# Improving Retrieval-Augmented Generation Performance Using the MAF-RAG Architecture, EVR–VOR Vector Retrieval, and Multi-Agent Fallback Reasoning

**Erlanda Galant Prasetio [1*], L. Budi Handoko [2*], Khafiizh Hastuti [3*]**
[*] Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Jawa Tengah, Indonesia
111202214194@mhs.dinus.ac.id [1] , handoko@dsn.dinus.ac.id [2], afis@dsn.dinus.ac.id [3]

**Article Info**

***Article history:***

Received 2025-11-10
Revised 2025-12-30
Accepted 2026-01-07

***Keyword:***

*LLM,*
*Multi-Agent Reasoning,*
*Multi-Phase Retrieval,*
*RAG Architecture,*
*Vector Search.*

**ABSTRACT**

Retrieval-Augmented Generation (RAG) AI chatbots have gained popularity for their effectiveness in producing accurate, fast, and reliable responses; however, they have faced critical challenges stemming from limited datasets, outdated documents, and noisy, unfiltered data. This study proposes a Multi-Agent Fallback in Retrieval Augmented Generation (MAF-RAG). This robust RAG system testing pipeline integrates three-phase retrieval, filtering, and re-ranking data, along with a multi-agent debating process to address these challenges. This study demonstrates MAF-RAG's ability to perform under a constrained dataset, using a near-deployment dataset of 1,100 real-world documents. The pipeline utilizes 150 testing queries, carefully selected to reflect real-world RAG-based chatbot scenarios. A sentence-transformers/all-MiniLM-L6-v encoder encodes various chunks of documents into a 384-dimensional query vector embedding, ensuring an accurate relationship between testing queries and vectorized documents. The results show that the proposed MAF-RAG significantly outperforms the baseline system, achieving a mean F1-score of 0.556, an improvement of 18.8% over the Enhanced Baseline (mean F1-score = 0.469) and a 70.0% improvement over the Legacy Baseline (mean F1-score = 0.327). MAF-RAG also achieves the highest success rate, with 78% of the queries, while other baseline systems manage only 34% and 62%, respectively. MAF-RAG also reduces the failure rate by 42.1%, significantly increasing system reliability. Although MAF-RAG exhibits an increase in latency of 4.9%, these trade-offs are outweighed by the significant improvements in system reliability and performance. These findings highlight the contribution of this study: by implementing a robust retrieval testing pipeline, system accuracy can be improved, reducing the presence of noisy and unfiltered documents, and increasing system performance even when faced with challenging and varied datasets, making it a suitable solution for a RAG-based chatbot system that faces dataset challenges.

## I. INTRODUCTION

Retrieval Augmented Generation (RAG) based AI chatbots are currently growing in popularity across various fields[1]. Different sectors, including government, education, healthcare, and individuals, utilize chatbots for 24/7 customer support, answering FAQs, and even as ordering assistants[2].

This widespread adoption of chatbot services may stem from the increasing popularity of AI-powered search engines and LLM-based chatbot interfaces, such as OpenAI's ChatGPT. [3], Anthropic Claude [4], and Google Gemini [5]. These LLM services help increase the dependability of the chatbots across multiple use cases.

While these advances have made chatbots more accessible, this rapid growth in AI chatbot adoption across industries[6] increases demand for a system that can handle multiple user queries, ranging from different complexities and challenges. Building successful enterprise RAG chatbots is challenging because they require meticulous engineering and complex agents to address diverse user queries effectively.[7] In a traditional RAG-based chatbot system, the LLM retrieves external datasets [1], and uses relevant data based on user queries to generate responses. This process creates a heavy dependence on the retrieved datasets [1], determining the quality, accuracy, and relevance of the LLM responses[8].

However, RAG-based chatbot system faces a challenge when the retrieved data contains ambiguous results, conflicting information from similar sources, and noisy, unreliable, or outdated data[9], [10]. These challenges degrade the quality and accuracy of the LLM's abilities to respond [10]. Another crucial limitation of the RAG-based chatbot system is :

Various research studies have been conducted to address data-driven challenges, for example, using Multi-Agent Filtering Retrieval Augmented Generation (MAIN-RAG) with multiple LLM agents and data retrieval filtering mechanisms. These studies have shown improvements in accuracy of 2-11% and a further reduction in noisy documents of 15-30%[1]. Other supporting studies have also reported similar improvements, such as Multi-Agent Debate and Argue Mechanism (MADAM), which employs multiple LLM agents engaging in multi-round debates over retrieved documents with an aggregator acting as a mediator between agent responses. These approaches also achieve improvements on ambiguous queries and a 15.8% improvement on conflicting documents [8]. While both methods lead to improvements over ambiguous, unfiltered, and noisy datasets, a fundamental issue exists within RAG-based chatbot systems: LLM Hallucination and LLM overconfidence in using incorrect data. LLM hallucination occurs when an LLM generates consistent answers that are factually incorrect[9]. More critically, when the system only retrieves incorrect data, LLM models will exhibit high confidence in generating responses using false information[8], [11]. Recent research experiments demonstrate that hallucination and overconfidence issues are not merely a practical failure of the system, but rather the inevitability of how LLM models are trained with a form of optimization and benchmark incentives that encourage LLM models to generate confident responses rather than acknowledging uncertainty about data, incorrect data, or data limitations [12]. These issues require an immediate solution, as they could lead to incorrect information and have the potential to influence the user's critical decisions. This risk is particularly dangerous for government, educational, and healthcare chatbot systems [11]. While the MADAM and MAIN-RAG methods filter out noisy documents and improve accuracy, neither approach directly addresses the hallucinations and overconfidence problems; both approaches can still produce responses that contain misinformation.

To address these issues, this study proposed a Multi-Agent Fallback for Retrieval-Augmented Generation (MAF-RAG). While previous approaches show promise, they face distinct trade-offs. MAIN-RAG[1] employs an adaptive filtering mechanism that effectively removes noise but lacks the deep reasoning capabilities required for complex multi-document synthesis. Conversely, MADAM [8] addresses hallucination through comprehensive multi-agent debate, but applies this computationally expensive process to every query, resulting in high latency that is impractical for real-time public services. A critical gap remains for a system that can balance these needs: delivering the speed of standard retrieval for simple queries while reserving deep multi-agent reasoning only for complex, ambiguous cases.

MAF-RAG introduces a hierarchical escalation protocol that dynamically switches strategies based on confidence. It uniquely positions the expensive Multi-Agent Debate phase as a tertiary fallback-triggered only when both Vector-Only Retrieval (VOR) and Enhanced Vector Retrieval (EVR) fail, but before resorting to external Internet Search. This system is designed for government, education, and healthcare agencies, employing a multi-phase retrieval process. MAF-RAG methods were chosen because of their unique three-phase retrieval system and multi-agent debate capabilities. This approach not only boosts confidence and accuracy but also minimizes hallucinations in LLM outputs. MAF-RAG effectively identifies incorrect documents and subjects them to multiple debates until the response is confirmed as correct and factual. Additionally, MAF-RAG includes an internet fallback to further verify the trustworthiness of documents before generating responses with LLM models.

Integrating MAF-RAG into this study creates a robust RAG testing pipeline. This system learns from hundreds or even thousands of queries from commercial and government sources. Unlike previous methods, which often rely on controlled datasets and sometimes on synthetic data. MAF-RAG shows a significant improvement, similar to previous approaches, while improving on other issues and maintaining dependability for real-world deployment[1], [8].

The novelty of this study is based on the implementation of MAF-RAG to an existing RAG-based chatbot, evaluating real-world feedback, various real-world queries, while testing the system's abilities when faced with real-time challenges such as dataset quality impurities. Unlike previous works that involve synthetic and controlled testing, the goal of this study is to demonstrate the proposed method's ability to address the challenges posed by RAG-based chatbot systems.

## II. METHODOLOGY

The proposed methods of the MAF-RAG retrieval systems follow a structured process to ensure accuracy and reliability during testing. It begins with organizing datasets through embedding and chunking, where approximately 1,100 documents, ranging in multiple variations, are organized and vectorized via Supabase for efficient and secure retrieval. The following step involves classifying each query, helping to categorize them into six semantic classes (Technical, Procedure, Licensing, NIB (business identification number) Support, General). This classification allows the system to use category-specific confidence measurements and retrieval strategies. After the query is classified, the system performs retrieval testing using 50 carefully selected queries to ensure its ability to handle various challenges and the complex nature of real-world queries.

To ensure the MAF-RAG system retrieves optimal and accurate testing performance, it uses a confidence scoring mechanism ranging from 0.0 (No Confidence) to 1.0 (Complete Confidence). This confidence scoring is fundamentally based on cosine similarity; the system converts both the testing queries and document chunks into dense numerical vectors. We selected `all-MiniLM-L6-v2` to enable fully local deployment on constrained hardware (e.g., single consumer GPU), ensuring data privacy for government applications by avoiding external API dependencies. While larger models like E5-Large offer higher resolution, they incur significantly higher latency (1.18 ms vs 0.26 ms per query, a 4.6x slowdown), which is impractical for the target deployment environment. This encoder transforms text inputs into fixed-size 384-dimensional embeddings, regardless of their original length. These embeddings position similar texts closer together in the vector space. For example, when a user enters a query about business licensing procedures into the system, it generates a 384-dimensional query vector and compares it against all document chunk vectors in the database using cosine similarity. If the query matches a document that describes the exact procedure and the system yields a cosine similarity of 0.82, this indicates strong semantic alignment between the query and the retrieved document.

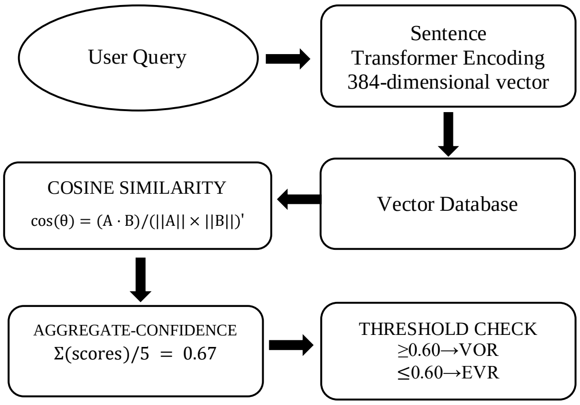


Figure 1. Cosine Similarity-Based Confidence Scoring Pipeline

Figure 1 illustrates the cosine-based confidence-scoring pipeline, comprising five main stages: query encoding, cosine similarity computation, confidence aggregation, threshold escalation logic, and confidence scoring. This process ensures accurate confidence scoring and semantic matching between queries and documents.

### *A. MAF-RAG Retrieval Testing*

Once the system reads the queries, the first retrieval phase is executed using Vector-Only Retrieval (VOR). Formally defined as

$$fVOR(q) = top - k(\cos\_sim(E(q),D), k = 12))$$

In this phase, the system searches through the datasets and retrieves the top 5 relevant documents. The system uses cosine similarity based on the queries with which it has been tested. VOR ensures accurate document selection by leveraging its paraphrasing and contextual capabilities to mitigate noisy and irrelevant documents. VOR applies a specific threshold for selecting documents. If the retrieval confidence rating falls below 0.6, the system advances to the second phase of retrieval testing, utilizing Enhanced Vector Retrieval (EVR), defined as

$$fEVR(q) = top - k\left(cos_{sim(E(q'),D)}, k = 24\right)$$

Where q' represents the query expanded with domain-specific synonyms (e.g., "SIUP" -> "SIUP Surat Izin Usaha Perdagangan") to maximize recall. The ranking function employs a weighted hybrid score:

$$S_{hybrid(d,q')} = \alpha * sim\big(E(q'),E(d)\big) + (1-\alpha) * KeyMatch(q',d)$$

with α = 0.7, prioritizing semantic vector similarity while explicitly boosting documents with exact keyword matches. Finally, the filtering step retains only documents where S_hybrid > 0.35, ensuring that the expanded retrieval does not introduce irrelevant noise. EVR implements this re-ranking mechanism and document filtering to further improve document relevance and accuracy, while reducing the presence of noisy documents.

If the confidence score on the EVR still doesn't meet a satisfactory threshold of the EVR (0.5), the system will escalate to the third and the most comprehensive phase of the retrieval system: MAF-RAG deploys collaborative reasoning efforts between multiple independent agents, each agent use a different document and analyses each document to engage in a debating phase until a convergence has been reached. This phase is designed to handle complex queries and gain a deeper understanding of information across different documents, where the similarity between the query and documents in the previous phase proves to be insufficient to produce a satisfactory confidence threshold level.

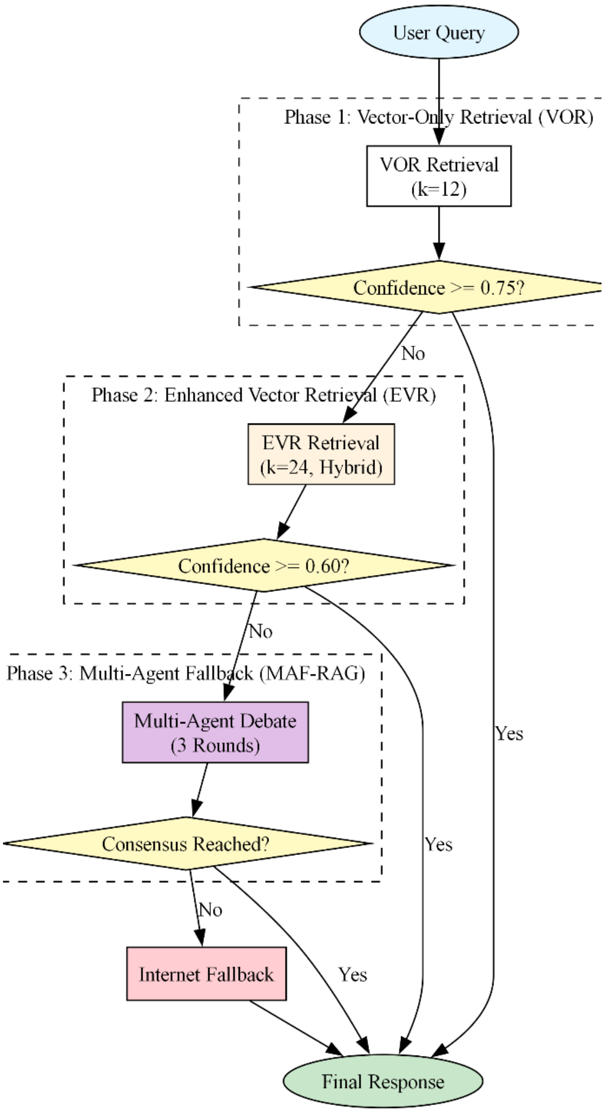

User Query
Phase 1: Vector-Only Retrieval (VOR)
VOR Retrieval (k=12)
Confidence >= 0.75?
No
Phase 2: Enhanced Vector Retrieval (EVR)
EVR Retrieval (k=24, Hybrid)
Confidence >= 0.60?
No
Phase 3: Multi-Agent Fallback (MAF-RAG)
Multi-Agent Debate (3 Rounds)
Consensus Reached?
No
Internet Fallback
Yes
Yes
Yes
Final Response

Figure 2. MAF-RAG Three-Phase Retrieval Escalation Pipeline

As shown in Figure 2, this progressive retrieval process ensures that simple, non-complex queries receive fast responses through VOR, while more complex and ambiguous queries receive more detailed processing with EVR and multi-agent debate via MAF-RAG.

## B. *Multi-Agent Debating Procedure*

The MAF-RAG debating process follows Algorithm 1: MAF-RAG Debate Protocol, formalized as:

$$Phase\ 1\ (Initial\ Response):$$
$$R\_i^1 = f_{(agent)(Q,D_i)}$$
$$Phase\ 2\ (Aggregation):$$
$$A_{agg}^t = f_{(agg)(R_1^t,\ldots,R_n^t)}$$
$$Phase\ 3\ (Update):$$
$$R_i^{t+1} = f_{(agent)(Q,D_i,A_{agg}^t)}$$

where t represents the debate round, converging when $\left|R^{\{(t+1)\}} - R^{\{(t)\}}\right| < \varepsilon\ or\ t = r_{max}$ MAF-RAG analyses the top four most relevant documents retrieved from the EVR phase ($k = 4$). This parameter was chosen for computational efficiency, as larger values would incur excessive LLM API costs and further increase processing time, while yielding diminishing returns in the results. Based on these four most relevant documents, each LLM agent receives one copy of each and begins the debate. The debate operates for a maximum of three rounds. $r\_\{max\} = 3$, during which the agents examine and analyse the document set $D = \{d\_1, d\_2, d\_3, d\_4\}$ And generate responses to the query Q.

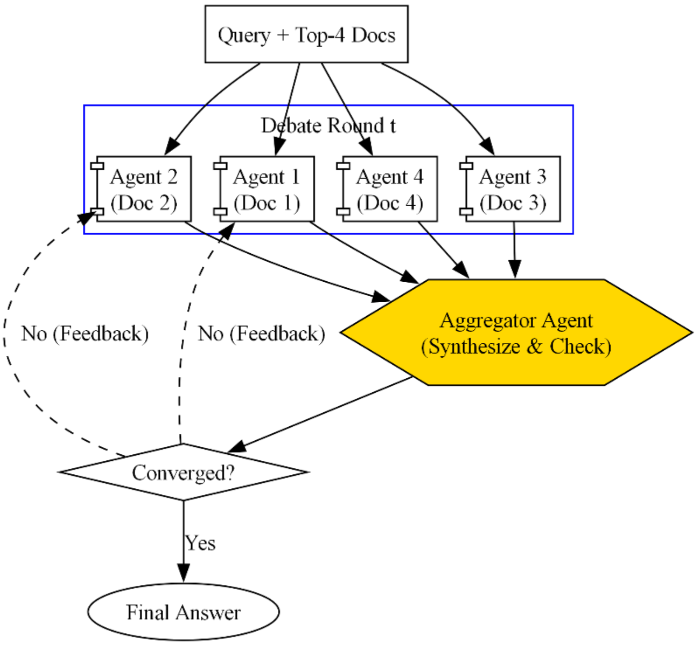


Figure 3. Multi-Agent Debate Architecture in MAF-RAG

Figure 3 illustrates collaborative reasoning through multiple debate rounds. In the first round of debate, each agent $i$ summarizes the retrieved document, where each agent $i$ is assigned to its own independent document $d_{\mathrm{i}}$ and generates an answer based on those documents $R\_i^1 =$

$$R\_i^1 = f\_{agent(Q,D_i)}$$

The function f_agent represents the agent's reasoning abilities implemented through a custom LLM adapter. An aggregator then aggregates these agents' responses to identify common patterns and generate an aggregated response. $A_{agg}^t$:

$$A_{agg}{}^{(t)} = f\_{agg}(R\_1^t, \ldots, R\_n^t)$$

where f_agg represents the aggregator function that synthesizes responses to check for convergence, in subsequent rounds (t > 1), agents refine their responses by considering both their assigned document and the reasoning provided by their peers in the previous round:

$$R\_i^{t+1} = \quad f\_agent(Q, D\_i, \{R\_j^t \mid j \neq i\})$$

This peer-to-peer context enables agents to critique and refine their reasoning based on conflicting or supporting evidence from other documents. The debate continues until convergence or the maximum round limit is reached. If the final aggregation remains "unknown," the system escalates to the Internet-Fallback phase..

TABLE I
CONFIDENCE SCORE THRESHOLDS AND PHASE ESCALATION LOGIC

| Phase | Confidence Threshold | Action If Met | Action If Not Met |
|---|---|---|---|
| VOR | ≥ 0.75 | Accept and return result | Escalate to EVR |
| EVR | ≥ 0.60 | Accept and return result | Escalate to MAF-RAG |
| MAF-RAG | ≥ 0.65 | Accept if the result not "unknown" | Escalate to Internet-Fallback |
| Internet-Fallback | ≥ 0.55 | Always accept if results found | Return failure message |

### C. Datasets and Data Context

This study uses a prototype RAG-based chatbot system under development for Central Java Investment Platform (CJIP)[13], Online Single Submission (OSS)[14], and Pelayanan Terpadu Satu Pintu (PTSP) in general [15]. The development of this chatbot aims to enhance responses and provide citizens with accurate, reliable information regarding government services, investment procedures, and business licensing, with the goal of reducing the time required for future investors to invest in Central Java. The development of this chatbot addresses a critical need for accessing vital and accurate information, as citizens often become confused when navigating complex administrative procedures. The urgent need for accurate and reliable information aligns perfectly with the foundation of MAF-RAG, which integrates directly with the knowledge base's systems to handle various and complex citizen queries in a near-production-like environment, where accuracy, reliability, and minimal misinformation are critical.

The MAF-RAG system was evaluated using a specialized dataset consisting of 1,100 documents and questions in the context of government and administrative procedures. Each dataset represents the authentic inquiry ecosystem that citizens might ask in a real deployment scenario. To further evaluate the MAF-RAG effectiveness, this study uses two sets of dataset configurations: **Legacy Dataset,** which represents the original knowledge base of the prototype system, sometimes consisting of outdated regulations, incomplete procedures and inconsistent formatting of the documents, This dataset demonstrate the main challenges of the RAG-based chatbot system, where a constantly updating datasets and sudden changes in procedural purposes could result in the system producing a dangerous responses. **Enhanced Dataset,** representing a newly improved dataset that is carefully chosen from up-to-date FAQs, revised procedural guidelines, and more standardized documentation formats. This dataset also represents how much the RAG-based chatbot system is entirely dependent on the quality of the dataset, which determines the system's ability to generate a response regardless of the sophistication of multi-agent debates or modern reasoning solutions.

While the quality of the datasets determines the quality of the system performance, a comprehensive document preprocessing is also crucial for the effectiveness of a retrieval system. Both Legacy and Enhanced datasets undergo consistent and standardized document processing. Preprocessing begins with extracting text from various formats present in the system knowledge base, including multiple formats such as PDF, DOCX, and HTML. Each format requires a thorough extraction method to convert into plain text while still preserving structural information, such as numbered lists or procedural steps.

Following text extractions, all documents are parsed to a textual content which then the system employs a hybrid chunking method combining both size segmentation and semantic detection, all documents are segmented into divided chunks of 1200 characters with 100 characters of overlap between consecutive chunks to maintain continuity across chunks. The chunked documents are then vectorized using the sentence-transformers/all-MiniLM-L6-v2 embedding model[16] This embedding model is executed locally on an Nvidia CUDA GPU to ensure data privacy and reduce external API cost, which is critical for government applications. All document vectors are indexed and stored in Supabase's pgvector extension that supports approximate nearest neighbor (ANN) search by using the Hierarchical Navigating Small World (HNSW) indexing algorithm for a sub-second retrieval latency.

For language generation, the MAF-RAG system uses Llama-3.3-70B-Versatile via Groq's API, with a maximum of 8,000 tokens to accommodate the MAF-RAG multi-agent debate and agent responses.

### D. Evaluation Metrics

The evaluation methods used to determine the effectiveness of the proposed MAF-RAG systems consist of three experiment configurations: a baseline RAG-based chatbot with the Legacy Dataset, a Baseline RAG-based chatbot with the Enhanced Dataset, and MAF-RAG.

All of the experiment configurations is tested using 50 carefully selected queries spanning between 6 semantic categories.

TABLE II
ESCALATION WORKFLOWS ACROSS SYSTEM CONFIGURATIONS

| Configuration | Dataset | Pipeline Sequence |
|---|---|---|
| Legacy Baseline | Old (Noisy) | VOR to EVR to Internet Fallback |
| Enhanced Baseline | New (Clean) | VOR to EVR to Internet Fallback |
| MAF-RAG | New (Clean) | VOR to EVR to Debate to Internet Fallback |

It is important to note that the Baselines do not include the Debate phase. All configurations utilize identical retrieval parameters: VOR (k=12) and EVR (k=24) with the hybrid scoring threshold set at 0.35, ensuring a fair comparison of the architectural impact. All experimental configurations are evaluated on 50 carefully selected queries across six semantic categories.

TABLE III
DETAILED SYSTEM CONFIGURATION PARAMETERS

| Component | Legacy | Enhanced | MAF-RAG |
|---|---|---|---|
| Dataset Quality | Outdated/noisy | Cleaned | Cleaned |
| Embedding Model | all-MiniLM-L6-v2 | all-MiniLM-L6-v2 | all-MiniLM-L6-v2 |
| Chunk Size | 1,200 chars | 1,200 chars | 1,200 chars |
| Chunk Overlap | 200 chars | 200 chars | 200 chars |
| VOR k / Threshold | 12 / 0.75 | 12 / 0.75 | 12 / 0.75 |
| EVR k / Threshold | 24 / 0.60 | 24 / 0.60 | 24 / 0.60 |
| Hybrid Score α | 0.7 | 0.7 | 0.7 |
| Debate Phase | No | No | Yes (k=4, r_max=3) |
| LLM Model | Llama-3.3-70B | Llama-3.3-70B | Llama-3.3-70B |
| Max Tokens | 8,000 | 8,000 | 8,000 |
| Temperature | 0.6 | 0.6 | 0.6 |

Retrieval performance is provided from a measurement of Precision, Recall, and F1-Score. Precision measures the fragmented data of the retrieved documents that are relevant in conjunction with the testing queries:

$$Precision = \frac{|R \cap G|}{|R|}$$

Where R denotes the set of retrieved documents, and G denotes the set of ground-truth relevant documents, the ground-truth set (G) was established without incurring excessive API costs by performing a Retrieval-Only Verification Run. In this phase, the system was executed using the corresponding dataset with the generative component disabled, isolating the retrieval mechanism. The document chunks identified during this optimal pass were recorded as the ground truth. This approach circumvents the need for manual annotation of thousands of query-document pairs while providing a consistent baseline derived from the system's best-case retrieval performance. Consequently, F1-scores below 1.0 in experimental runs reflect deviations from the optimal retrieval standard, attributable to factors such as dataset noise (Legacy) or algorithmic variations. The model response is expected to remain within the ground truth. Whereas Recall measures the fraction of data of the relevant documents that are successfully retrieved:

$$Recall = \frac{|R \cap G|}{|G|}$$

F1-score combines both precision and recall into a single metric through their harmonic mean:

$$F_1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

In addition to retrieval performance metrics, this study also employs system reliability metrics, which provide insight into real-world performance, including Response Time, Category Performance, System Failure and Success Rates, and Phase Utilization. Since this study evaluates MAF-RAG across both Legacy and Enhanced Datasets, the metrics were calculated for each dataset, allowing a detailed assessment of how data quality impacts, and isolating MAF-RAG's effectiveness in handling such challenges.

## III. RESULTS AND DISCUSSION

This chapter presents the findings of the comparative evaluation across three experiment configurations. Each configuration uses 50 test queries measuring retrieval performance through precision, recall, and F1-score, and system reliability metrics through failure and success rate, phase utilization, and category performance.

## *A. Dataset Quality Impact on Retrieval Performance*

Before examining the RAG system, it is worth mentioning the foundational role of dataset quality in its performance. A comparison between two configurations provides critical insights into how data quality impacts system performance. The Legacy Dataset baseline system achieved a mean F1-score of 0.32, with 66.0% of the queries resulting in failure (F1 = 0.0) and a 34% success rate (F1 > 0 ). While the Enhanced Dataset Baseline system achieved a mean F1-score of 0.468 and a success rate of 62%, representing improvements of 43.1%

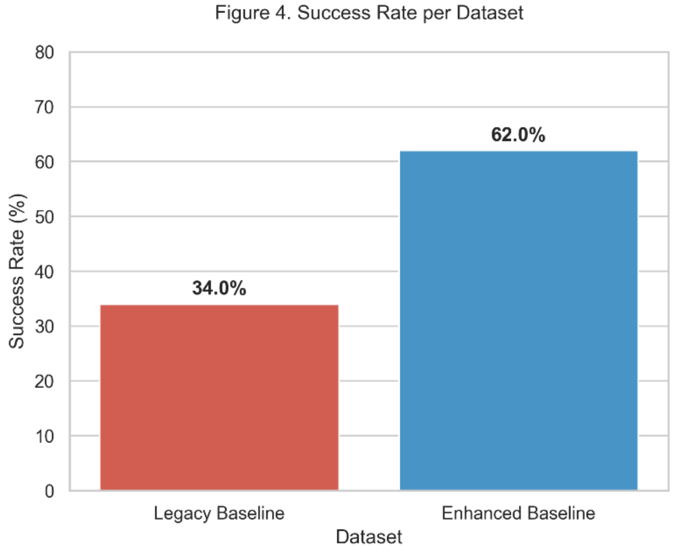


Figure 4. Success Rate per Dataset: Legacy 34%, Enhanced 62%

As shown in figure 4, these findings directly prove data quality enhancement, such as updated regulatory information, standardized document format, and updated documents, substantially improve performance and reliability on baseline retrieval systems.

TABLE IV
BASELINE PERFORMANCE COMPARISON BASED ON DATASETS QUALITY

| Metric | Legacy | Enhanced | Improvement |
|---|---|---|---|
| Avg F1-Score | 0.327 | 0.468 | +43.1% |
| F1=1.0 Retrieval | 16 (32.7%) | 18 (36.7%) | +12.5% |
| Succes Rate | 34.0% | 62.0% | +82.5% |
| Failed Retrieval | 33 (66.0%) | 19(38.0%) | -42.4% |

Further improvements are evident in category-specific metrics. The Enhanced Dataset baseline system benefited most from data organization, such as NIB queries, which achieved a mean F1-score of 0.750, up from 0.250 on the Legacy Dataset baseline system—a 240% improvement due to enhanced dataset quality. Other categories also showed significant gains, with Technical and Procedural queries improving by 95% and 100%, respectively.

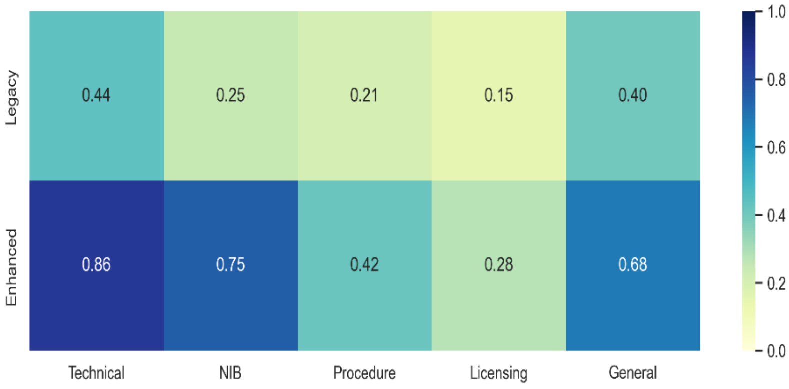


Figure 5. Category Performance Heatmap: Technical, NIB, Procedure, Licensing, General

From the heatmap visualization in Figure 5, we can see the improvement between the two baseline systems on category-specific metrics, with the Technical and NIB categories showing the most pronounced improvements.

These results establish a critical principle: regardless of the sophistication of the retrieval system, it cannot compensate for poor dataset quality. This improvement was achieved without modifying the baseline system's algorithm; this also demonstrates that the quality of the RAG-based chatbot's knowledge base is essential to system accuracy.

## *B. Document Retrieval Effectiveness Comparison*

Having established that dataset quality can improve system accuracy, the proposed MAF-RAG three-phase system adds an algorithmic aspect on top of data quality. The MAF-RAG system achieved a mean F1-score of 0.556, a 18.8% improvement over the Enhanced Dataset baseline system and a further 70.0% over the Legacy Dataset baseline system. A paired t-test confirms this improvement is statistically significant [ p = 0.009 ]. This significant performance improvement demonstrates that further enhancements to algorithmic factors, such as a more sophisticated retrieval system and a multi-agent refinement and validation process, provide a measurable gain beyond data quality improvements alone..

TABLE V
OVERALL SYSTEM RETRIEVAL PERFORMANCE COMPARISON

| Metric | Legacy | Enhanced | MAF-RAG | Δ |
|---|---|---|---|---|
| Avg F1-Score | 0.327 | 0.468 | 0.556 | +18.8 % |
| F1=1.0 Retrieval | 16 (32.7%) | 18 (36.7%) | 19(39.8%) | +5.4% |
| Succes Rate | 34.0% | 62.0% | 78.0% | +26% |
| Failed Retrieval | 33 (66.0%) | 19(38.0%) | 11(22.0%) | -42.1% |
| Avg.Time (s) | 14.5696 | 11.0838 | 11.6270 | +4.9% |

As shown in Table V, the proposed MAF-RAG consistently outperforms both the Legacy and Enhanced datasets across all retrieval performance metrics. The improvement in the mean F1-score indicates that the MAF-RAG system retrieves more relevant documents with higher precision and recall. In addition, it achieved the highest F1 score of 1.0 ever recorded on any system configuration. The MAF-RAG achieves a 26% improvement over the Enhanced Baseline and a significant increase in success rate compared with the Legacy Baseline, while maintaining the lowest failure rate, with a 42% reduction relative to the Enhanced Baseline.

TABLE VI
FAILURE ANALYSIS BY SEMANTIC CATEGORY

| Categories | Failures | % of Total Failures | Failure Rate |
|---|---|---|---|
| Procedure | 7 | 63.6% | 29.2% (7/24) |
| Licensing | 3 | 27.3% | 60.0% (3/5) |
| General | 1 | 9.1% | 14.3% (1/7) |

Table VI details the remaining 22% of failures (11 queries). The majority (7/11) stemmed from Procedural queries, in which the system struggled to synthesize steps from multiple conflicting regulations. However, the Licensing category exhibited the highest failure rate (60%), indicating a specific weakness in handling highly specialized regulatory definitions. This suggests that while Debate improves reasoning, extremely complex multi-document synthesis and niche regulatory disambiguation remain challenges.

Although the MAF-RAG is 4.9% slower than the Enhanced Baseline in the Avg Time metrics, it improves by 20.2% over the Legacy Baseline. In a government advisory context, the cost of misinformation (hallucination) far outweighs the price of a 0.5s delay. Users prefer a correct answer in 11s over a wrong answer in 10s. Thus, this trade-off is minimal compared to the advantages that MAF-RAG gained over both baseline systems, which further improve accuracy and reliability.

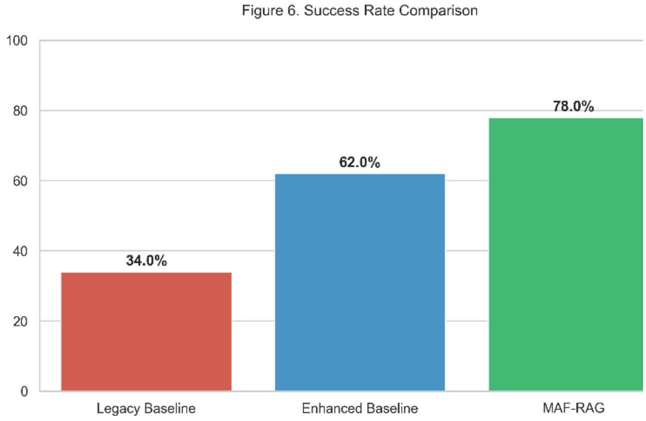


Figure 6. Success Rate Comparison: Legacy 34%, Enhanced 62%, MAF-RAG 78%

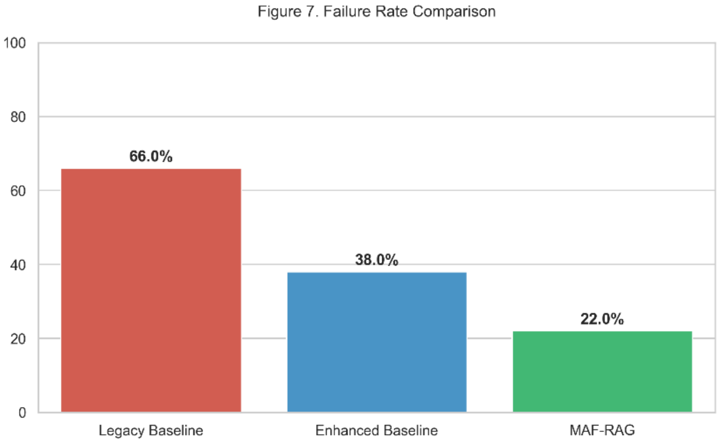


Figure 7. Failure Rate Comparison: Legacy 66%, Enhanced 38%, MAF-RAG 22%

Figure 6 and 7 further illustrates the success and failure rate comparison across three experiment configurations. With MAF-RAG outperforming both Legacy and Enhanced Baseline.

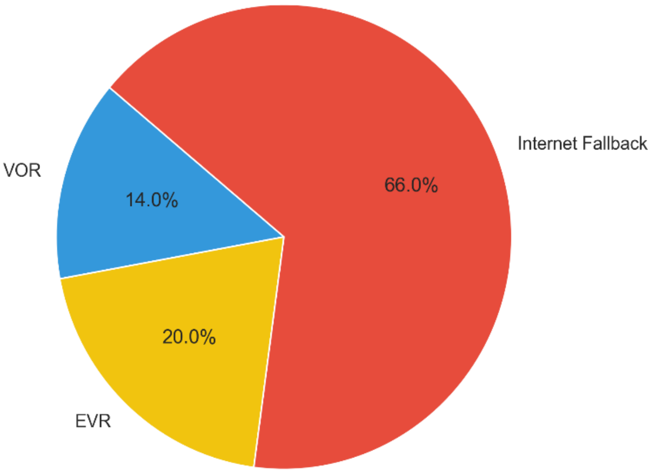
TABLE VII
PHASE UTILIZATION ACROSS ALL THREE SYSTEM

| Retrieval Phase | Legacy | Enhanced | MAF-RAG |
|---|---|---|---|
| EVR | 20.0% | 14.0% | 38.0% |
| VOR | 14.0% | 30.0% | 30.0% |
| MAF-RAG | N/A | N/A | 30.0% |
| Internet Fallback | 66.0% | 38.0% | 2% |

From Table VII, the phase utilization is the next key factor in understanding the system's effectiveness in retrieving relevant and accurate data for each testing query, which shows the system's reliance on the most used phase.

Figure 8. Legacy Baseline Phase Utilization Chart

As shown in Figure 8. The Legacy Baseline is overreliant on internet fallback (66%), indicating that the knowledge base's unable to answer most queries using internal data. EVR and VOR, respectively handled the remaining 20% and 14%. Higher EVR usage also means the system cannot search relevant data through the knowledge base of the Legacy Baseline system



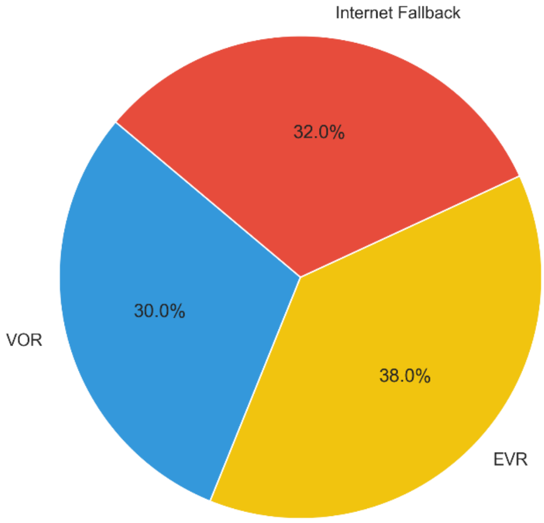


Figure 9. Enhanced Baseline Phase Utilization Chart

Figure 9 shows that the Enhanced Baseline reduces internet fallback reliance to 32%, while EVR and VOR together handle 68% of queries internally. This phase distribution further proves that data quality reduces system dependence on last-resort fallback without requiring changes in the algorithmic aspect.

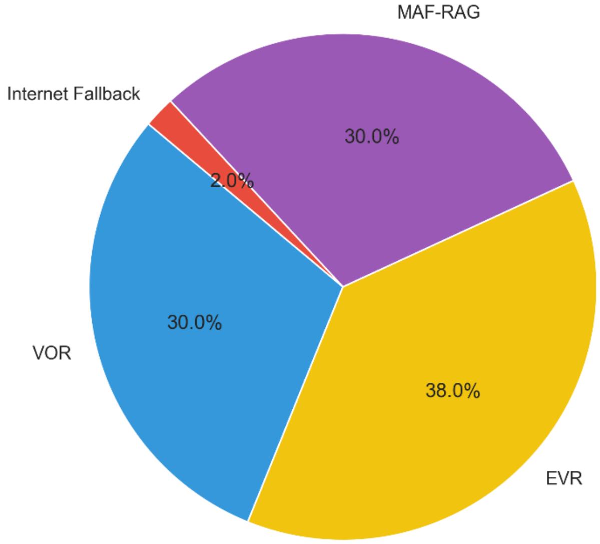


Figure 10. MAF-RAG Phase Utilization Chart

Figure 10 illustrates the phase utilization of the MAF-RAG system. The system successfully resolves the majority of queries internally, with VOR handling 30.0% and EVR handling 38.0%. The Multi-Agent Fallback (MAF-RAG) phase resolves an additional 30.0% of complex queries that failed the initial confidence checks. Consequently, the reliance on Internet Fallback is minimized to only 2.0%, demonstrating the system's ability to handle a wide range of query complexities without resorting to external search.

TABLE VIII
MAF-RAG INTERNAL PHASE PERFORMANCE

| **Retrieval Phase** | **Queries** | **% of Total** | **Mean F1** | **F1 = 1.0** | **Success Rate** | **Failed Rate** |
|---|---|---|---|---|---|---|
| EVR | 19 | 38.0 % | 0.737 | 11 | 89.5 % | 2 |
| VOR | 15 | 30.0 % | 0.627 | 6 | 93.3 % | 1 |
| MAF-RAG | 16 | 32.0 % | 0.275 | 2 | 50.0 % | 8 |

From Table VIII, we can focus on the internal phase performance of MAF-RAG. EVR handled 19 queries, achieving a mean score of 0.737 and an 89.5% success rate, whereas VOR handled 15 queries, achieving a mean score of 0.627 and a higher success rate of 93.3%. MAF-RAG handles the more complex queries that didn’t pass EVR and VOR confidence thresholds, with MAF-RAG achieving only a 50%

success rate. This also indicates that the system resolves half of otherwise unresolvable queries, which would otherwise fail.

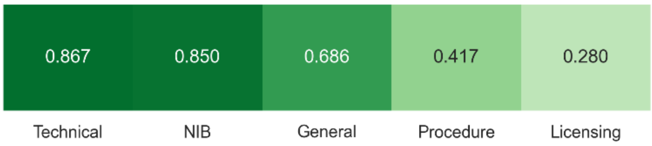


Figure 11. MAF-RAG Heatmap of Categorical Performance

Figure 11 illustrates a category-specific analysis revealing MAF-RAG further delivers benefits across all categories. Technical queries achieved a mean F1 score of 0.867, an improvement of 35% over the Enhanced baseline, and NIB queries also achieved a mean F1 score of 0.850, another improvement of 13% over the Enhanced baseline. These results represent the most substantial performance improvement over both baseline systems, suggesting a technical and dataset enhancement benefit substantially from a multi-agent debate to produce information and validate both accuracy and consistency across retrieved documents.

### *C. System Reliability Analysis*

Beyond the raw F1-score, phase utilization, and categorical performance, a system reliability analysis is required to further understand the MAF-RAG performance distribution or to handle test queries and real-world behavior.

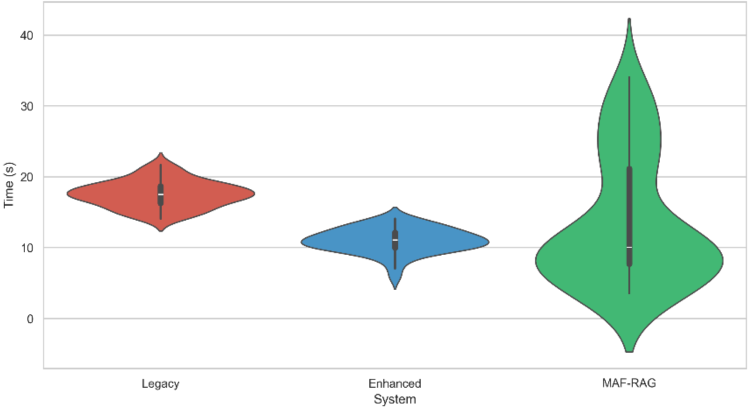


Figure 12. Retrieval Time Distribution Across All Three Systems

Figure 12 illustrates a violin plot of the distribution of retrieval times across three system configurations, highlighting the wider range of retrieval times for each system. The Legacy Baseline shows a relatively more compact distribution of retrieval times, with most clustering between 15 and 22 seconds. In contrast, the Enhanced Baseline demonstrates improvements in responsiveness with its retrieval times shifting to a lower time spectrum, suggesting that the improved data quality contributes significantly to more efficient retrieval. The MAF-RAG exhibits a wide spread of retrieval times, ranging from 3 to 8 seconds, but also occasionally goes to high-latency above the 25-second range. This wider range of time spread reflects the design of MAF-RAG itself, where easy, non-complex queries are resolved in the earlier phase, and harder, more complex queries are handled through a multi-agent process, which requires more time in general.

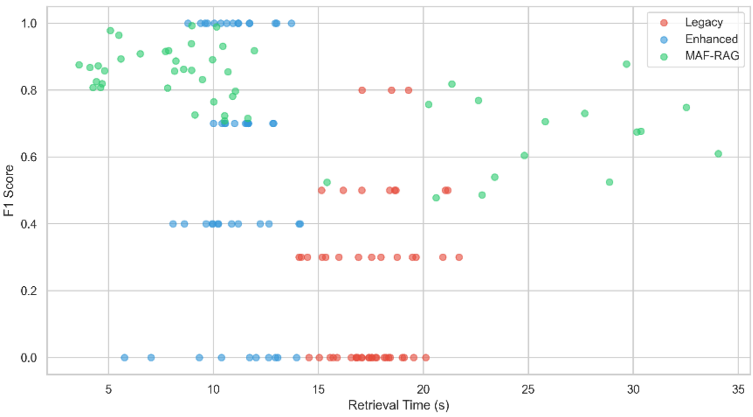


Figure 13. Time Scatter Plot Showing Performance Trade-Offs

Figure 13 illustrates the scatter plot of the relationship between retrieval accuracy measures by F1-score and retrieval latency across the three system configurations. The Legacy Baseline shows a dense cluster of points between 18 and 21 seconds, with a significant portion of queries achieving a 0.0 rate, which correlates to a 66% failure rate. In contrast, the Enhanced Baseline presents a faster and more balanced distribution of time clusters, with more points achieving higher F1-score values. This more balanced distribution reflects the addition of standardized document formats, updated content, and revised procedural guidelines, resulting in lower latency and improved retrieval performance.

The MAF-RAG system exhibits the widest spread in retrieval time, as illustrated in the previous Figure 12. However, this does not imply reduced accuracy, unlike the baselines, a high F1-score result is still consistently achieved across the entire latency spectrum. Overall, the scatter plot confirms that retrieval accuracy is not directly related to latency, but rather to retrieval strategy and data quality. The MAF-RAG system delivers the strongest performance across all baseline systems, achieving higher F1-scores and a marginally higher latency trade-off than the baseline systems.

This work's practical contribution is twofold. First, it establishes a replicable and validated machine learning testing pipeline for the future development of an RAG-based chatbot system, enabling the testing of system performance under limited dataset constraints —a methodological solution for developers who want to optimize the retrieval system. Second, it demonstrates that data quality improvements alone yield a 43.4% performance gain, and also in reducing 42% in failures rates, confirming the importance of system knowledge base before changing algorithmic complexity. In conclusion, this study positions the MAF-RAG as a solution for addressing challenges that persist in a RAG-based chatbot system, a solution that already proven to outperforms any of the baseline system in a near-deployment like chatbot system, a clear model of choice for any future works revolving around RAG-based chatbot systems where a dataset quality, algorithmic design, and real-world operational are a constraining factors.

### *D. Cross-Domain Generalization and Modularity*

The MAF-RAG architecture is domain-agnostic by design. Adapting the system to new domains (e.g., medical, legal, e-commerce) requires only replacing the dataset, re-embedding, and updating the externalized domain configuration; the retrieval logic, debate protocol, and confidence thresholds remain unchanged. All domain-specific terminology is externalized to configuration files rather than hardcoded, enabling rapid deployment across sectors. For instance, transitioning from government administrative data to hospital procedure documentation would require re-vectorizing the new knowledge base with the same embedding model (sentence-transformers/all-MiniLM-L6-v2) and adjusting the domain-specific configuration parameters, without modifying the three-phase escalation (VOR → EVR → MAF-RAG) or the multi-agent debate logic. This modularity addresses a critical concern about cross-domain applicability, positioning MAF-RAG as a general-purpose framework rather than a domain-specific solution.

## IV. CONCLUSION

This study successfully implemented and evaluated an extensive retrieval testing pipeline, examining its impact on dataset quality. It also implemented Multi-Agent Fallback in Retrieval-Augmented Generation (MAF-RAG), combining data retrieval filtering, multi-agent debate, and answer aggregation to provide accurate and reliable information. The datasets used in this test consisted of 1,100 documents. These documents were then text-extracted and divided into 150 selected test queries. Each query was then segmented into 1,200-character chunks, with 100 characters of overlap between consecutive chunks to maintain continuity.

Results demonstrate that MAF-RAG delivers substantial improvements and outperforms both baseline systems, achieving a mean F1-score of 0.556. MAF-RAG surpasses Enhanced Baseline by 18.8% and Legacy Baseline by 70.0%. It also achieves a staggering success rate of 78% the highest recorded in retrieval testing, while maintaining a low failure rate, confirming that multi-agent debate and fallback mechanisms effectively address complex queries where traditional approaches cannot resolve. The MAF-RAG system exhibits slightly higher latency than the Enhanced baseline, with a slight 4.9% trade-off. However, it still outperforms the legacy baseline by completing the retrieval test 20.2% faster. This trade-off is marginal compared to the significant gains in retrieval accuracy, reliability, and failure rates.

In a real-world deployment environment, the findings of this research provide a significant practical and realistic contribution to the development of a retrieval-augmented generation chatbot system. The ability to accurately retrieve and validate responses before generating final responses is a strategic advantage for the end user. This study demonstrates the influence of dataset quality, highlighting the challenges that even a sophisticated retrieval system crumbles under a poorly structured knowledge base. This MAF-RAG system serves as a validated proof of concept for an algorithmic architectural improvement that provides an accurate, noise-free, and hallucination-risk-free system.

Unlike many RAG approaches that rely on large-scale datasets and extensive use of LLMs, this study highlights the effectiveness of the three main components of the architecture (Data Improvement, Three-Phase Retrieval System, and Multi-Agent Debate and Fallback Mechanism) in addressing various data-quality issues, outdated content, and ambiguity. By implementing a three-phase retrieval system (VOR, EVR, MAF-RAG) it demonstrates effectiveness in managing queries of varying complexity. Simple queries resolve rapidly via VOR with a success rate of 93.3%; moderate-complex queries via EVR with a success rate of 89.5%; and most complex queries via MAF-RAG with a 50% success rate.

This contribution to the understanding of practicality and efficiency of the proposed MAF-RAG system, particularly for a government, education, and healthcare deployment where data quality is left in an unregulated, unstandardized state, and the cost of inaccuracy, misinformation, and hallucination far exceeds the system's inability to generate a reliable and trusted answer, as even a single inaccurate response could affect critical decision.

Future work will focus on four key areas to further enhance system efficiency and adaptability. First, Heterogeneous Agents will be introduced, with each agent deploying a different LLM model, replacing the current homogeneous setup with specialized roles to improve debate quality. Second, Adaptive Retrieval will be implemented to dynamically switch between "Deep" and "Normal" retrieval modes based on query complexity, optimizing resource usage. Third, a Dynamic Phase Selection mechanism will be developed, using a router model to predict the necessary retrieval phase (VOR, EVR, or Debate) immediately upon query receipt, thereby reducing latency by bypassing unnecessary steps. Finally, Continuous Learning via Reinforcement Learning from Human Feedback (RLHF) will be integrated to fine-tune the router and agents based on real-world user interactions, ensuring the system evolves with changing information needs.

## REFERENCES

[1] C. Chang *et al.*, “MAIN-RAG : Multi-Agent Filtering Retrieval-Augmented Generation,” 2024.

[2] R. Bommasani *et al.*, “On the Opportunities and Risks of Foundation Models,” pp. 1–214, 2022, [Online]. Available: http://arxiv.org/abs/2108.07258

[3] OpenAI, “ChatGPT,” 2025, GPT-5. [Online]. Available: https://openai.com/index/introducing-gpt-5/

[4] Anthropic, “Claude 4,” 2025, Claude 4.5 Sonnet. [Online]. Available: https://www.anthropic.com/news/claude-4

[5] Google., “Gemini,” 2025, Gemini 2.5. [Online]. Available: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/

[6] N. Alharbi, F. Ud Din, D. Paul, and E. Sadgrove, “Driving AI chatbot adoption: A systematic review of factors, barriers, and future research directions,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 3, p. 100590, 2025, doi: 10.1016/j.joitmc.2025.100590.

[7] R. Akkiraju *et al.*, “FACTS About Building Retrieval Augmented Generation-based Chatbots,” 2024, [Online]. Available: http://arxiv.org/abs/2407.07858

[8] H. Wang and E. Stengel-eskin, “Retrieval-Augmented Generation with Conflicting Evidence,” pp. 1–22, 2025.

[9] I. Augenstein *et al.*, “Factuality Challenges in the Era of Large Language Models,” pp. 1–13, 2023.

[10] W. Chen, Y. Pan, and L. Pan, “On the Risk of Misinformation Pollution with Large Language Models,” no. 2, pp. 1389–1403, 2023.

[11] J. Zhou and A. G. Parker, “Synthetic Lies : Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions,” 2023, doi: 10.1145/3544548.3581318.

[12] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, “Why Language Models Hallucinate,” vol. 3, no. May, pp. 1–36, 2025, [Online]. Available: http://arxiv.org/abs/2509.04664

[13] Pemerintah Provinsi Jawa Tengah, “Central Java Investment Platform.” [Online]. Available: https://cjip.jatengprov.go.id/

[14] Pemerintah Indonesia, “Online Single Submission (OSS-RBA).” [Online]. Available: https://oss.go.id/

[15] D. J. Tengah, “Dinas Penanaman Modal dan Pelayanan Terpadu Satu Pintu Provinsi Jawa Tengah.” [Online]. Available: https://dpmptsp.jatengprov.go.id/

[16] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.