JNEST EISSN: 2961-8738 ● PISSN: 2961-8916 ● DOI: 10.56741/jnest.v5i03.2329
Article Received 27 May 2026, Revised 31 July 2026

Multi-Stage Hybrid Retrieval with Neural Re-Ranking and Rule-Based Context Validation for Domain-Specific Retrieval-Augmented Generation

Suharyadi* and Irwansyah Saputra

***Abstract*—Retrieval-Augmented Generation (RAG) has emerged as a widely adopted approach for grounding responses generated by large language models (LLMs) in external document corpora. However, conventional single-stage retrieval pipelines frequently fail to surface the most relevant context at top-ranked positions and lack explicit mechanisms to verify evidence adequacy prior to generation. This paper proposes a multi-stage RAG pipeline integrating three sequential components: hybrid retrieval combining BM25 and dense vector search via Reciprocal Rank Fusion (RRF), neural re-ranking using a cross-encoder transformer model, and a rule-based context validation gate that classifies retrieved context into three categories OK, WEAK, or ABSTAIN before answer generation. The system is evaluated on a domain-specific administrative corpus from the Indonesian AID Scholarship (TIAS) program using 50 annotated queries across four question types. Results demonstrate that hybrid retrieval improves recall at rank 10 (Recall@10) from 0.6391 to 0.8097 compared with BM25 alone, while neural re-ranking further increases mean reciprocal rank at rank 10 (MRR@10) from 0.6702 to 0.8040 ($p = 0.0043$, Wilcoxon signed-rank test). The complete pipeline achieves token-level F1 scores of 0.4429 and 0.4999 for the 4-billion-parameter (4B) and 12-billion-parameter (12B) model configurations, respectively, with a faithfulness score of 1.00. In comparison, the LLM-only baselines obtain token-level F1 scores below 0.09. The context validation mechanism correctly triggers ABSTAIN on 90% of out-of-domain queries, confirming its utility as a reliability control layer. These findings demonstrate that multi-stage retrieval with explicit context validation substantially improves both ranking quality and answer reliability in domain-specific RAG deployments.**

***Index Terms*— BM25, Context Validation, Dense Retrieval, Hybrid Retrieval, Neural Re-Ranking, Retrieval-Augmented Generation, Question Answering**

I. Introduction

The proliferation of large language models (LLMs) has created new opportunities for automated question answering over organizational knowledge bases. Models built on transformer architectures can generate fluent and contextually coherent responses, yet their reliance on parametric knowledge introduces well-documented limitations. Information encoded in model weights during training is static, potentially outdated, and prone to factual hallucination when applied to domain-specific queries that fall outside the model's training distribution [1]–[2].

Retrieval-Augmented Generation (RAG) addresses this limitation by coupling a retrieval component with a generative model. Rather than relying solely on internal knowledge, a RAG system first retrieves relevant document passages from an external corpus and then conditions the generator on those passages to produce responses that are grounded and verifiable [3]. This paradigm has proven effective across open-domain question answering benchmarks as well as domain-specific information systems.

Despite these advantages, conventional RAG implementations carry two structural weaknesses that limit their reliability in high-stakes administrative settings. The first concerns retrieval quality. Single-stage retrieval, whether lexical or semantic, imposes a ceiling on the relevance of the context supplied to the generator. Lexical approaches such as BM25 [4] excel at exact-term matching but fail to capture semantic equivalence across paraphrases and synonyms. Dense retrieval methods based on bi-encoder architectures [5]–[6] improve semantic coverage but can overlook specific administrative terminology, numeric constraints, and formal vocabulary that appear verbatim in policy documents. Neither approach alone is sufficient for corpora that are simultaneously formal in structure and varied in natural language expression.

The second weakness concerns evidence verification. Most RAG pipelines pass retrieved context directly to the generator without any intermediate assessment of evidence quality. This creates a non-trivial risk of generating responses that are grounded in incomplete, ambiguous, or internally conflicting passages, a failure mode that carries serious consequences in policy and administrative domains where precision and accountability are required [7].

These limitations are illustrated concretely by the operational context motivating this study. The TIAS (The Indonesian AID Scholarship) program, administered by Indonesia's International Development Cooperation Fund (LDKPI), provides scholarships for civil servants from partner countries to pursue postgraduate study at Indonesian universities. The program's official booklet contains densely structured information covering eligibility conditions, age and language proficiency thresholds, lists of partner universities, available study programs, and administrative procedures, presented in a combination of narrative prose and tabular formats. The program has grown significantly in recent years, from 316 applicants in 2024 to 1,858 in 2026, creating substantial demand for automated, document-grounded information services that cannot be met through manual responses alone.

Applicant queries over such a corpus frequently combine semantic intent with domain-specific constraints. A question such as "what master programs are available at

Suharyadi and **Irwansyah Saputra** are with the Department of Computer Science, Universitas Nusa Mandiri, Jakarta, Indonesia.
Corresponding author: *14240001@nusamandiri.ac.id

© 2026 IISTR

a specific partner university" requires both semantic understanding and precise term matching against tabular content. This dual requirement exposes the limitations of any single-stage retrieval strategy and motivates a pipeline architecture that addresses both coverage and ranking quality in sequence.

To address these challenges, this paper proposes a three-stage pipeline that improves retrieval quality and answer reliability in a principled, incremental manner. In the first stage, hybrid retrieval combines BM25-based lexical retrieval and Facebook AI Similarity Search (FAISS)-based dense retrieval using Reciprocal Rank Fusion (RRF) to produce a candidate set with broader recall coverage than either retrieval method alone. In the second stage, neural re-ranking applies a cross-encoder transformer to rescore candidates and promote the most semantically relevant passages to the top positions. In the third stage, context validation applies a rule-based gate that evaluates evidence sufficiency and consistency, classifying the retrieved context as OK, WEAK, or ABSTAIN before generation proceeds.

Unlike prior RAG systems that rely primarily on learning-based context assessment or evaluate performance mainly through generative metrics, this study emphasizes a transparent rule-based validation gate and reports retrieval-level improvements using Recall@10, MRR@10, and Hit@10 in a semi-structured administrative corpus.

The contributions of this paper are as follows. First, a complete multi-stage RAG pipeline integrating hybrid retrieval, neural re-ranking, and context validation is proposed and evaluated end-to-end on a domain-specific administrative corpus. Second, empirical evidence is provided that each pipeline component contributes incrementally to retrieval quality, with statistical significance tested through Wilcoxon signed-rank tests across all pairwise comparisons. Third, a rule-based context validation mechanism is introduced that identifies insufficient or out-of-domain context and triggers transparent refusal responses rather than hallucinated answers. Fourth, a comparative analysis against LLM-only and single-stage retrieval baselines demonstrates the practical advantage of the proposed approach in terms of both answer quality and computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work covering lexical and dense retrieval, hybrid retrieval strategies, neural re-ranking, RAG architectures, and context validation mechanisms. Section 3 describes the proposed multi-stage pipeline in detail. Section 4 presents the experimental setup, including the dataset, evaluation metrics, and baseline configurations. Section 5 reports and discusses the experimental results. Section 6 concludes the paper and outlines directions for future work.

II. Related Work

Research relevant to this paper spans five interconnected areas: lexical and dense retrieval, hybrid retrieval and rank fusion, neural re-ranking, retrieval-augmented generation, and context validation mechanisms. This section reviews key developments in each area and identifies the gap that the proposed approach addresses.

A. Lexical and Dense Retrieval

Information retrieval research has long relied on term-based matching models. BM25, introduced by Robertson and Zaragoza [4], remains one of the most widely adopted lexical retrieval models. It scores documents based on term frequency and inverse document frequency while normalizing for document length, making it effective for queries that share explicit vocabulary with the target corpus. Its computational efficiency and strong baseline performance have kept it relevant in modern retrieval systems despite being a pre-neural method.

The emergence of transformer-based language models introduced a fundamentally different retrieval paradigm. Karpukhin [5] proposed Dense Passage Retrieval (DPR), demonstrating that dense vector representations learned by a bi-encoder architecture can outperform BM25 on open-domain question answering tasks, achieving a top-5 accuracy of 65.2% on standard benchmarks. Reimers and Gurevych [6] further contributed to this direction through Sentence-BERT, a framework that produces semantically meaningful sentence embeddings suitable for efficient similarity search. To support large-scale dense retrieval in practice, Johnson [8] and Douze [9] developed FAISS, a library enabling efficient nearest-neighbor search in high-dimensional vector spaces. Subsequent work by Xiong [10] and Izacard [11] refined dense retrieval training through contrastive learning and hard negative mining, further improving retrieval accuracy.

Despite these advances, dense retrieval methods have a known weakness in domain-specific settings where precise terminology, numeric values, or administrative vocabulary must be matched exactly. In such cases, BM25 often retains an advantage, motivating the combination of both approaches.

B. Hybrid Retrieval and Rank Fusion

The complementary strengths of lexical and dense retrieval have motivated considerable interest in hybrid approaches. Ma [12] demonstrated in a replication study that combining sparse and dense retrieval improves accuracy by approximately 3% over dense retrieval alone across several benchmarks, confirming the complementary effect of the two paradigms. A key technical challenge in hybrid retrieval is how to merge ranked lists from heterogeneous sources whose score distributions are incompatible. Cormack [13] proposed Reciprocal Rank Fusion (RRF), a score-free fusion method that aggregates rankings based on document position rather than raw scores. RRF has since proven consistently effective as a simple yet robust baseline for combining retrieval results from multiple systems.

C. Neural Re-Ranking

While hybrid retrieval improves candidate coverage, it does not guarantee that the most relevant documents appear at the top of the ranked list. Neural re-ranking addresses this by applying a more computationally intensive model to rescore a shortlist of candidates.

Nogueira and Cho [14] demonstrated that using Bidirectional Encoder Representations from Transformers

(BERT) as a cross-encoder re-ranker, which jointly encodes the query and passage for direct relevance scoring, improves Mean Reciprocal Rank (MRR) by up to 27% on passage-ranking benchmarks. Nogueira et al. [15] extended this approach through multi-stage pipelines combining multiple BERT-based rankers. Khattab and Zaharia [16] proposed ColBERT, a late-interaction model that balances the expressiveness of cross-encoders with the efficiency of bi-encoders, achieving an MRR at rank 10 (MRR@10) of approximately 0.36–0.38 on the Microsoft Machine Reading Comprehension (MS MARCO) passage-ranking dataset. More recently, Sun et al. [17] demonstrated that large language models can serve as re-ranking agents, with GPT-4-based ranking improving normalized Discounted Cumulative Gain (nDCG) by 2.7 points over supervised methods. Collectively, these studies establish re-ranking as an effective strategy for improving the relevance of top-ranked passages, particularly when the candidate set is produced by a less computationally intensive first-stage retriever.

D. Retrieval-Augmented Generation

Lewis [3] introduced the foundational RAG framework, integrating a non-parametric retrieval component with a parametric sequence-to-sequence generator to improve factual accuracy on knowledge-intensive NLP tasks. This work established the core principle that external retrieval can substantially reduce hallucination and support verifiable generation. Subsequent work has extended RAG in several directions. Izacard and Grave [18] proposed Fusion-in-Decoder (FiD), which processes multiple retrieved passages jointly in the decoder and achieves an exact match score of 46.5 on Natural Questions. Ram [19] explored in-context retrieval-augmented language modeling, showing that retrieval augmentation reduces perplexity across language modeling tasks without fine-tuning. Siriwardhana [20] investigated domain adaptation of RAG models and showed consistent accuracy improvements through domain-specific fine-tuning. Gao [7] provided a comprehensive survey of RAG techniques, categorizing advances in retrieval, augmentation, and generation strategies and identifying reliability as an open challenge. Suharyadi and Saputra [21] applied hybrid ensemble RAG to Indonesian legal document consultation, demonstrating the viability of hybrid retrieval in a formal, domain-specific setting closely related to the present work.

E. Context Validation in RAG Systems

A growing body of work recognizes that retrieval quality alone does not guarantee answer reliability. Yoran [22] showed that irrelevant retrieved passages can actively degrade model performance even when a strong retriever is used, establishing the need for context quality control. Yu [23] proposed Chain-of-Note, which instructs the model to generate reading notes for each retrieved passage before producing an answer, improving accuracy by 7.9% and reducing sensitivity to noisy context. Asai [24] introduced SELF-RAG, which trains a model to generate critique tokens that adaptively decide whether retrieval is needed and whether retrieved passages are relevant and supportive. Yan [25] proposed Corrective RAG (CRAG), which evaluates retrieved document quality and triggers corrective retrieval when the initial context is assessed as insufficient. Wang [26] developed REAR, a relevance-aware framework that incorporates retrieval quality signals directly into the generation process. Edge [27] explored graph-based RAG for query-focused summarization, adding structural reasoning over retrieved content.

These approaches collectively demonstrate the importance of verifying context quality before generation. However, methods such as SELF-RAG and CRAG rely on learning-based evaluation models that introduce additional training requirements and reduce system transparency. In administrative and policy-oriented deployments, interpretability and consistency of validation decisions are often as important as accuracy. This creates a practical need for deterministic, rule-based validation that can be audited and adjusted without retraining.

More recent RAG systems have introduced adaptive and structurally richer mechanisms for controlling retrieval and context use. Self-RAG trains the generative model to retrieve on demand and to produce reflection tokens for evaluating retrieved evidence and generated responses [24]. CRAG employs a retrieval evaluator that determines whether the initially retrieved documents are adequate and may trigger corrective retrieval actions when the evidence is insufficient [25]. Adaptive-RAG dynamically selects among no-retrieval, single-step retrieval, and iterative retrieval strategies according to predicted query complexity [28]. GraphRAG constructs an entity-based graph and community summaries to support global query-focused summarization over large document collections [27]. More recently, SetR reframes passage re-ranking as a set-selection problem, selecting passages that collectively satisfy the information requirements of complex multi-hop queries [29].

These approaches are not directly interchangeable with the proposed pipeline. Self-RAG and Adaptive-RAG require learned control components, CRAG assumes corrective retrieval actions that may extend beyond a closed corpus, GraphRAG is primarily designed for global sensemaking over large collections, and SetR focuses on multi-hop evidence coverage. In contrast, the present study investigates a locally executed pipeline for factual question answering over a compact administrative corpus, with a deterministic context-validation gate that requires no additional task-specific training and can be audited through explicit rules. Accordingly, the comparison in this study is conducted at the architectural and evaluation-design levels. Direct experimental benchmarking against these newer systems remains an important direction for future work.

F. Research Gap

As shown in Table I, the landscape of RAG research has evolved considerably, with several recent systems incorporating multiple pipeline components that overlap with the present work. However, a careful examination across three dimensions reveals a gap that none of the existing systems fully addresses.

The first dimension concerns the nature of context validation. Systems such as SELF-RAG [24], CRAG [25],

Chain-of-Note [23], and REAR [26] all integrate a form of context assessment within their pipelines. However, each of these systems relies on learning-based mechanisms to perform this assessment. SELF-RAG trains the generator to produce critique tokens that evaluate retrieval necessity and passage relevance. CRAG employs a trained document quality estimator that triggers corrective retrieval when the retrieved context is deemed insufficient. Chain-of-Note fine-tunes the generator to produce structured reading notes for each passage prior to answering. REAR integrates relevance-aware signals directly into the model training objective. These learning-based approaches introduce additional training requirements, specialized model components, and validation decisions that are difficult to audit or interpret in practice. In administrative and government-facing deployments, where the consistency and traceability of system decisions are often operational requirements rather than optional features, a deterministic rule-based validation mechanism represents a meaningful and practical alternative that none of the above systems provide.

The second-dimension concerns evaluation methodology. Prior systems with context validation report performance almost exclusively through generative metrics such as Exact Match and Accuracy on open-domain QA benchmarks. Retrieval-specific metrics such as Recall@k and MRR@k are rarely reported as primary evaluation targets, despite the fact that retrieval quality is the direct upstream determinant of generation quality in any RAG system. The absence of retrieval-level evaluation in these systems makes it difficult to isolate the individual contribution of each pipeline component, assess where improvements actually originate, and understand the conditions under which each component is most beneficial.

The third dimension concerns the evaluation domain. Existing systems are predominantly benchmarked on open-domain QA datasets such as Natural Questions and TriviaQA, which consist of factual queries over encyclopedic corpora with well-defined answer spans. Administrative document corpora differ substantially in structure, containing a mixture of formal narrative prose and tabular information, domain-restricted vocabulary, and numeric constraints that must be matched with precision. Whether multi-stage retrieval pipelines transfer effectively to such settings has not been systematically studied.

It is also worth distinguishing the proposed approach from two related but fundamentally different lines of work. Yoran [22] investigate the robustness of RAG systems to irrelevant context, demonstrating that noisy retrieval can actively degrade model performance. Their work is diagnostic in nature and does not propose a deployable pipeline with a proactive context gating mechanism. GraphRAG [27] introduces a graph-based retrieval architecture that constructs a hierarchical knowledge graph from the corpus through community detection and summarization. While effective for global query-focused summarization over large corpora, this approach requires substantial graph construction overhead and is not designed for factual question answering over compact, semi-structured administrative documents. Neither system incorporates neural re-ranking, and neither evaluates retrieval quality through ranking-specific metrics.

TABLE I. COMPARISON OF PRIOR RAG SYSTEMS

Recent methods provide learning-based critique, corrective retrieval, relevance-aware generation, or graph-based retrieval, whereas the proposed work combines hybrid retrieval, neural re-ranking, deterministic context validation, and retrieval-level evaluation for a compact administrative corpus.

Study	Year	Hybrid Retrieval	Re-Ranking	Context Validation	RAG	Metric	Score
Lewis et al. [3]	2020	No	No	No	Yes	EM / F1	SOTA
Nogueira and Cho [14]	2020	No	Yes	No	No	MRR@10	+27%
Cormack et al. [13]	2009	Yes	No	No	No	MAP	0.36
Suharyadi and Saputra [21]	2025	Yes	No	No	Yes	N/A	N/A
Karpukhin et al. [5]	2020	No	No	No	No	Recall	Top-5 accuracy: 65.2%
Izacard and Grave [18]	2021	Yes	No	No	Yes	EM	46.5 (NQ)
Asai et al. [24]	2023	No	No	Yes (learning-based self-reflection)	Yes	Multi-metric	N/A
Reimers & Gurevych [6]	2019	No	No	No	No	Correlation	rho 75.9
Nogueira et al. [15]	2019	No	Yes	No	No	MRR	N/A
Robertson & Zaragoza [4]	2009	No	No	No	No	N/A	N/A
Khattab & Zaharia [16]	2020	No	Yes	No	No	MRR@10	0.36 - 0.38
Xiong et al.[10]	2020	No	No	No	No	Accuracy	Comp. SOTA
Izacard et al. [11]	2022	No	No	No	No	Recall	Better than BM25
Ram et al. [19]	2023	Yes	No	No	Yes	Perplexity	Improvement
Yu et al. [23]	2024	No	No	Yes (learning-based context assessment)	Yes	Accuracy	+7.9
Yoran et al. [22]	2024	No	No	No (diagnostic analysis)	Yes	Accuracy	Decrease observed
Sun et al. [17]	2024	No	Yes	No	No	nDCG	+2.7
Siriwardhana et al. [20]	2023	Yes	No	No	Yes	Accuracy	Improvement
Wang et al. [26]	2024	No	No	Yes (learned relevance assessment)	Yes	Accuracy	Improvement

Study	Year	Hybrid Retrieval	Re-Ranking	Context Validation	RAG	Metric	Score
Yan et al. [25]	2024	No	No	Yes (retrieval evaluator and corrective actions)	Yes	Accuracy	Improvement
Edge et al. [27]	2025	No (graph-based retrieval)	No	No	Yes	Qualitative	Improvement
Ma et al. [12]	2021	Yes	No	No	No	Accuracy	+3%
Jeong et al. [28]	2024	No	No	No (Adaptive retrieval control)	Yes	Accuracy / Efficiency	Improvement over adaptive baselines
Lee et al. [29]	2025	No	Yes (set-wise passage selection)	No	Yes	Retrieval quality / Answer correctness	Outperforms re-ranking baselines

Abbreviations: EM, Exact Match; F1, F1 score; MAP, Mean Average Precision; MRR, Mean Reciprocal Rank; nDCG, normalized Discounted Cumulative Gain; NQ, Natural Questions; N/A, not available; SOTA, state of the art; and ρ, Spearman's rank-correlation coefficient.

III. Proposed Methodology

This section describes the proposed multi-stage RAG pipeline for domain-specific question answering over administrative documents. The pipeline is designed to address two structural limitations identified in conventional RAG systems: the inability of single-stage retrieval to consistently surface the most relevant context at top-ranked positions, and the absence of an explicit mechanism for verifying evidence adequacy prior to generation.

The proposed approach organizes the retrieval and generation process into three sequential stages. The first stage performs hybrid retrieval combining lexical and semantic search to maximize recall coverage. The second stage applies neural re-ranking to reorder candidates and promote the most relevant passages to the positions that directly influence generation. The third stage introduces a rule-based context validation gate that assesses evidence sufficiency and consistency before the generation stage is reached. The following subsections detail the system architecture, document ingestion and indexing process, each pipeline stage, and the answer generation procedure.

A. System Architecture Overview

The proposed system is a multi-stage RAG pipeline designed for document-grounded question answering over a domain-specific administrative corpus. The pipeline processes each user query through three sequential stages before generating a response. In the first stage, hybrid retrieval combines lexical and semantic search to produce a broad candidate set with high recall coverage. In the second stage, neural re-ranking rescores the candidate set using a cross-encoder model to promote the most semantically relevant passages to the top positions. In the third stage, a rule-based context validation gate evaluates the evidence quality of the top-ranked passages before the generation stage is reached. Fig. 1 illustrates the overall architecture.

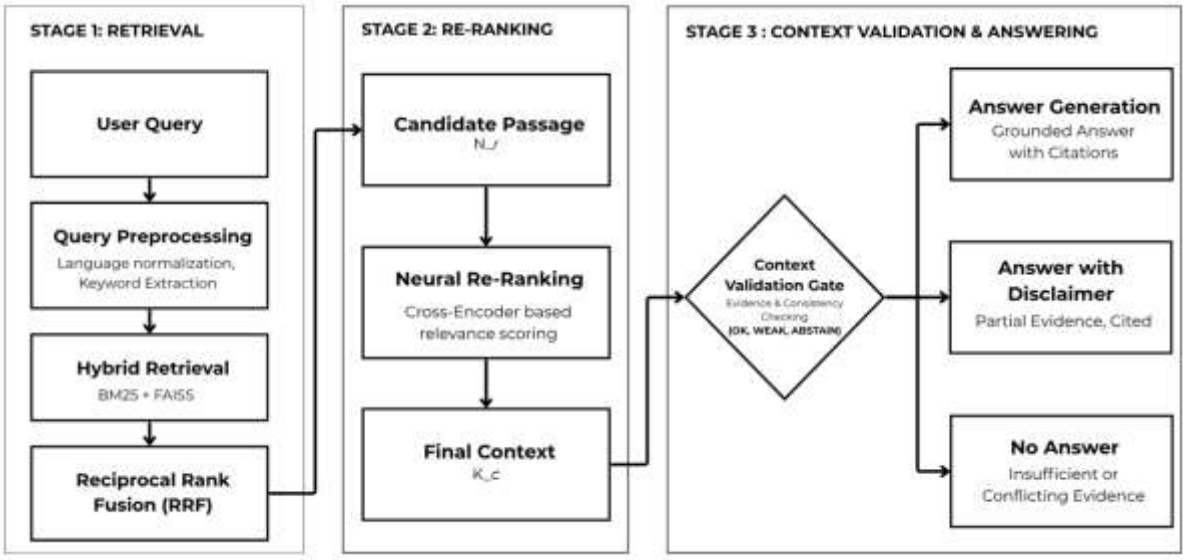


Fig. 1. Architecture of the proposed multi-stage RAG pipeline

BM25 and dense retrieval maximize candidate coverage, RRF fuses their ranks, the cross-encoder improves top-ranked relevance, and the rule-based gate labels the final context as OK, WEAK, or ABSTAIN before generation.

The pipeline operates in a stateless single-turn configuration, where each query is processed independently without maintaining conversational history. This design choice ensures that retrieval and generation

behavior is fully attributable to the pipeline components rather than accumulated conversational context, which is particularly important for controlled experimental evaluation.

B. Document Ingestion and Indexing

The knowledge base for this system consists of the official TIAS program booklet, a 77-page document containing program descriptions, eligibility requirements, partner university listings, available study programs, selection procedures, and administrative obligations. The document combines narrative prose sections with tabular content, a structural characteristic that directly motivates the hybrid retrieval approach adopted in this work.

Document preparation follows a structure-based chunking strategy, where boundaries are defined by section and subsection headings rather than fixed token counts. This approach preserves the semantic integrity of each chunk, ensuring that administrative rules, numeric constraints, and program listings are not split across chunk boundaries. Each chunk is assigned metadata including page number, section heading, subsection heading, content type (narrative or table row), and study level (S1 or S2), which is subsequently used by the context validation component.

The ingestion process produces 201 chunks from the source document. Of these, 130 chunks (64.7%) are narrative passages and 71 chunks (35.3%) are individual table rows derived from program listing tables. The average chunk length is 56.7 words, with variability reflecting the natural length difference between narrative paragraphs and table rows. No overlap is applied between adjacent chunks, as the structure-based segmentation already ensures that each chunk represents a logically self-contained unit of information.

Two separate indexes are constructed from the chunk collection. For lexical retrieval, each chunk is tokenized and indexed using BM25. For dense retrieval, each chunk is encoded into a 768-dimensional vector representation using the all-MiniLM-L6-v2 model from SentenceTransformers [6], chosen for its strong performance on multilingual semantic similarity tasks covering both Indonesian and English. The resulting vectors are indexed using FAISS [8]–[9] to support efficient nearest-neighbor search at query time.

C. Hybrid Retrieval with Reciprocal Rank Fusion

When a query is submitted, it undergoes lightweight preprocessing including language normalization and keyword extraction before being issued to both retrieval channels in parallel. The BM25 channel retrieves candidates based on term frequency and inverse document frequency weighting, following the probabilistic relevance framework of Robertson and Zaragoza [4]. BM25 assigns higher scores to documents containing query terms with high frequency relative to the corpus, while normalizing for document length to avoid bias toward longer passages. The dense retrieval channel encodes the query using the same embedding model used during indexing and retrieves the nearest neighbors in the FAISS index using cosine similarity.

The results from both channels are merged using Reciprocal Rank Fusion [13]. RRF assigns a score to each document based on its rank position in each ranked list rather than its raw relevance score, which avoids the need to normalize incompatible score distributions across retrieval methods. The RRF score for document d is computed as:

$$\mathrm{RRF}(d) = \sum_{j=1}^{M} \frac{1}{c_{\mathrm{RRF}} + \mathrm{rank}_j(d)} \quad (1)$$

where d denotes a candidate passage, M is the number of ranked lists being fused, and j indexes the retrieval channel. In the proposed system, $M = 2$, corresponding to the BM25 and FAISS ranked lists. The term $\mathrm{rank}_j(d)$ denotes the one-based rank position of passage d in ranked list j, while c_{RRF} is the rank-stabilization constant, which is set to 60 following established RRF practice [13]. A passage that does not occur in a particular ranked list receives no contribution from that list.

The top N_r passages from the fused ranking are forwarded to the neural re-ranking stage. In the implementation, N_r corresponds to the parameter TOP_N_RERANK and is set to 30 based on the parameter-sensitivity experiment reported in Section V-B.

D. Neural Re-Ranking

The hybrid retrieval stage is optimized for recall coverage, meaning its primary objective is to ensure that relevant passages are present somewhere in the candidate set. However, the ranking order produced by RRF is not guaranteed to place the most relevant passages at the top positions, which is critical in RAG systems where the generator operates on only a small number of top-ranked passages. Neural re-ranking addresses this by applying a more expressive relevance model to reorder the candidate set.

Each query passage pair from the N_r hybrid-retrieval candidates is independently evaluated using a cross-encoder transformer model. Unlike a bi-encoder, which encodes the query and passage separately, the cross-encoder processes the query and passage jointly as a single input sequence. This joint encoding allows token-level interactions between the query and passage to be considered when producing the relevance score. The cross-encoder model used in this study is ms-marco-MiniLM-L-6-v2, selected for its balance between passage-ranking accuracy and inference efficiency.

After all, N_r candidates have been rescored and reordered, the highest-ranked K_c passages are retained as the final context set for the validation and answer-generation stages. In the implementation, K_c corresponds to the parameter TOP_K and is set to 10 in the primary experiments. This setting reflects the practical assumption that most administrative queries in the evaluated corpus can be answered using a limited number of highly relevant passages.

E. Context Validation

Context validation is introduced as an intermediate control layer between re-ranking and generation. Its purpose is to evaluate whether the top-ranked passages

collectively provide sufficient, consistent, and relevant evidence to support a reliable answer before the generation stage is reached. This component is designed as a deterministic rule-based mechanism, ensuring that validation decisions are fully transparent, reproducible, and free from dependency on additional trained models.

The validation procedure operates on the top K_c passages from the re-ranked context set, where $K_c = 10$ in the primary experiments, and evaluates three aspects. The first is evidence sufficiency, which assesses whether the context contains explicit information that directly addresses the query. For numeric-type queries, this requires the presence of at least one numeric value in the context. For list-type queries, this requires either multiple retrieved passages or the presence of enumeration patterns. For general queries, sufficiency is assessed based on the presence of requirement-related keywords and minimum context length. The second aspect is consistency, which checks for potential conflicts between passages, such as contradictory numeric values or incompatible statements regarding the same topic. The third aspect is structural adequacy, which verifies that the retrieved set is not too sparse or too short to constitute a meaningful response basis.

Based on these assessments, each query-context pair is assigned one of three validation labels. A label of OK is assigned when the context provides sufficient evidence and contains no detected conflicts, indicating that the generator may proceed with full confidence. A label of WEAK is assigned when the context is present and partially relevant but falls below the evidence sufficiency threshold, indicating that the generator should proceed but qualify its response with an explicit acknowledgment of limited evidence. A label of ABSTAIN is assigned when the context fails minimum evidence criteria, contains unresolvable conflicts, or when no relevant passages are retrieved, indicating that the system should withhold a substantive response to prevent misinformation.

TABLE II. CONTEXT-VALIDATION DECISIONS
The rule-based gate permits generation for sufficient context, qualifies answers supported by partial evidence, and abstains when minimum evidence requirements are not met.

Label	Condition	System Action
OK	Sufficient evidence, no conflict	Generate grounded answer
WEAK	Partial evidence	Generate answer with disclaimer
ABSTAIN	Insufficient/out-of-domain/conflicting evidence	Refuse to answer

The decision procedure is formalized as follows:

Algorithm 1 uses descriptive variable names to improve readability. The variable query denotes the user question, while top_contexts contains the K_c highest-ranked passages produced by neural re-ranking and ordered by their relevance scores. The variable top_score represents the re-ranking score of the highest-ranked passage. For detailed validation, the procedure selects the first three passages as contexts_to_check. The Boolean variable evidence_found indicates whether explicit supporting evidence is available, whereas validation_issue_found indicates that the retrieved context is structurally insufficient for the query. The minimum_context_length parameter represents an empirically determined threshold for assessing contextual adequacy. The procedure returns a validation_label with one of three possible values: OK, WEAK, or ABSTAIN.

Algorithm 1. Rule-based Context Validation Procedure

```
Input:
query:
  user query to be evaluated
top_contexts:
  K_c highest-ranked passages produced by neural re-ranking
top_score:
  re-ranking score of the highest-ranked passage
minimum_context_length:
  context-length threshold

Output:
  validation_label: one of {OK, WEAK, ABSTAIN}

Procedure:

1. Initialize:
    evidence_found ← FALSE
    validation_issue_found ← FALSE

2. Check whether retrieved context is available:
    If top_contexts is empty:
      validation_label ← ABSTAIN
      Return validation_label

3. Select contexts for detailed validation:
    contexts_to_check ← first three passages from top_contexts

4. Evaluate evidence sufficiency:
    If query is a numeric-type query:

      If any numeric value is found in contexts_to_check:
        evidence_found ← TRUE

    Else if query is a list-type query:
      If number of passages in contexts_to_check ≥ 2
        OR a list-like pattern is found in contexts_to_check:
        evidence_found ← TRUE

    Else if requirement-related terms are found
        in contexts_to_check:

      evidence_found ← TRUE

    Else if length of the combined contexts_to_check
        ≥ minimum_context_length:

      evidence_found ← TRUE

    Else:

      evidence_found ← FALSE

5. Detect context insufficiency:
    If query is a list-type query
      AND number of passages in contexts_to_check < 2:
      validation_issue_found ← TRUE

    Else if the highest-ranked passage is shorter
        than minimum_context_length
        AND query requires a detailed answer:

      validation_issue_found ← TRUE
```

```
    Else:

        validation_issue_found ← FALSE

6. Assign validation label:
    If evidence_found = TRUE
      AND validation_issue_found = FALSE:
       If top_score > 0:
          validation_label ← OK
       Else:
          validation_label ← WEAK

    Else if evidence_found = TRUE
         AND validation_issue_found = TRUE:

        validation_label ← WEAK
    Else:
        validation_label ← ABSTAIN

7. Return validation_label
```

The thresholds applied in the validation procedure, including minimum_context_length, are determined empirically based on the observed distribution of context lengths and re-ranking scores in the evaluation dataset rather than through optimization on a separate training set. This approach preserves the deterministic and transparent nature of the validation mechanism. An OK label indicates that sufficient evidence is available without detected validation issues. A WEAK label indicates that some relevant evidence is available but the context may be incomplete or have a low top-ranked relevance score. An ABSTAIN label indicates that no adequate evidence is available to support a reliable answer.

F. Answer Generation

Queries that receive an OK or WEAK validation label proceed to the answer generation stage. The top K_c validated passages are concatenated with the user query and supplied as a structured prompt to the generative model, following the standard RAG prompting convention introduced by Lewis [3]. The model is instructed to generate a response grounded strictly in the provided context and to include citations indicating the source passages used. For WEAK-labeled queries, the prompt additionally instructs the model to acknowledge the limited nature of the available evidence within the generated response.

Queries that receive an ABSTAIN label do not proceed to generation. Instead, the system returns a fixed explanatory response indicating that the query falls outside the scope of the available document corpus or that insufficient evidence was found to support a reliable answer. This behavior is a deliberate design choice to prevent the generator from producing hallucinated responses when no adequate grounding is available.

Two model configurations are evaluated in this work: Gemma 3:4B and Gemma 3:12B, both executed locally using the Ollama runtime to ensure reproducibility in a controlled computational environment without dependency on external application programming interface (API) services.

IV. Experimental Setup

This section describes the experimental configuration used to evaluate the proposed pipeline. The evaluation is designed to assess the incremental contribution of each pipeline component across two dimensions: retrieval quality and answer generation quality. The following subsections detail the document corpus and annotated query dataset, the evaluation metrics applied at each stage, the system configurations used for comparison, and the implementation parameters adopted throughout the experiments.

A. Dataset and Corpus

The document corpus used in this study consists of the official TIAS program booklet, as described in Section Document Ingestion. The evaluation dataset comprises 50 annotated queries constructed to represent the range of information needs that prospective applicants are likely to express when consulting the program documentation. Queries are formulated in both English and Indonesian to reflect realistic usage patterns, and include paraphrase variations to test robustness across lexical surface forms.

The queries are distributed across four categories based on their structural complexity. Factoid queries require a single explicit answer retrievable from a specific passage, such as a numeric age limit or a language proficiency threshold. List queries require the system to enumerate multiple items, such as available study programs at a specific partner university. Constraint-based queries impose one or more conditions that must be satisfied simultaneously. Multi-condition queries combine multiple constraints across different attributes. The distribution is presented in Table III.

TABLE III. Distribution of the 50 Evaluation Queries
Factoid and list questions account for 88% of the dataset, while constraint-based and multi-condition questions are less represented.

Category	Count	Percentage
Factoid	29	58%
List	15	30%
Constraint-based	5	10%
Multi-Condition	1	2%
Total	**50**	**100%**

Ground truth for each query consists of two components: a set of relevant chunk identifiers indicating which passages in the corpus contain the information needed to answer the query, and a reference answer constructed directly from the source document. Ground truth annotation was performed by two annotators with domain knowledge of the TIAS program and administrative document structures. Each annotator first assessed relevance independently, after which disagreements were resolved through structured discussion to produce a consensus-based final annotation. The resulting ground truth is used as the reference for all retrieval and generation evaluation metrics.

The 50-query evaluation set is intended as a controlled, domain-specific test set rather than a large-scale benchmark. Its design prioritizes bilingual query coverage, passage-level relevance annotation, consensus-based reference answers, and representation of four

administrative question types. Consequently, the reported results should be interpreted as evidence of the behavior of individual pipeline components within the TIAS corpus, rather than as a general estimate of RAG performance across domains or document collections.

B. Evaluation Metrics

System performance is evaluated across two dimensions: retrieval quality and answer generation quality.

1) Retrieval Metrics

Three ranking-based metrics are used to evaluate retrieval performance, consistent with standard information retrieval evaluation practices [30]–[31]. Recall at rank k(Recall@k) measures the average proportion of relevant passages retrieved within the first k positions:

$$\text{Recall@}\,k = \frac{1}{Q}\sum_{i=1}^{Q}\frac{\left|R_i \cap D_i^{(k)}\right|}{|R_i|} \quad (2)$$

where (Q) is the total number of evaluation queries, (R_i) is the set of passages annotated as relevant to query (i), and $(D_i^{(k)})$ is the set of the top-(k) passages returned by the system for query (i). Recall@(k) therefore measures how much of the available relevant information is successfully retrieved within the specified ranking cutoff.

MRR@k measures the average reciprocal rank of the first relevant passage:

$$\text{MRR@}\,k = \frac{1}{Q}\sum_{i=1}^{Q}\begin{cases}\frac{1}{r_i}, & r_i \le k\\ 0, & r_i > k\end{cases} \quad (3)$$

where (r_i) is the rank position of the first relevant passage retrieved for query (i). If no relevant passage appears within the top-(k) results, the reciprocal-rank contribution of that query is zero. MRR@(k) is particularly informative in the RAG setting because the generative model primarily relies on the highest-ranked passages.

Hit@k indicates whether at least one relevant passage occurs within the top-(k) results:

$$\text{Hit@}\,k = \frac{1}{Q}\sum_{i=1}^{Q} I\left(R_i \cap D_i^{(k)} \neq \emptyset\right) \quad (4)$$

where $I(\cdot)$ is an indicator function that returns 1 when at least one relevant passage is found within the top-k results and 0 otherwise. Unlike Recall@k, which measures the proportion of all relevant passages retrieved, Hit@k represents retrieval success as a binary outcome for each query. All three-retrieval metrics are evaluated at $k = 10$ in the primary experiments.

2) Generation Metrics

Answer quality is evaluated using three metrics. Token-level F1 measures the token overlap between the generated answer and the reference answer, thereby capturing partial matches that Exact Match scoring would not identify:

$$F1_{\text{token}} = \frac{2P_{\text{token}}R_{\text{token}}}{P_{\text{token}} + R_{\text{token}}} \quad (5)$$

where $P_{\text{token}}$is the proportion of tokens in the generated answer that also occur in the reference answer, and $R_{\text{token}}$is the proportion of tokens in the reference answer recovered in the generated answer. The subscripts distinguish token-level precision and recall from the retrieval metrics used elsewhere in the manuscript.

Faithfulness is assessed through manual annotation and indicates whether all claims in the generated answer can be traced back to the provided context, without the introduction of information from the model's parametric knowledge. A faithfulness score of 1 indicates full grounding in the supplied context and 0 indicates the opposite.

Hallucination rate measures the proportion of generated answers containing at least one statement that cannot be attributed to the retrieved context. While faithfulness provides a holistic assessment, hallucination rate identifies the presence of specific unsupported claims at a more granular level.

3) Statistical Testing

To validate that observed performance differences between system configurations are not attributable to chance, pairwise comparisons are conducted using the Wilcoxon signed-rank test [32], a non-parametric test appropriate for comparing paired per-query metric scores without assuming a normal distribution.

C. System Configuration

Five system configurations are evaluated to assess the incremental contribution of each pipeline component:

1. Configuration 1

BM25 uses BM25 as the sole retrieval method without re-ranking or context validation. This configuration serves as the lexical retrieval baseline.

2. Configuration 2

Dense Retrieval uses FAISS-based dense retrieval as the sole method, representing the semantic retrieval baseline.

3. Configuration 3

Hybrid Retrieval combines BM25 and dense retrieval outputs using RRF, without re-ranking or context validation. This configuration isolates the contribution of fusion over single-stage retrieval.

4. Configuration 4

Hybrid Retrieval with Neural Re-Ranking extends Configuration 3 by applying cross-encoder re-ranking to the N_r candidates produced by hybrid retrieval. This configuration isolates the contribution of neural re-ranking over hybrid retrieval alone.

5. Configuration 5

Full Pipeline extends Configuration 4 with rule-based context validation prior to generation. This is the complete proposed system.

In addition, two LLM-only baselines using Gemma 3:4B and Gemma 3:12B without any retrieval are included to quantify the improvement attributable to retrieval augmentation over parametric generation alone.

D. Implementation Details

All experiments are conducted in a controlled local computing environment to ensure reproducibility. The system is implemented in Python. Dense retrieval uses the all-MiniLM-L6-v2 model from SentenceTransformers [6], which produces 768-dimensional embeddings and supports multilingual semantic similarity across Indonesian and English. Vector indexing and search are handled by FAISS [8]–[9]. Neural re-ranking uses the ms-marco-MiniLM-L-6-v2 cross-encoder model, selected for its balance between passage ranking accuracy and inference efficiency. Answer generation is performed by Gemma 3:4B and Gemma 3:12B, both executed locally through the Ollama runtime. The key parameter settings used across experiments are summarized in Table IV.

TABLE IV. EXPERIMENTAL PARAMETERS
Both retrievers return 50 candidates, 30 fused candidates are re-ranked, and the 10 highest-scoring passages form the final context.

Parameter	Value
BM25 candidate-pool size, K_{BM25}	50
Dense-retrieval candidate-pool size, K_{dense}	50
RRF stabilization constant, c_{RRF}	60
Re-ranking candidate-pool size, N_r(TOP_N_RERANK)	30
Final context size, K_c(TOP_K)	10
Embedding model	sentence-transformers/all-MiniLM-L6-v2
Re-ranking model	cross-encoder/ms-marco-MiniLM-L-6-v2
Generation models	Gemma 3:4B, Gemma 3:12B

The re-ranking candidate-pool size $N_r$is set to 30 based on an empirical sweep over $N_r \in \{20, 30, 40\}$. Increasing $N_r$from 20 to 30 produces measurable improvements across the retrieval metrics, whereas increasing it from 30 to 40 yields no further gain, indicating performance saturation. The parameter-sensitivity results are reported in Section V-B.

V. RESULT AND DISCUSSION

This section presents and interprets the experimental results across the two evaluation dimensions introduced in Section Experimental Setup. The analysis proceeds in stages that mirror the pipeline architecture, beginning with retrieval performance across all configurations, followed by the specific contribution of neural re-ranking, the behavior of the context validation mechanism, and end-to-end answer generation quality. A comparison against LLM-only baselines and a latency analysis across all system configurations are also reported.

A. Retrieval Performance Analysis

Table V presents retrieval performance for the four retrieval-stage configurations. The fifth configuration, which adds context validation, is evaluated separately in terms of validation behavior and answer generation quality because it does not alter the initial ranked retrieval output.

TABLE V. RETRIEVAL PERFORMANCE AT (k =10)
Hybrid retrieval improves coverage over BM25, while neural re-ranking yields the largest improvement in first-relevant-passage rank.

Configuration	Recall@10	MRR@10	Hit@10	Lat (s)
BM25	0.6391	0.5447	0.76	0.0001
Dense (FAISS)	0.7747	0.6473	0.90	0.0058
Hybrid	0.8097	0.6702	0.92	0.0063
Hybrid + Re-Ranking	0.8136	0.8040	0.94	0.2454

The results show a consistent and progressive improvement across all three metrics as pipeline components are added. BM25 alone achieves a Recall@10 of 0.6391, reflecting its well-known limitation in capturing semantic equivalence across query and document vocabulary. Dense retrieval improves Recall@10 to 0.7747 and MRR@10 to 0.6473, demonstrating the benefit of semantic representation for queries that do not share exact vocabulary with the target passages. Hybrid retrieval further improves Recall@10 to 0.8097 and MRR@10 to 0.6702, confirming the complementary effect of combining lexical and semantic retrieval channels. The addition of neural re-ranking produces the most significant jump in MRR@10, from 0.6702 to 0.8040, while Recall@10 remains largely unchanged at 0.8136. Fig. 2 visualizes these trends across all four configurations, showing the stepwise improvement from BM25 through to the full hybrid pipeline with neural re-ranking.

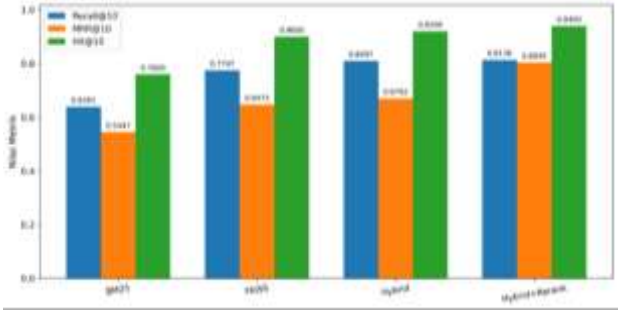


Fig. 2. Retrieval performance across the four pipeline configurations

Hybrid fusion provides the main Recall@10 gain over BM25, whereas neural re-ranking produces the largest MRR@10 improvement, from 0.6702 to 0.8040.

To assess whether these differences are statistically meaningful, pairwise Wilcoxon signed-rank tests are conducted. The results are presented in Table VI.

TABLE VI. PAIRED WILCOXON SIGNED-RANK TESTS
Hybrid retrieval significantly improves BM25 performance, and neural re-ranking significantly improves MRR@10 but not Recall@10.

Comparison	Metric	p-value	Significance
BM25 vs Hybrid	Recall@10	0.0005	Significant
BM25 vs Hybrid	MRR@10	0.0083	Significant
FAISS vs Hybrid	Recall@10	0.2858	Not Significant
FAISS vs Hybrid	MRR@10	0.7572	Not Significant
Hybrid vs Hybrid + Rerank	Recall@10	0.6465	Not Significant
Hybrid vs Hybrid + Rerank	MRR@10	0.0043	Significant

The statistical tests reveal two important patterns. First, the improvement from BM25 to hybrid retrieval is significant on both Recall@10 and MRR@10, confirming that the fusion of lexical and semantic retrieval produces a genuine and consistent gain over lexical retrieval alone. Second, the improvement from FAISS to hybrid retrieval is not statistically significant, suggesting that the primary

contribution of hybrid retrieval in this domain comes from adding lexical coverage on top of dense retrieval, rather than from combining two equally strong signals. Third, the addition of neural re-ranking produces a statistically significant improvement in MRR@10 but not in Recall@10, which is consistent with the role of cross-encoder re-ranking as a precision optimizer that improves the ranking order of already-retrieved candidates rather than expanding the set of relevant documents found.

To examine performance across query types, Table VII reports retrieval metrics broken down by query category for the full pipeline configuration.

TABLE VII. RETRIEVAL PERFORMANCE BY QUERY CATEGORY
Factoid and constraint-based queries achieve the strongest results, list queries show lower Recall@10 and MRR@10, and the multi-condition result is based on only one query.

Category	Count	Recall@10	MRR@10
Factoid	29	0.8465	0.8333
List	15	0.7088	0.7222
Constraint-based	5	0.9000	0.8400
Multi-condition	1	1.0000	1.0000

Factoid queries achieve strong performance on both metrics, consistent with the observation that explicit single-fact queries map well to specific passages in the corpus. Constraint-based queries perform comparably well, suggesting that the pipeline handles conditional filtering effectively when constraints are explicitly stated. List queries show a relative decline on both metrics, which reflects the inherent difficulty of retrieving all relevant table rows needed to construct a complete enumerated answer from a single query. The multi-condition result should be interpreted cautiously given that only one query of this type is present in the evaluation dataset.

B. Impact of Neural Re-Ranking

The most pronounced effect of neural re-ranking is on MRR@10, which increases by approximately 20 percentage points relative to hybrid retrieval alone, from 0.6702 to 0.8040. This improvement indicates that relevant documents, which were already present in the candidate set after hybrid retrieval, are now positioned substantially closer to the top of the ranked list. In the context of RAG, where the generator is conditioned on only a small number of top-ranked passages, this improvement in ranking precision has a direct downstream effect on generation quality.

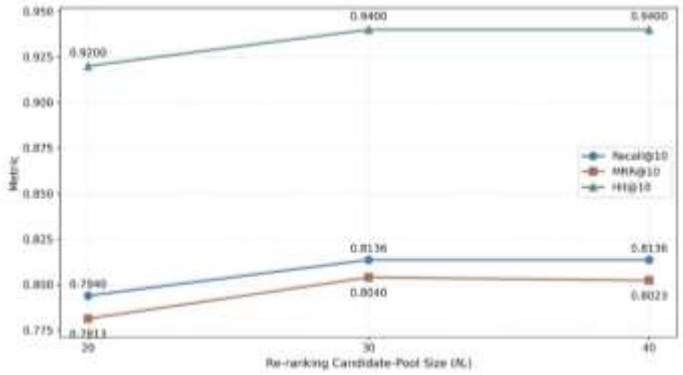


Fig. 3. Effect of the re-ranking candidate-pool size (N_r) on Recall@10, MRR@10, and Hit@10
Performance improves from 20 to 30 candidates and plateaus at 40, supporting the use of (N_r=30).

To examine the sensitivity of re-ranking performance to the candidate-pool size, an additional experiment evaluates N_r at 20, 30, and 40 passages. Increasing N_r from 20 to 30 produces measurable gains across the evaluated retrieval metrics, whereas increasing it from 30 to 40 produces no further improvement. This plateau indicates that the re-ranking model reaches a practical performance ceiling at approximately 30 candidates, possibly because lower-ranked candidates provide diminishing relevance differentiation. The result also shows that strong re-ranking performance can be achieved without incurring the additional scoring cost associated with a larger candidate pool.

The latency cost of neural re-ranking is approximately 0.25 seconds per query, which is a meaningful but acceptable overhead relative to the magnitude of the MRR improvement obtained. This trade-off is discussed further in Section Latency and Efficiency.

C. Context Validation Analysis

Fig. 4 shows the distribution of context validation labels across the 50 evaluation queries. Of these, 80% receive an OK label and 20% receive a WEAK label. No queries in the primary evaluation dataset receive an ABSTAIN label.

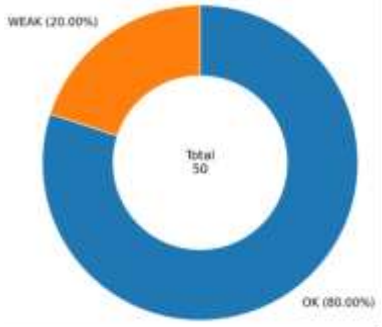


Fig. 4. Distribution of context-validation decisions for the 50 in-domain queries

40 were labeled OK, 10 WEAK, and none ABSTAIN. In a separate out-of-domain test, the validator correctly assigned ABSTAIN to 9 of 10 queries.

The absence of ABSTAIN cases in the primary evaluation is explained by two factors. First, all 50 queries are drawn from within the domain of the TIAS booklet corpus, meaning the system always retrieves some degree of relevant context. Second, the validation thresholds are calibrated to prioritize answer availability over conservative refusal when at least partial evidence is present. The ABSTAIN label is therefore reserved for genuinely out-of-domain queries or cases where retrieval returns no usable content.

To validate the ABSTAIN mechanism under out-of-domain conditions, a supplementary test is conducted using 10 queries explicitly constructed to fall outside the scope of the TIAS corpus, covering topics such as admission procedures at non-partner universities, other scholarship programs, and general knowledge questions. The results show that the system correctly assigns ABSTAIN to 9 of the 10 queries, yielding a 90% detection rate. The single false positive, where the system assigned an OK label to an out-of-domain query, is attributed to superficial lexical overlap between the query and passages

in the corpus. This case illustrates the known limitation of rule-based validation in distinguishing shallow keyword matches from genuine topical relevance, and motivates the consideration of model-based validation approaches in future work.

The WEAK category captures queries where context is retrieved but does not fully satisfy the evidence sufficiency criteria. These cases arise primarily from list-type queries where the retrieved passages provide partial rather than complete enumeration, and from queries that require synthesis across multiple passages with varying levels of specificity. In these cases, the system proceeds to generation with an explicit prompt instruction to acknowledge the limited evidence, resulting in responses that accurately reflect the uncertainty of the available context rather than overstating the completeness of the answer.

D. Answer Generation Quality

Table VIII presents end-to-end generation performance for four retrieval configurations, all using the Gemma 3:4B model to isolate the effect of retrieval quality on generation outcomes.

TABLE VIII. END-TO-END ANSWER GENERATION WITH GEMMA 3:4B
The complete pipeline attains the highest Token F1 while retaining full context faithfulness in this evaluation.

Configuration	Token F1	Faithfulness	Hallucination	Latency(s)
BM25 + LLM	0.3608	0.9800	0.0200	4.7025
FAISS + LLM	0.4358	0.9400	0.0600	3.0022
Hybrid + LLM	0.4155	1.0000	0.0200	3.6327
Hybrid + Re-ranking + Context Validation	0.4429	1.0000	0.0200	3.2033

The full pipeline achieves the highest Token F1 at 0.4429 and a faithfulness score of 1.00, indicating that all generated answers are fully grounded in the retrieved context. The FAISS configuration achieves a slightly higher Token F1 than the Hybrid configuration at 0.4358 versus 0.4155, a result that warrants interpretation. RRF combines results from both BM25 and dense channels, which can introduce passages with high lexical overlap but lower semantic density from the BM25 component. These passages are less informative for the generator, which may need to process a more heterogeneous context before extracting the relevant content. This trade-off is reflected in a marginally lower Token F1 for the Hybrid configuration compared to FAISS alone, but the full pipeline recovers and surpasses this figure through the filtering effect of re-ranking and context validation, which removes lower-quality candidates before they reach the generator.

Qualitative analysis of generation errors indicates that the majority of cases with low Token F1 originate from the generation stage rather than retrieval failure. For list-type queries in particular, the generator tends to produce summarized or partial enumerations rather than complete lists, even when the retrieved context contains all the required items. This suggests that the remaining performance gap is attributable to generation-side limitations rather than retrieval inadequacy.

E. Comparison with LLM-Only Baseline

Table IX presents a direct comparison between the full RAG pipeline and LLM-only baselines across both model sizes.

TABLE IX. LLM-ONLY VERSUS FULL-RAG PERFORMANCE
Retrieval augmentation substantially improves answer overlap and grounding for both model sizes.

Configuration	*Exact Match*	*Token F1*	*Faithfulness*	*Hallucination*	*Latency* (s)
LLM-only 4B	0.0000	0.0814	N/A	N/A	18.6504
LLM-only 12B	0.0000	0.0704	N/A	N/A	66.8392
RAG 4B	0.0000	0.4429	1.0000	0.0200	3.2033
RAG 12B	0.0200	0.4999	1.0000	0.0400	8.6971

The RAG pipeline outperforms the LLM-only baseline by a substantial margin across all evaluated configurations. The 4B RAG model achieves a Token F1 of 0.4429, representing more than a five-fold improvement over the LLM-only 4B baseline at 0.0814. The 12B RAG model achieves a Token F1 of 0.4999 and is the only configuration to produce any non-zero Exact Match score, albeit at a low rate of 0.02. These results confirm that the TIAS corpus contains highly domain-specific information that is not adequately represented in the parametric knowledge of either model, making retrieval augmentation an essential component rather than an optional enhancement.

The faithfulness and hallucination metrics cannot be computed for LLM-only configurations because there is no retrieved context against which to assess grounding. However, qualitative examination of LLM-only outputs shows that both models tend to produce generic responses based on general knowledge about scholarship programs rather than providing answers grounded in TIAS-specific policy details, which explains the very low Token F1 values.

An additional observation concerns end-to-end latency. Despite incorporating retrieval, re-ranking, and validation stages, the RAG pipeline is substantially faster than both LLM-only baselines. The RAG 4B configuration completes in 3.20 seconds compared to 18.65 seconds for LLM-only 4B, and RAG 12B completes in 8.70 seconds compared to 66.84 seconds for LLM-only 12B. This counterintuitive result is explained by the role of retrieved context in constraining the generation process. When the model receives a focused, relevant context, it can produce a concise and targeted response more efficiently than when it must generate an answer entirely from internal

knowledge over a broader reasoning space.

F. Latency and Efficiency Analysis

Fig 5 presents the per-stage latency breakdown across all pipeline components. The retrieval stages, both BM25 and FAISS, contribute negligible latency at 0.0001 and 0.0058 seconds respectively. The context validation stage adds only 0.0001 seconds, confirming that the rule-based mechanism imposes essentially no computational overhead. The neural re-ranking stage contributes approximately 0.25 seconds, which represents the primary additional cost introduced by the multi-stage pipeline relative to single-stage retrieval. The dominant component in end-to-end latency is the generation stage, accounting for 3.20 seconds with the 4B model and 8.70 seconds with the 12B model.

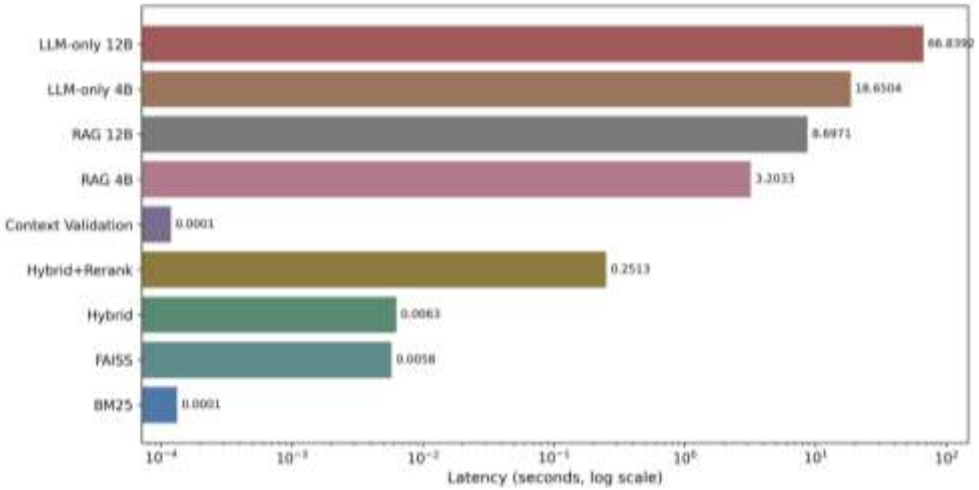


Fig. 5. Mean latency across retrieval stages and end-to-end configurations

Retrieval and rule-based context validation add negligible overhead, neural re-ranking adds approximately 0.25 s, and generation dominates total response time.

These measurements indicate that the performance gains achieved through hybrid retrieval and neural re-ranking come at a modest latency cost that is dwarfed by the generation latency. The practical implication is that optimizing the retrieval and re-ranking stages beyond what is already achieved would have a limited effect on overall system responsiveness, and that efforts to improve end-to-end latency should focus primarily on the generation stage, for example through model quantization or smaller specialized models.

Fig. 6 illustrates the accuracy-latency trade-off across configurations. The RAG 4B model occupies a favorable position in this space, offering strong generation quality at substantially lower latency than both the RAG 12B and both LLM-only configurations. The RAG 12B model provides a modest further improvement in Token F1 at the cost of approximately 2.7 times higher generation latency compared to RAG 4B. For deployment in interactive information services such as TIAS, where response time is a practical requirement, the 4B configuration represents the more suitable choice given its accuracy-efficiency balance.

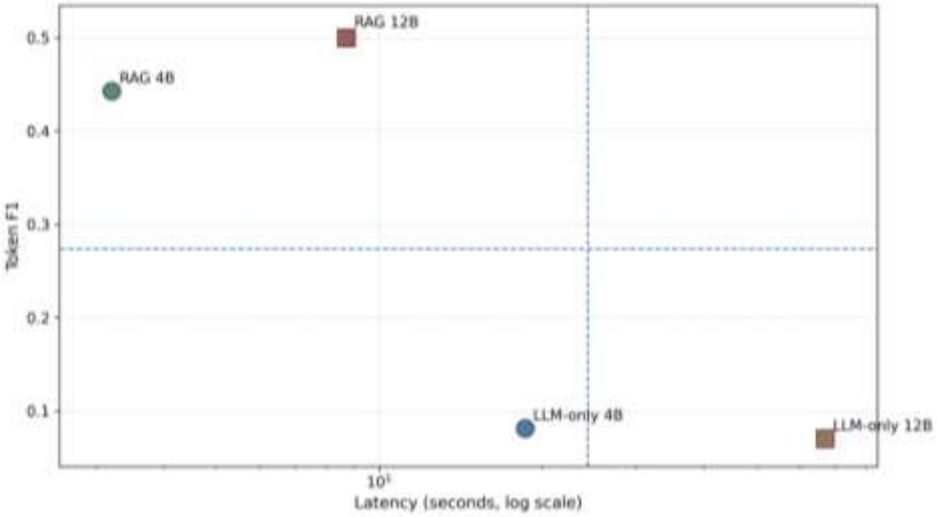


Fig. 6. Token F1–latency trade-off across RAG and LLM-only configurations

RAG 4B provides the strongest efficiency balance, whereas RAG 12B achieves the highest Token F1 at approximately 2.7 times the end-to-end latency of RAG 4B.

G. Limitations

This study has several limitations that should be considered when interpreting the results. First, the evaluation dataset consists of only 50 annotated queries derived from the TIAS administrative document corpus. Although the queries were designed to represent realistic information needs across factoid, list, constraint-based, and multi-condition question types, the dataset remains small compared with recent large-scale RAG evaluation benchmarks that contain substantially more diverse questions and grounding annotations [33]. The query distribution is also imbalanced: factoid questions account for 58% of the evaluation set, whereas the multi-condition category contains only one instance. These characteristics may contribute to the relatively strong retrieval and faithfulness results because factoid queries frequently correspond to explicit passages in the source document. The reported findings should therefore be interpreted as evidence of effectiveness within a controlled administrative corpus rather than as a general claim across document-grounded question-answering settings. Future evaluation should expand the number and balance of queries, include additional administrative documents, and incorporate independent annotation with inter-annotator agreement reporting.

Second, the proposed context validation mechanism is implemented using deterministic rule-based logic. This design provides transparency, reproducibility, and low computational overhead, which are desirable properties for administrative and government-facing deployments. However, rule-based validation may be less effective in identifying deeper semantic conflicts, implicit entailment relations, or cases where superficial lexical overlap masks true out-of-domain intent. This limitation is reflected in the supplementary out-of-domain test, where the system correctly assigned ABSTAIN to 9 out of 10 queries but produced one false positive due to shallow lexical or semantic overlap with the corpus. Future work may therefore incorporate Natural Language Inference (NLI), LLM-based verification, or hybrid validation mechanisms to improve the system's ability to distinguish genuine evidence adequacy from surface-level relevance.

Third, the evaluation of answer faithfulness relies on manual assessment against the retrieved context and reference answers. While this approach enables detailed inspection of whether generated claims are supported by the provided evidence, it may introduce confirmation bias because the evaluation process is closely tied to the same ground truth construction framework used in the study. The faithfulness score of 1.00 achieved by the full pipeline should therefore be interpreted as evidence of internal consistency between generated answers and retrieved context, rather than as an externally validated measure of factual correctness. The absence of independent annotators and inter-annotator agreement measures, such as Cohen's Kappa, limits the ability to quantify the objectivity and reproducibility of the human evaluation process.

Fourth, this study focuses on a single administrative document corpus, namely the official TIAS program booklet. The corpus has a semi-structured format consisting of narrative passages and tabular information, with 201 chunks produced through structure-based segmentation. Although this setting is representative of many administrative information systems, the findings may not directly generalize to larger, noisier, or more heterogeneous document collections. Further evaluation on other domains, such as legal, healthcare, education policy, or multi-document government archives, is required to assess the robustness and transferability of the proposed pipeline.

Fifth, the study does not perform an extensive comparison across alternative embedding models, re-ranking models, or generative models. The selected models were chosen based on practical considerations, including multilingual capability, availability, local execution feasibility, and computational efficiency. As a result, the reported performance should not be interpreted as the optimal configuration for all domain-specific RAG systems. Future work should include systematic ablation studies over different embedding models, cross-encoder architectures, retrieval fusion strategies, and prompt designs to identify more generalizable configuration patterns.

The study also does not include direct experimental comparisons with recent adaptive or graph-based RAG architectures such as Self-RAG, CRAG, Adaptive-RAG, GraphRAG, or SetR. These systems introduce substantially different assumptions, including task-specific model training, learned retrieval control, external corrective retrieval, knowledge-graph construction, or set-wise evidence selection. Implementing them within the present experiment would therefore require changes beyond a like-for-like replacement of the retrieval or validation component. The current comparison is consequently limited to architectural characteristics and reported results from the literature. Future work should evaluate comparable implementations under a shared corpus, query set, generative model, and evaluation protocol.

Finally, although latency is evaluated and discussed, the system remains a research prototype rather than a production-ready deployment. The experiments are conducted in a controlled local environment, and aspects such as concurrent user access, authentication, security hardening, infrastructure scalability, monitoring, and long-term maintenance are outside the scope of this study. Therefore, while the results demonstrate the practical feasibility of the proposed multi-stage RAG pipeline, further engineering and operational validation are required before deployment in a real-world public service environment.

VI. CONCLUSION

This paper proposed a multi-stage Retrieval-Augmented Generation pipeline that integrates hybrid retrieval, neural re-ranking, and rule-based context validation for domain-specific question answering over administrative scholarship documents. The proposed approach was motivated by two limitations commonly found in conventional RAG pipelines: the inability of single-stage retrieval to consistently surface the most relevant context at top-ranked positions, and the absence of an explicit

mechanism for assessing evidence adequacy before answer generation.

The experimental results demonstrate that combining BM25 and dense retrieval through Reciprocal Rank Fusion improves retrieval coverage compared to single-stage lexical retrieval. Hybrid retrieval increases Recall@10 from 0.6391 to 0.8097 compared to BM25 alone, confirming the complementary effect of lexical and semantic retrieval in a semi-structured administrative corpus. The addition of neural re-ranking further improves ranking quality, particularly in terms of MRR@10, which increases from 0.6702 to 0.8040. The Wilcoxon signed-rank test confirms that this improvement is statistically significant, indicating that neural re-ranking consistently promotes relevant passages to higher-ranked positions rather than merely expanding retrieval coverage.

Beyond retrieval performance, the proposed context validation mechanism contributes to system reliability by evaluating whether retrieved passages provide sufficient and consistent evidence before generation. The validation mechanism classifies retrieved context into OK, WEAK, or ABSTAIN categories, allowing the system to generate full answers when evidence is sufficient, provide qualified responses when evidence is partial, and withhold answers when the available context is inadequate. The supplementary out-of-domain evaluation shows that the system correctly triggers ABSTAIN in 90% of out-of-domain cases, demonstrating the practical value of context validation as a reliability control layer.

The end-to-end evaluation further confirms the advantage of retrieval augmentation over LLM-only baselines. The proposed RAG pipeline achieves substantially higher Token F1 scores than LLM-only configurations, while also producing grounded responses with high faithfulness. These results indicate that domain-specific information in administrative documents cannot be reliably captured through parametric model knowledge alone, and that retrieval-based grounding is essential for generating accurate and verifiable answers. In addition, the latency analysis shows that the additional cost introduced by hybrid retrieval, neural re-ranking, and rule-based validation remains modest compared to the overall generation latency, making the proposed pipeline practically feasible for interactive information service scenarios.

Overall, the findings suggest that reliable domain-specific RAG systems require more than a retrieval-and-generation architecture. Retrieval coverage, ranking precision, and evidence validation should be treated as complementary components within a unified pipeline. By integrating these components, the proposed approach improves not only retrieval performance but also the trustworthiness and operational reliability of generated answers. Although the study is limited to a single administrative corpus and a relatively small evaluation set, it provides empirical evidence that multi-stage retrieval with explicit context validation is a promising direction for building accountable, document-grounded question answering systems.

Future work should extend the evaluation to larger and more diverse document collections, incorporate independent human evaluation with inter-annotator agreement, and explore more advanced validation mechanisms such as Natural Language Inference or LLM-based verification. Further optimization of embedding models, re-ranking architectures, prompt strategies, and adaptive context selection may also improve the generalizability and efficiency of the proposed pipeline in broader real-world deployments.

REFERENCES

[1] A. Vaswani *et al.*, "Attention is all you need," in *31st Conf. Adv. Neural Inf. Processing Systems*, vol. 30, pp. 5998–6008, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *34th Int. Conf. Adv. Neural Inf. Processing Systems*, Curran Associates, Inc., pp. 9459–9474, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3495724.3496517

[4] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends® in Inf. Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[5] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. the 2020 Conf. Empirical Methods in Natural Lang. Processing (EMNLP)*, pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Lang. Processing and the 9th Int. Joint Conf. on Natural Lang. Processing (EMNLP-IJCNLP),* pp. 3982-3992, Aug. 2019. [Online]. Available: https://aclanthology.org/D19-1410/

[7] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *Comput. Lang.* Mar. 2024, doi: 10.48550/arXiv.2312.10997.

[8] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.

[9] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, and P.-E. Mazaré, "The Faiss library," *IEEE Trans. on Big Data,* vol. 12, no. 2, pp. 346-361, Oct. 2025, doi: 10.1109/TBDATA.2025.3618474.

[10] L. Xiong *et al.*, "Approximate nearest neighbor negative contrastive learning for dense text retrieval," *Inf. Retrieval,* Oct. 2020, doi: 10.48550/arXiv.2007.00808.

[11] G. Izacard *et al.*, "Unsupervised dense information retrieval with contrastive learning," *Inf. Retrieval,* Aug. 2022, doi: 10.48550/arXiv.2112.09118.

[12] X. Ma, K. Sun, R. Pradeep, and J. Lin, "A replication study of dense passage retriever," *Comput. Lang.,* Apr. 2021, doi: 10.48550/arXiv.2104.05740.

[13] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proc. the 32nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, pp. 758–759, Jul. 2009, doi: 10.1145/1571941.1572114.

[14] R. Nogueira and K. Cho, "Passage re-ranking with BERT," *Inf. Retrieval,* Apr. 2020, doi: 10.48550/arXiv.1901.04085.

[15] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-stage document ranking with BERT," Oct. 2019, doi: 10.48550/arXiv.1910.14424.

[16] O. Khattab and M. Zaharia, "ColBERT," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, pp. 39–48, Jul. 2020, doi: 10.1145/3397271.3401075.

[17] W. Sun *et al.*, "Is ChatGPT good at search? Investigating large language models as re-ranking agents," in *Proc. 2023 Conf. Empirical Methods in Natural Lang. Processing*, pp. 14918–14937, 2023, doi: 10.18653/v1/2023.emnlp-main.923.

[18] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguist.,* pp. 874–880, 2021, doi: 10.18653/v1/2021.eacl-main.74.

[19] O. Ram *et al.*, "In-context retrieval-augmented language models," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1316–1331, Nov. 2023, doi: 10.1162/tacl_a_00605.
[20] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of Retrieval Augmented Generation (RAG) models for open domain question answering," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1–17, Jan. 2023, doi: 10.1162/tacl_a_00530.
[21] Suharyadi and I. Saputra, "Hybrid ensemble retrieval-augmented generation for Indonesian legal consultation with keyword boosting," *J. Novel Eng. Sci. Technol.*, vol. 4, no. 02, pp. 71–85, Jul. 2025, doi: 10.56741/jnest.v4i02.1042.
[22] O. Yoran, T. Wolfson, O. Ram, and J. Berant, "Making retrieval-augmented language models robust to irrelevant context," *Comput. Lang.,* May 2024, doi: 10.48550/arXiv.2310.01558.
[23] W. Yu *et al.*, "Chain-of-note: Enhancing robustness in retrieval-augmented language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, 2024, doi: 10.18653/v1/2024.emnlp-main.813.
[24] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," *Comput. Lang.,* Oct. 2023, doi: 10.48550/arXiv.2310.11511.
[25] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *Comput. Lang.,* Oct. 2024, doi: 10.48550/arXiv.2401.15884.
[26] Y. Wang, R. Ren, J. Li, W. X. Zhao, J. Liu, and J.-R. Wen, "REAR: A relevance-aware retrieval-augmented framework for open-domain question answering," Nov. 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.321/
[27] D. Edge *et al.*, "From local to global: A graph RAG approach to query-focused summarization," *Comput. Lang.,* Feb. 2025, doi: 10.48550/arXiv.2404.16130.
[28] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. Park, "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," in *Proc. 2024 Conf. North Am. Chapter Assoc. Comput. Linguist.,* pp. 7036–7050, 2024, doi: 10.18653/v1/2024.naacl-long.389.
[29] D. Lee, Y. Jo, H. Park, and M. Lee, "Shifting from ranking to set selection for retrieval augmented generation," in *Proc. 63rd Annual Meeting Assoc. Comput. Linguist.,* pp. 17606–17619, 2025, doi: 10.18653/v1/2025.acl-long.861.
[30] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2009.
[31] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search.* New York: ACM Press, 2011.
[32] R. F. Woolson, "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*, Wiley, pp. 1–3, 2008, doi: 10.1002/9780471462422.eoct979.
[33] I. T. Sorodoc, L. F. R. Ribeiro, R. Blloshmi, C. Davis, and A. de Gispert, "GaRAGe: A benchmark with grounding annotations for RAG evaluation," in *Findings of the Association for Computational Linguistics,* pp. 17030–17049, 2025, doi: 10.18653/v1/2025.findings-acl.875.



Suharyadi is currently a Master's student in Computer Science at Universitas Nusa Mandiri, Jakarta. He has been working as an Information and Communication Technology Specialist at the Ministry of Finance, Republic of Indonesia, since 2010. His research interests include information retrieval, machine learning, deep learning, and technology innovation in public sector institutions. (email: 14240001@nusamandiri.ac.id).



Irwansyah Saputra is a lecturer at the Faculty of Computer Science, Universitas Nusa Mandiri, Jakarta. He obtained his Doctoral degree from IPB University (Institut Pertanian Bogor), Indonesia, with a research focus in blockchain technology. His academic interests include distributed systems, blockchain applications, cybersecurity, and emerging technologies in computer science. He has published several works related to blockchain and continues to contribute to research and development in secure and decentralized computing systems. (email: irwansyah.iys@nusamandiri.ac.id).