.

BM25 preprocessing included Unicode NFKC normalization, lowercase conversion, conversion of decimal commas to decimal points, and treating hyphens and forward slashes as token separators. Tokenization was performed using the following pattern:

$$[a-z0-9]+(?:\.[a-z0-9]+)*$$

No stemming, stopword removal, synonym expansion, intent boosting, brand boosting, or filtering based on product or service attributes was applied.

2. Dense Semantic Retrieval Using Multilingual-E5

The second approach employed intfloat/multilingual-e5-base. The E5 model family was developed through contrastive pre-training to produce text representations that can be used for various embedding and retrieval tasks [14], while Multilingual-E5 extends this approach to multilingual representations [8].

$$\text{sim}(q, d) = \frac{q \cdot d}{\|q\|_2 \|d\|_2} \quad (2)$$

Queries were prefixed with query: and documents with passage:, followed by mean pooling and L2 normalization. Cosine similarity was calculated using Equation (2). Each query was compared exhaustively against all 145 document embeddings without using approximate nearest neighbor (ANN) search or a vector database. Therefore, the resulting rankings were not affected by approximate index configurations.

3. Hybrid Retrieval Using Reciprocal Rank Fusion

The hybrid retrieval approach was constructed by combining the BM25 and Multilingual-E5 rankings using Reciprocal Rank Fusion (RRF). RRF combines multiple ranked lists based on document positions without requiring normalization of the raw scores produced by different retrievers [15].

$$\text{RRF}(d) = \sum_{m=1}^M \frac{1}{K + r_m(d)} \quad (3)$$

The RRF score for a document d was calculated using Equation (3). In Equation (3), $r_m(d)$ denotes the rank position of document d in the m -th retrieval method, M represents the number of ranked lists being fused, and K denotes the RRF constant [15]. This study used $K = 60$ and assigned equal weights of 1:1 to BM25 and Multilingual-E5. Fusion was performed over the complete ranking of all 145 documents. No reranker, oracle filter, service-type filter, capacity filter, or intent boosting was applied.

D. Retrieval Evaluation

Each query produced a ranked list of documents for each retrieval method. Evaluation was conducted using Hit@1, Hit@3, Hit@5, Recall@1, Recall@3, Recall@5, and MRR@5.

Hit@K was assigned a value of 1 if the relevant document appeared within the top- K results and 0 otherwise, as defined in Equation (4)

$$\text{Hit}@K(q) = \begin{cases} 1, & r_q \leq K \\ 0, & r_q > K \text{ or target not retrieved} \end{cases} \quad (4)$$

Because each query in the benchmark had exactly one relevant document under a binary relevance scheme, Recall@K was numerically equivalent to Hit@K in this study. This condition limits the benchmark design because Recall@K provides no information beyond Hit@K. Therefore, the evaluation of the relevant document's ranking position was complemented using MRR@5. Future work may extend the benchmark to include multiple relevant documents or graded relevance levels so that metrics such as Recall@K and nDCG can provide richer evaluation information.

To evaluate the ranking position of the relevant document, Mean Reciprocal Rank over the top five results (MRR@5) was used, as defined in Equation (5).

$$\text{MRR}@5 = \frac{1}{N} \sum_{i=1}^N \begin{cases} \frac{1}{r_i}, & r_i \leq 5 \\ 0, & r_i > 5 \end{cases} \quad (5)$$

where $N = 40$ represents the number of queries, and r_i denotes the rank position of the relevant document for the i -th query.

The resulting rankings were also analyzed by query category and domain to identify potential performance trends across different types of information needs.

E. Context Construction and RAG Answer Generation

For the main RAG experiment, the top five documents retrieved by each method were used as context for answer generation. The retrieved entities were supplied to the generator in their original retrieval order, from rank 1 to rank 5. While retrieval was performed using index_text, the generator received the corresponding context_text representation so that the complete product or service information available for each retrieved entity could be used during generation.

All generation settings were held constant across BM25, Multilingual-E5, and hybrid RRF to isolate the effect of the retrieval method. Gemini 3.6 Flash was used as the generator through the Google GenAI SDK. The thinking level was set to low, the maximum output length was set to 2048 tokens, and one response candidate was generated per inference. Temperature, top-p, top-k, and random seed were not explicitly overridden and therefore followed the model defaults. Automatic function calling was disabled so that generation relied only on the supplied prompt and retrieved contexts.

The same system instruction and user-prompt template were used for every query and retrieval method. The system instruction required the generator to answer only from the supplied context, satisfy all explicit user constraints, avoid combining information from different products or services, avoid fabricating unavailable attributes, and explicitly state when the available context was insufficient. Responses were required to be concise and written in Indonesian, and the model was instructed not to expose retrieval methods, rankings, scores, ground-truth information, or other experimental metadata.

The user prompt consisted of the retrieved context block followed by the benchmark query and an instruction to answer based on the provided information. The context block was constructed from the context_text of the Top-5 retrieved entities in rank order. For each generation instance, the model received only the query and retrieved contexts; it was not provided with the retrieval-method identity, retrieval scores, relevance judgments, or ground-truth target document. The exact prompt templates and generation configuration are available in the public repository accompanying this study.

With 40 benchmark queries and three retrieval methods, the main Top-5 experiment produced 120 generated answers. Apart from the retrieved contexts themselves, the generation model, prompt instructions, context ordering, and inference configuration were identical across all retrieval methods.

TABLE III
MAIN EXPERIMENTAL CONFIGURATION

Component	Configuration
Number of Documents	145
BM25	BM25Okapi
k_1	1.5
b	0.75
ϵ	0.25
Dense Retrieval	multilingual-e5-base
Similarity	Exact cosine similarity
Hybrid	RRF
RRF constant K	60
Fusion weights	1:1
Reranker	Not used
Generator context	Top-5 per method
Context representation	context_text
Context ordering	Retrieval rank 1-5
Generator	Gemini 3.6 Flash
Thinking level	Low
Sampling parameters	Model defaults
Max Output tokens	2048
Candidate count	1
Function calling	Disabled
Number of generated answers	120

F. RAG Answer Quality Evaluation

A total of 40 reference answers were manually constructed from the ground-truth relevant documents before evaluating the generated responses. The reference answers were prepared without consulting any generated answer to reduce evaluation leakage. For each benchmark query, the same reference answer was used to evaluate the outputs produced by BM25, Multilingual-E5, and hybrid RRF.

All 120 generated answers were evaluated automatically using `ragas==0.4.3` with three complementary metrics: Faithfulness, Answer Relevancy, and Answer Accuracy. The evaluation used Gemini 3.7 Flash as the LLM judge through InstructorLLM, while gemini-embedding-001 was used only for Answer Relevancy and was not involved in document retrieval. Thus, the answer-quality evaluation employed models that were separate from the retrieval methods being compared.

Faithfulness evaluated whether factual claims in the generated answer were supported by the retrieved contexts. Its evaluation inputs were the user query, generated response, and retrieved contexts. Answer Relevancy evaluated how directly the generated response addressed the original user query and used the query and generated response as inputs; the metric was configured with `strictness = 3` and used gemini-embedding-001 for semantic comparison. Answer Accuracy evaluated agreement between the generated response and the manually constructed reference answer using the user query,

generated response, and reference answer as inputs. Therefore, the three metrics assessed different aspects of the RAG pipeline: contextual grounding, responsiveness to the query, and agreement with the expected answer, respectively.

The metrics were executed independently for each query-method pair using the asynchronous `.ascore()` interface rather than a single batch evaluation. Intermediate scores were checkpointed to allow interrupted evaluations to resume without recomputing previously completed samples. This resulted in 360 evaluation tasks, consisting of 120 generated responses evaluated across three answer-quality dimensions.

G. Context-Size Robustness Analysis

To evaluate the sensitivity of downstream RAG performance to the number of retrieved contexts, an additional robustness experiment was conducted using Top-1, Top-3, and Top-5 contexts. For each query-retrieval-method pair, the Top-1 and Top-3 contexts were constructed as strict prefixes of the corresponding Top-5 retrieval ranking, thereby preserving both the retrieved documents and their original rank order.

All generation settings were kept identical to those used in the main experiment, including the generator model, system instruction, prompt template, context representation, context ordering, thinking level, maximum output length, and sampling configuration. All three context-size conditions, including Top-5, were regenerated as part of the dedicated robustness experiment rather than reusing the outputs of the main experiment. This design ensured that Top-1, Top-3, and Top-5 were compared within the same robustness run.

With 40 benchmark queries, three retrieval methods, and three context-size conditions, the robustness experiment produced 360 generated answers. Each answer was evaluated using the same reference answers and the same Faithfulness, Answer Relevancy, and Answer Accuracy configuration used in the main experiment. This resulted in 1,080 metric-evaluation tasks.

H. Statistical and Comparative Analysis

Because the same 40 benchmark queries were evaluated under all three retrieval methods, all inferential comparisons followed a paired repeated-measures design, with each benchmark query serving as the unit of analysis. Statistical significance was assessed at $\alpha = 0.05$.

Binary Hit@K outcomes were compared across the three retrieval methods using Cochran's Q test. Exact McNemar tests were used for pairwise follow-up comparisons. MRR@5 and the three answer-quality metrics Faithfulness, Answer Relevancy, and Answer Accuracy were analyzed using the Friedman test, followed by pairwise Wilcoxon signed-rank tests. Holm correction was applied to control the family-wise error rate across multiple pairwise comparisons [16].

Effect sizes were reported alongside significance tests because the benchmark contained only 40 queries. Kendall's W was used as the effect-size measure for Friedman tests, while rank-biserial correlation was used for pairwise

Wilcoxon comparisons. Additional analyses included descriptive statistics and stratification by query category and catalog domain. Spearman rank correlation was used to examine the association between the ground-truth document's exact rank position and each answer-quality metric. This analysis was restricted to query-method pairs in which the ground-truth document appeared within the Top-5 context, because an exact context position was undefined when the target was absent. Correlations were computed separately for each retrieval method, and Holm correction was applied across the nine method-metric correlation tests.

For the context-size robustness analysis, Top-1, Top-3, and Top-5 answer-quality scores were compared separately for each retrieval method and answer-quality metric using the Friedman test, with the benchmark query as the repeated-measures unit. Kendall's W was reported as the corresponding effect-size measure. Because nine method-metric omnibus tests were conducted, Holm correction was applied across these nine tests to control the family-wise error rate. Pairwise Wilcoxon signed-rank tests were planned as follow-up analyses only when the Holm-adjusted omnibus result remained statistically significant.

III. RESULTS AND DISCUSSION

A. Retrieval Evaluation Results

Retrieval evaluation was conducted on 40 queries using three methods: BM25, Multilingual-E5, and hybrid RRF. Performance was measured using Hit@1, Hit@3, Hit@5, and MRR@5. Because each query had exactly one relevant document, Recall@K was numerically identical to Hit@K and is therefore not reported separately in Table IV.

TABLE IV
OVERALL RETRIEVAL EVALUATION RESULTS

Method	Hit@1	Hit@3	Hit@5	MRR@5
BM25	0.775	0.875	0.925	0.8279
Multilingual-E5	0.650	0.900	0.950	0.7737
Hybrid RRF	0.725	0.900	0.925	0.8133

BM25 achieved the strongest early-ranking performance, with the highest Hit@1 and MRR@5 scores, whereas Multilingual-E5 achieved the highest top-5 coverage. Hybrid RRF generally performed between the two individual methods across most metrics. These results highlight the distinction between placing the target document at an early rank and ensuring that it remains available within the top-5 context. They also indicate that combining lexical and semantic retrieval does not necessarily produce the highest performance across all metrics [7], [9].

TABLE V
STATISTICAL TEST RESULTS FOR RETRIEVAL PERFORMANCE

Metric	Statistic	p - value
Hit@1	$Q = 3.1667$	0.2053
Hit@3	$Q = 0.2857$	0.8669

Hit@5	$Q = 0.6667$	0.7165
MRR@5	$\chi^2 = 2.8148$	0.2448

All p values were greater than 0.05, indicating no statistically significant differences in retrieval performance among the three methods. Kendall's $W = 0.0352$ for MRR@5 also indicated a small effect size, and all pairwise comparisons remained non-significant after Holm correction. Therefore, the differences reported in Table IV are interpreted as descriptive trends rather than statistically significant differences.

B. Retrieval Analysis by Query Category and Domain

To examine performance characteristics in greater detail, the evaluation was conducted across four query categories: exact, semantic, multi-constraint, and fine-grained. Table VI summarizes the results.

TABLE VI
RETRIEVAL PERFORMANCE BY QUERY CATEGORY

Category	Method	Hit@1	Hit@5	MRR@5
Exact	BM25	0.900	0.900	0.900
	E5	0.700	1.000	0.850
	Hybrid	0.800	0.900	0.850
Semantic	BM25	0.900	1.000	0.950
	E5	0.700	0.900	0.770
	Hybrid	0.800	1.000	0.853
Multi-constraint	BM25	0.800	0.900	0.833
	E5	0.900	1.000	0.933
	Hybrid	0.800	0.900	0.850
Fine-grained	BM25	0.500	0.900	0.628
	E5	0.300	0.900	0.542
	Hybrid	0.500	0.900	0.700

Across query categories, BM25 performed most strongly on exact and semantic queries, Multilingual-E5 performed best on multi-constraint queries, and hybrid RRF achieved the highest MRR@5 on fine-grained queries. These findings differ from those of Karpukhin et al., who reported that dense retrieval outperformed BM25 on several open-domain question-answering datasets [12]. This difference further suggests that the effectiveness of lexical and dense retrieval depends on the corpus characteristics and the information needs being evaluated. In contrast, E5's stronger performance on multi-constraint queries suggests the potential benefit of semantic representations when multiple conditions must be considered simultaneously.

The decrease in Hit@1 for fine-grained queries indicates that all three methods had greater difficulty distinguishing between highly similar entities. Hybrid RRF achieved the highest MRR@5 in this category. These category-level patterns are interpreted as descriptive trends.

Domain-based analysis also revealed different performance characteristics, as shown in Table VII.

TABLE VII
RETRIEVAL PERFORMANCE BY DOMAIN

Domain	Method	Hit@1	Hit@5	MRR@5
AC Products	BM25	0.917	1.000	0.944
	E5	0.583	1.000	0.764
	Hybrid	0.833	1.000	0.917
AC Services	BM25	0.625	0.812	0.693
	E5	0.688	0.938	0.768
	Hybrid	0.625	0.812	0.708
CCTV	BM25	0.833	1.000	0.892
	E5	0.667	0.917	0.792
	Hybrid	0.750	1.000	0.850

By domain, BM25 performed best on AC Products and CCTV, whereas Multilingual-E5 performed best on AC Services. This pattern suggests that explicit product attributes are well suited to lexical matching, while more descriptive service queries may benefit more from semantic representations. Differences in answer quality among the methods within each domain were not statistically significant after Holm correction.

C. RAG Answer Quality Evaluation Results

After retrieval, the top five documents retrieved by each method were used as context for answer generation. A total of 120 answers were evaluated using Faithfulness, Answer Relevancy, and Answer Accuracy. This multidimensional evaluation approach is consistent with recent RAG evaluation literature, which emphasizes assessing retrieval quality together with generation dimensions such as faithfulness, relevance, and correctness [17].

TABLE VIII
RAG ANSWER QUALITY EVALUATION RESULTS

Metric	BM25	Multilingual-E5	Hybrid RRF
Faithfulness	0.9380	0.9667	0.9647
Answer Relevancy	0.8418	0.8446	0.8451
Answer Accuracy	0.7813	0.7938	0.7938

Descriptively, Multilingual-E5 achieved the highest Faithfulness score (0.9667), while hybrid RRF achieved the highest Answer Relevancy score (0.8451). For Answer Accuracy, Multilingual-E5 and hybrid RRF obtained the same score of 0.7938. However, the differences among the methods were relatively small.

To determine whether these differences were systematic, a Friedman test was conducted, as shown in Table IX.

TABLE IX
FRIEDMAN TEST RESULTS FOR ANSWER QUALITY

Metric	χ^2	p - value	Kendall's W
Faithfulness	0.2456	0.8844	0.0031
Answer Relevancy	0.4000	0.8187	0.0050
Answer Accuracy	0.6667	0.7165	0.0083

The Friedman test produced p-values greater than 0.05 for all metrics, with very small Kendall's W values, indicating no statistically significant differences in answer quality among the three retrieval methods. Pairwise Wilcoxon signed-rank tests with Holm correction were also non-significant. The large number of identical scores across methods suggests that differences in retrieval performance do not necessarily translate into downstream differences in answer quality when the target document remains available within the top-5 contexts.

D. Relationship Between Retrieval Success and Answer Quality

To analyze the relationship between retrieval performance and the generation stage, the answers were grouped according to whether the relevant document was present within the top-5 retrieved contexts.

TABLE X
ANSWER QUALITY BASED ON TARGET PRESENCE IN THE TOP-5 CONTEXT

Method	Metric	Target in Top-5	Target not in Top-5	Difference
BM25	Faithfulness	0.9600	0.6667	+0.2933
	Answer Relevancy	0.8834	0.3277	+0.5557
	Answer Accuracy	0.8176	0.3333	+0.4842
Multilingual-E5	Faithfulness	0.9781	0.7500	+0.2281
	Answer Relevancy	0.8891	0.0000	+0.8891
	Answer Accuracy	0.8355	0.0000	+0.8355
Hybrid RRF	Faithfulness	0.9618	1.0000	-0.0382
	Answer Relevancy	0.8870	0.3277	+0.5593
	Answer Accuracy	0.8311	0.3333	+0.4977

Table X shows that Answer Relevancy and Answer Accuracy were descriptively higher when the target document was available within the top-5 contexts. This pattern was observed across all three methods. However, the Target OUT groups consisted of only three cases for BM25, two for E5, and three for Hybrid RRF; therefore, these results were not used as inferential evidence.

Faithfulness did not always follow the same pattern. For example, Hybrid RRF achieved a Faithfulness score of 1.0000 when the target document was not present. This does not

imply that the generated answers were more accurate, because Faithfulness measures whether the answer is supported by the provided context rather than whether the retrieved context itself is correct [11]. The importance of context quality was

also highlighted by Knollmeyer et al., who noted that overly long or loosely related contexts may distract the generator and degrade answer quality [18].

TABLE XI
SPEARMAN CORRELATION BETWEEN TARGET RANK AND ANSWER QUALITY WITHIN TOP-5

Method	Metric	<i>n</i>	ρ	<i>P</i>	<i>P</i> _{Holm}
BM25	Faithfulness	37	-.039	.818	1.000
	Answer Relevancy	37	-.089	.602	1.000
	Answer Accuracy	37	-.117	.491	1.000
E5	Faithfulness	38	.315	.054	.488
	Answer Relevancy	38	.029	.861	1.000
	Answer Accuracy	38	-.126	.450	1.000
Hybrid	Faithfulness	37	-.165	.329	1.000
	Answer Relevancy	37	-.188	.265	1.000
	Answer Accuracy	37	-.072	.674	1.000

To formally examine whether the exact position of the relevant document within the Top-5 context was associated with downstream answer quality, Spearman rank correlations were computed separately for each retrieval method. The analysis included 37 BM25, 38 Multilingual-E5, and 37 Hybrid RRF cases in which the target document was present within the Top-5 context. No significant correlations were observed for Faithfulness, Answer Relevancy, or Answer Accuracy after Holm correction. For BM25, Spearman's ρ ranged from -0.117 to -0.039; for Multilingual-E5, from -0.126 to 0.315, and for Hybrid RRF, from -0.188 to -0.072. The largest association was observed between target rank and Faithfulness for Multilingual-E5 ($\rho = 0.315$, $p = 0.054$, $p_{\text{Holm}} = 0.488$), but it was not statistically significant.

These results provide no statistical evidence that moving the relevant document among ranks 1–5 systematically changes answer quality under the evaluated configuration. In contrast, the descriptive comparison in Table X shows

substantially lower Answer Relevancy and Answer Accuracy when the target document is absent from the Top-5 context. However, because the Target OUT groups contain only three BM25 cases, two Multilingual-E5 cases, and three Hybrid RRF cases, the presence-absence comparison remains descriptive rather than inferential. Accordingly, the results suggest, rather than establish, that ensuring target inclusion within the Top-5 context may be more consequential than its exact position once retrieved.

E. Diagnostic Case Analysis

A case-level analysis was conducted to better understand how differences in retrieval affected answers to individual queries. Of the 40 queries, one failed across all three methods, three showed different top-5 retrieval patterns across methods, and 12 showed differences in Hit@1.

TABLE XII
EXAMPLES OF KEY DIAGNOSTIC CASES

Query	BM25	E5	Hybrid	Accuracy Bm25	Accuracy E5	Accuracy Hybrid
Q005	Not in Top-5	Rank 2	Not in Top-5	1.00	1.00	1.00
Q019	Rank 1	Not in Top-5	Rank 5	0.75	0.00	0.75
Q026	Not in Top-5	Rank 3	Not in Top-5	0.00	0.75	0.00
Q034	Not in Top-5	Not in Top-5	Not in Top-5	0.00	0.00	0.00

Table XII illustrates the relationship between retrieval outcomes and answer quality at the individual-query level.

For Q019, BM25 (rank 1) and Hybrid RRF (rank 5) both achieved an Accuracy score of 0.75, whereas E5, which failed

to retrieve the target document within the top-5 results, scored 0.00. The opposite pattern was observed for Q026, where only E5 retrieved the target document (rank 3) and achieved an Accuracy score of 0.75. Q034 represents a case in which all three methods failed to retrieve the target document and all produced an Accuracy score of 0.00. Q005, however, was an exception: BM25 and Hybrid RRF still produced accurate answers even though the primary target document was not present in the top-5 results, suggesting that alternative retrieved documents may sometimes provide sufficient information to generate a correct answer.

F. Context-size Robustness Analysis

To examine whether the downstream results were sensitive to the number of retrieved contexts, answer generation and evaluation were repeated using Top-1, Top-3, and Top-5 contexts for each retrieval method. Table XIII summarizes the mean answer-quality scores together with the corresponding Friedman tests. The Top-5 results in this robustness analysis were obtained from the dedicated context-size experiment and were therefore analyzed independently from the main Top-5 experiment reported in Table VIII.

TABLE XIII
CONTEXT-SIZE ROBUSTNESS ANALYSIS

Method	Metric	Top-1	Top-3	Top-5	χ^2	p	p_{Holm}	W
BM25	Faithfulness	0.9609	0.9604	0.9588	0.8333	0.6592	1.0000	0.0104
BM25	Answer Relevancy	0.7356	0.7965	0.8344	0.4060	0.8163	1.0000	0.0051
BM25	Answer Accuracy	0.6750	0.7563	0.7813	2.4615	0.2921	0.8762	0.0308
Multilingual-E5	Faithfulness	0.9718	0.9633	0.9480	4.5085	0.1050	0.5248	0.0564
Multilingual-E5	Answer Relevancy	0.5947	0.8056	0.8528	6.6715	0.0356	0.2491	0.0834
Multilingual-E5	Answer Accuracy	0.5563	0.7438	0.8000	9.8689	0.0072	0.0648	0.1234
Hybrid RRF	Faithfulness	0.9420	0.9476	0.9822	3.2069	0.2012	0.8048	0.0401
Hybrid RRF	Answer Relevancy	0.6850	0.8207	0.8446	7.4091	0.0246	0.1969	0.0926
Hybrid RRF	Answer Accuracy	0.6313	0.7688	0.7813	6.0000	0.0498	0.2987	0.0750

Descriptively, increasing the number of retrieved contexts was associated with higher Answer Relevancy and Answer Accuracy for all three retrieval methods. For BM25, Answer Relevancy increased from 0.7356 at Top-1 to 0.8344 at Top-5, while Answer Accuracy increased from 0.6750 to 0.7813. The largest descriptive changes were observed for Multilingual-E5, whose Answer Relevancy increased from 0.5947 to 0.8528 and Answer Accuracy from 0.5563 to 0.8000. Hybrid RRF showed a similar pattern, with Answer Relevancy increasing from 0.6850 to 0.8446 and Answer Accuracy from 0.6313 to 0.7813.

Faithfulness remained comparatively high across context sizes and did not exhibit the same monotonic pattern. BM25 remained nearly unchanged across Top-1, Top-3, and Top-5, while Multilingual-E5 showed a small descriptive decrease as additional contexts were included. In contrast, hybrid RRF showed higher Faithfulness at Top-5 than at Top-1.

Several context-size comparisons produced nominal Friedman p -values below 0.05, including Answer Relevancy and Answer Accuracy for Multilingual-E5 and hybrid RRF. However, after Holm correction across the nine method-metric omnibus tests, none remained statistically significant at $\alpha = 0.05$. The smallest adjusted p -value was observed for Multilingual-E5 Answer Accuracy ($p_{\text{Holm}} = 0.0648$). Kendall's W values ranged from 0.0051 to 0.1234, indicating small effect sizes under the evaluated configuration.

The descriptive pattern also indicates that much of the improvement occurred between Top-1 and Top-3, whereas the additional gain from Top-3 to Top-5 was generally smaller. This suggests diminishing returns from adding more retrieved contexts once sufficient relevant information becomes available. Nevertheless, because none of the context-size effects remained statistically significant after multiple-comparison correction, these differences should be interpreted as robustness trends rather than evidence of a definitive optimal context size.

G. Implications for Retrieval Method Selection

Overall, no single method showed statistically significant superiority across all evaluation metrics. BM25 achieved the strongest early-ranking performance, Multilingual-E5 provided the highest top-5 coverage and competitive downstream answer quality, while hybrid RRF showed descriptive advantages on fine-grained queries. For top-5 RAG, coverage appears important because failing to include the target document was descriptively associated with lower Answer Relevancy and Answer Accuracy. In contrast, the target document's exact position within the top-5 results did not show a significant relationship with answer quality.

The context-size robustness analysis further indicates that answer quality should not be considered independently of the amount of retrieved context. Answer Relevancy and Answer

Accuracy generally increased descriptively from Top-1 to Top-3 and Top-5, while Faithfulness remained comparatively high. Much of the descriptive improvement occurred between Top-1 and Top-3, with smaller additional gains from Top-3 to Top-5. However, none of these context-size effects remained statistically significant after Holm correction; therefore, the findings do not establish a universally optimal context size.

Under an operational criterion emphasizing Top-5 target coverage together with downstream answer quality, Multilingual-E5 showed a favorable profile in the evaluated setting because it achieved the highest Hit@5, the highest Faithfulness in the main Top-5 experiment, and tied for the highest Answer Accuracy. This should not be interpreted as statistically significant or universal superiority. BM25 remains relevant when early-ranking performance and explicit lexical attributes are prioritized, while hybrid RRF showed descriptive advantages for fine-grained retrieval.

H. Limitations

Several limitations should be considered when interpreting the findings of this study. First, the evaluation was conducted on a relatively compact corpus of 145 product and service documents obtained from a single partner company in the AC and CCTV domains. The corpus was primarily written in Indonesian with English technical terms and brand names. Consequently, the findings characterize retrieval behavior within this specific domain and should not be generalized directly to large-scale, open-domain, multilingual, or substantially different catalog collections.

Second, the benchmark contained 40 manually constructed queries, with exactly one relevant document assigned to each query under a binary relevance scheme. Query construction, ambiguity checking, hard-negative identification, and relevance judgments were performed by the first author. Although the query-target mappings were manually rechecked before the benchmark was frozen, no independent annotator or inter-annotator agreement measurement was available. The single-relevant-document design also makes Recall@K numerically identical to Hit@K and does not capture scenarios in which multiple documents may provide partially or fully relevant information.

Third, the experimental configuration evaluated one dense retrieval model, Multilingual-E5, and one fixed RRF configuration with $K = 60$ and equal fusion weights. No hyperparameter tuning, alternative dense embedding models, learned fusion, or reranking stage was included. The results therefore should not be interpreted as representing all lexical, dense, or hybrid retrieval approaches.

Fourth, downstream answer generation used a single generator model and a fixed prompting configuration. Sampling parameters followed the model defaults and no random seed was explicitly specified; therefore, some stochastic variation between generation runs may remain. The context-size robustness experiment mitigated this issue by regenerating all Top-1, Top-3, and Top-5 conditions within a

dedicated experiment, but the findings still reflect the behavior of the evaluated generator configuration.

Finally, answer quality was assessed automatically using RAGAS with a single LLM judge configuration rather than through independent human evaluation. Although Faithfulness, Answer Relevancy, and Answer Accuracy provide complementary perspectives, automated LLM-based evaluation may inherit model-specific biases and does not fully capture human judgments of usefulness, completeness, or preference. Future work should therefore validate the findings on larger and more diverse corpora, use independently annotated query sets with multiple relevant documents or graded relevance, compare additional retrieval and reranking approaches, and complement automatic metrics with human evaluation.

IV. CONCLUSION

This study compared BM25, Multilingual-E5, and hybrid RRF for retrieval and downstream RAG answer generation in a compact product-and-service catalog. BM25 achieved the strongest early-ranking performance, with the highest Hit@1 and MRR@5, whereas Multilingual-E5 achieved the highest Top-5 target coverage. Hybrid RRF showed descriptive advantages for fine-grained queries. However, no statistically significant differences were observed among the three methods for retrieval performance or main Top-5 answer quality.

The downstream analysis showed that Answer Relevancy and Answer Accuracy were descriptively lower when the relevant document was absent from the Top-5 context, whereas the target document's exact position within ranks 1–5 was not significantly correlated with answer quality. The context-size robustness analysis further showed descriptive improvements in Answer Relevancy and Answer Accuracy when moving from Top-1 to Top-3 and Top-5 contexts, with comparatively smaller additional gains from Top-3 to Top-5. Nevertheless, none of the nine context-size effects remained statistically significant after Holm correction. Thus, the results support interpreting context coverage and context size as relevant operational factors, but they do not establish a universally optimal retrieval method or context size.

Under an operational criterion emphasizing Top-5 target coverage and downstream answer quality, Multilingual-E5 showed a favorable profile in the evaluated configuration, while BM25 remained advantageous for early ranking and explicit lexical attributes and hybrid RRF showed useful descriptive behavior for fine-grained queries. Future work should extend the evaluation to larger and more diverse corpora, independently annotated query sets with multiple or graded relevance judgments, additional embedding and fusion configurations, reranking approaches, and human evaluation of generated answers.

DATA AND CODE AVAILABILITY

The benchmark data, source code, experimental configurations, and evaluation results used in this study are publicly available in the GitHub repository: <https://github.com/MahirFadha/rag-retrieval-benchmark.git>

The benchmark data were constructed from product and service catalog information publicly available on the official website of the partner company. The repository includes README documentation describing the data structure, dependencies, experimental configurations, and the steps required to reproduce the retrieval, RAG generation, evaluation, and statistical analysis processes.

REFERENCES

- [1] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [2] I. Pujiono, I. M. Agtyaputra, and Y. Ruldeviyani, "Implementing Retrieval-Augmented Generation And Vector Databases For Chatbots In Public Services Agencies Context," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 10, no. 1, pp. 216–223, 2024, doi: 10.33480/jitk.v10i1.5572.
- [3] A. N. Hadie, I. Tahyudin, and T. Hariguna, "Enhancing Accessibility in Local Government Data Portals via Retrieval-Augmented Generation: A Case Study on Satu Data Indonesia in Banyumas Regency," *J. Tek. Inform.*, vol. 6, no. 4, pp. 2420–2433, 2025, doi: 10.52436/1.jutif.2025.6.4.5153.
- [4] I. P. H. Putro, J. Antoni, M. K. Adhitya, N. A. Herawati, A. Purwarianti, and N. P. Utama, "Retrieval-Augmented Generation (RAG) Chatbot for Handling Customer Complaints in the Energy Sector," *J. Infomedia Tek. Inform. Multimedia, dan Jar.*, vol. 10, no. 2, pp. 105–111, 2025, doi: 10.30811/jim.v10i2.7169.
- [5] S. Wu *et al.*, "Retrieval-augmented generation for natural language processing: a survey," *Artif. Intell. Rev.*, vol. 59, p. 192, 2026, doi: 10.1007/s10462-026-11605-7.
- [6] M. R. Hajar, E. Utami, and A. H. Muhammad, "A Systematic Literature Review of Retrieval-Augmented Generation: Methods, Applications, and Future Research Directions," *J. Appl. Comput. Sci. Technol.*, vol. 6, no. 2, pp. 115–128, 2025, doi: 10.52158/jacost.v6i2.1170.
- [7] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, 2021.
- [8] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," *arXiv:2402.05672*, 2024, doi: 10.48550/arXiv.2402.05672.
- [9] P. Zhao *et al.*, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," *Data Sci. Eng.*, vol. 11, no. 1, pp. 1–29, 2026, doi: 10.1007/s41019-025-00335-5.
- [10] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 16, pp. 17754–17762, 2024, doi: 10.1609/aaai.v38i16.29728.
- [11] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, N. Aletras and O. De Clercq, Eds., St. Julians, Malta: Association for Computational Linguistics, 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.
- [12] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.
- [13] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Found. Trends Inf. Retr.*, vol. 4, no. 1–2, pp. 1–174, 2009, doi: 10.1561/15000000019.
- [14] L. Wang *et al.*, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *arXiv:2212.03533*, Dec. 2022, doi: 10.48550/arXiv.2212.03533.
- [15] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: Association for Computing Machinery, 2009, pp. 758–759. doi: 10.1145/1571941.1572114.
- [16] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979, Accessed: Sep. 06, 2026. [Online]. Available: <http://www.jstor.org/stable/4615733>
- [17] A. Brown, M. Roman, and B. Devereux, "A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges," *Big Data Cogn. Comput.*, vol. 9(12), no. 12, p. 320, 2025, doi: 10.3390/bdcc9120320.
- [18] S. Knollmeyer *et al.*, "Evaluating Retrieval Augmented Generation: A Comprehensive Review of Evaluation Dimensions, Question Types, and Application," *SN Comput. Sci.*, vol. 7, no. 5, p. 576, 2026, doi: 10.1007/s42979-026-05134-x.