Weighted Reciprocal Rank Fusion RAG for Context-Aware DoS Attack Mitigation

Abdullahil Kafi
Dept. of Computer Science & Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh.
mail.at.kafi@gmail.com

Sajal Saha, *Senior Member, IEEE*
Dept. of Computer Science
University of Northern British Columbia
Prince George, Canada
sajal.saha@unbc.ca

Nashid Shahriar
Dept. of Computer Science
University of Regina
Regina, Canada
nashid.shahriar@uregina.ca

***Abstract*—Modern cybersecurity systems rely increasingly on machine learning (ML) for threat detection, yet they often fall short in delivering context-specific mitigation strategies. To bridge this gap, we propose an explanation-aware Retrieval-Augmented Generation (RAG) framework that tightly integrates supervised ML-based attack detection with Large Language Model (LLM)-driven mitigation guidance. We propose a Weighted Reciprocal Rank Fusion (WRRF)—a novel ranking method that enhances multi-query retrieval by incorporating retriever-side confidence scores. This ensures that semantically relevant and high-confidence documents from cybersecurity knowledge bases (ENISA, NIST, CISA) are prioritized during response generation. Our system begins by classifying suspicious network traffic using a Random Forest classifier trained on the UNSW-NB15 dataset. It then constructs explanation-rich prompts grounded in key anomalous features to query semantically indexed domain-specific documents. Using a multi-query strategy, the framework retrieves diverse candidate documents, which are then aggregated using WRRF to improve contextual alignment and ranking fidelity. Experimental evaluations across multiple response-generation baselines—including OpenAI, standard RAG, and RRF—demonstrate that WRRF achieves superior performance in mitigation accuracy, semantic relevance to traffic indicators, document source diversity, and response precision.**

***Index Terms*—DoS Detection, Retrieval-Augmented Generation, Large Language Models, Cybersecurity, Semantic Ranking**

I. Introduction

Cybersecurity has shifted from static rules to ML-driven detection that learns patterns in network traffic (e.g., DoS, phishing, ransomware). Yet a key gap remains: models classify attacks but rarely provide actionable, context-aware mitigation. Closing this gap requires coupling detections with reasoning and domain knowledge. LLMs show promise for malware analysis, threat reporting, and incident response [1], but general-purpose models often produce vague or incorrect mitigation guidance due to hallucination [2], stale knowledge, and limited grounding—risks that are unacceptable in operational cybersecurity. To address these shortcomings, several strategies have been explored, including fine-tuning, in-context learning (ICL) [3], and Retrieval-Augmented Generation (RAG) [4]. Among them, RAG offers a scalable, interpretable solution by retrieving relevant knowledge chunks from external corpora to ground LLM outputs in authoritative information. However, traditional RAG applications in cybersecurity often function as generic question-answering systems, lacking integration with detection models and failing to provide threat-specific, operationally actionable responses.

One critical component of RAG systems is the ranking mechanism used to select the most relevant documents for generation. Typically, documents are ranked by their vector similarity to a query. However, when multiple reformulations of a query are used—as in RAG-Fusion [5]—it becomes necessary to merge multiple ranked lists into a unified set. Reciprocal Rank Fusion (RRF) is a widely adopted method for this task, assigning each document a fused score based on its rank across multiple lists. RRF has proven effective in improving answer comprehensiveness by considering multiple perspectives of a query. However, it suffers from key limitations. First, it treats all queries as equally important, failing to account for the semantic quality or relevance of individual query formulations. As a result, low-quality or noisy queries may dilute the ranking process, reducing the precision of the final document set. Second, RRF is agnostic to the confidence of the retriever, which could otherwise provide useful signals for weighting documents retrieved with higher certainty. To overcome these limitations, we introduce a novel ranking method: *Weighted Reciprocal Rank Fusion (WRRF)*. Our method incorporates retriever-side confidence scores into the fusion process. This allows our system to amplify the influence of high-confidence, semantically aligned queries while suppressing the effect of less relevant or ambiguous ones. By doing so, WRRF improves both retrieval precision and generation relevance, making the final output more trustworthy for high-stakes cybersecurity contexts.

In this work, we present a comprehensive pipeline that tightly integrates machine learning-based attack detection with WRRF-enhanced Retrieval-Augmented Generation. Our system translates detection outputs into interpretable, explanation-aware prompts, which are used to query a semantically indexed cybersecurity knowledge base composed of documents from European Union Agency for Cybersecurity (ENISA), the National Institute of Standards and Technology (NIST), and the Cybersecurity and Infrastructure Security Agency (CISA). Retrieved documents are reranked via WRRF and passed to an LLM for mitigation strategy generation. Our main contributions are summarized as follows:

- We present a framework that integrates ML-based attack detection, explanation-aware prompt construction, and RAG for LLM-driven mitigation.
- We introduce *Weighted Reciprocal Rank Fusion (WRRF)*, which leverages retriever confidence to refine document selection.
- Experiments on multiple dimensions—ranking resolution, source diversity, semantic relevance, and strategy adaptation—show that WRRF outperforms OpenAI, RAG, Multi-Query, and RRF.

The rest of the paper is organized as follows: Section II reviews related work; Section III details our proposed methodology; Section IV presents our experimental results; and Section V concludes the paper and outlines future directions.

II. Literature Review

The integration of Large Language Models (LLMs) into cybersecurity workflows has gained significant momentum in recent years, offering promising advances in threat detection, response automation, and vulnerability analysis. However, while general-purpose LLMs demonstrate strong language understanding, they often fall short in producing precise, context-specific mitigation strategies for evolving cyber threats. This limitation has motivated the development of Retrieval-Augmented Generation (RAG) frameworks, which enhance LLM outputs by grounding them in external, authoritative knowledge. Our proposed framework builds upon this foundation by introducing a novel ranking mechanism—Weighted Reciprocal Rank Fusion (WRRF)—to improve the selection of relevant documents and deliver more targeted, context-aware mitigation guidance. Ferrag et al. [6] provided a comprehensive review of LLM applications in cybersecurity, including their use in intrusion detection, malware analysis, and real-time threat response via RAG techniques. Xu et al. [7] conducted a systematic study of 185 publications, highlighting growing applications of LLMs in tasks such as vulnerability detection, phishing, and network intrusion analysis, while emphasizing the importance of fine-tuning and prompt engineering. Singh and Alfardan [8] explored LLMs for vulnerability management, advocating for context-aware threat mitigation. SecurityLLM [9], a BERT-based system, achieved 98% accuracy in classifying 14 attack types. Paul et al. [10] proposed LLM-integrated RAG pipelines that leverage continuous threat intelligence for adaptive response. Shao et al. introduced the *NYU CTF Bench*, while Gandhi et al. [11] developed SHIELD, combining anomaly detection with LLM reasoning for APT detection. RAG, introduced by Lewis et al. [4], bridges the knowledge limitations of LLMs by retrieving external context, enabling factually accurate and task-relevant responses. Gao et al. [12] surveyed the evolution of RAG and emphasized its potential in knowledge-intensive domains. Recent innovations include CyRAG and GraphCyRA [13], which fuse LLMs with structured databases and knowledge graphs to improve the granularity of cyber threat analysis. GraphCyRA, in particular, excels at identifying hidden relationships between attack vectors and suggesting mitigation priorities. VUL-RAG [14] addresses the misclassification of patched code as vulnerabilities, while Rajapaksha et al. [15] designed a RAG-based QA system tailored for cyber-attack attribution, showing enhanced results when few-shot examples are included. Despite these advancements, existing RAG systems typically treat all retrieved documents equally or rank them solely by vector similarity, which may dilute the contextual relevance of the final output. Our WRRF-enhanced framework addresses this gap by incorporating retriever-side confidence scores into the ranking process. This allows our system to elevate documents with high semantic relevance, thereby producing mitigation strategies that are both precise and grounded in expert knowledge—advancing the practical deployment of LLMs in high-stakes cybersecurity environments.

III. Proposed Framework and Methodologies

This section presents the architecture of our Explanation-Aware RAG framework for cyber threat mitigation, enhanced with Confidence-Weighted Reciprocal Rank Fusion (WRRF). As shown in Figure 1, the system proceeds through attack detection, explanation-aware prompt generation, query construction, expert retrieval with WRRF, and mitigation with evaluation. Each component is detailed in the following subsections.

A. Attack Classifier

We first train a Random Forest (RF) classifier for attack detection and categorization using the *UNSWNB15* [16] datasets include features such as time-to-live values, byte load, connection counts, and packet rates, which capture crucial aspects of malicious traffic patterns. Before training, we apply standard preprocessing: numeric features are scaled (e.g., standardized to zero mean and unit variance) and categorical features (such as *proto* for protocol type) are encoded into numeric form. We then train a RF classifier with $T = 100$ decision trees (estimators) on the processed feature set. Each decision tree h_t in the forest produces a predicted class for an input feature vector $\mathbf{x}$ (one of the nine attack categories or normal). The RF aggregates these predictions via majority voting to output the final predicted label $\hat{y}$.

B. Knowledge Embedding for Retrival

To ground the system's responses in expert knowledge, we construct an authoritative knowledge corpus by combining guidelines from the ENISA, the NIST, and the CISA. These documents collectively encapsulate best practices for cyber threat mitigation. We merge the documents into one comprehensive text and then segment it into overlapping chunks of 750 characters (with an overlap of 100 characters) to ensure continuity of context between chunks. Each chunk d_i of the knowledge corpus is then transformed into a high-dimensional vector representation $\mathbf{z}_i$ using OpenAI's text embedding model (accessed via LangChain's *OpenAIEmbeddings*). This embedding function $E(\cdot)$ maps a text string to a d-dimensional

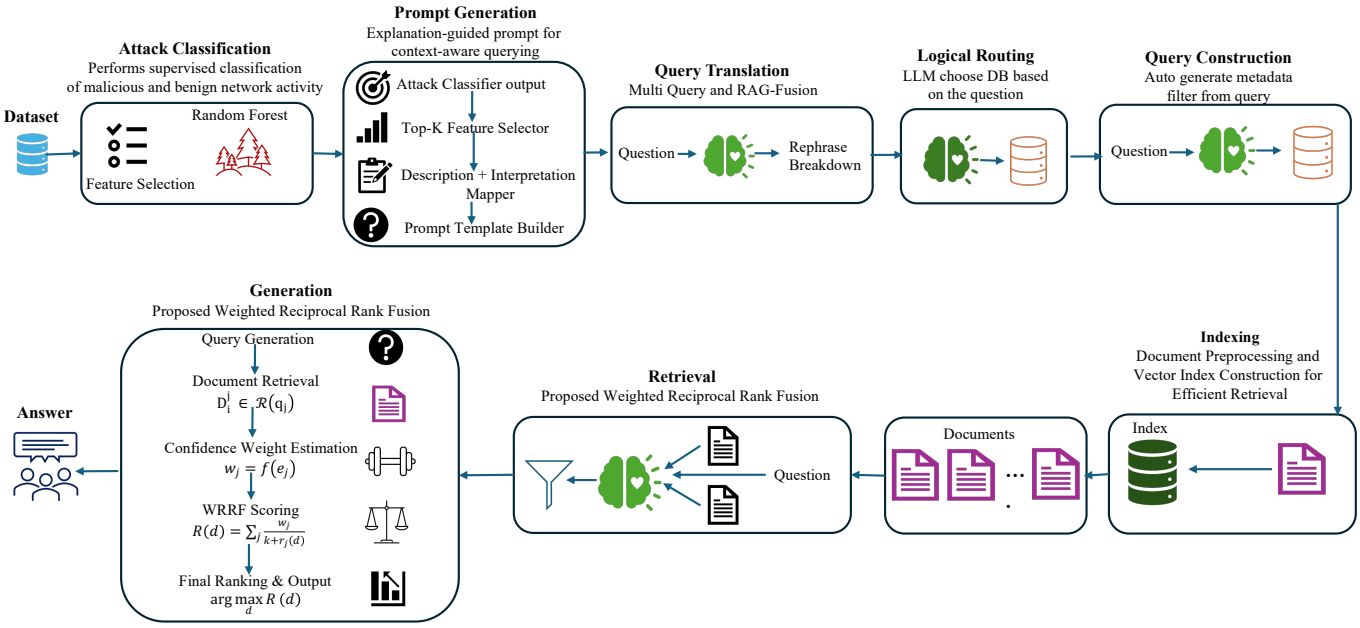

Fig. 1. The proposed explanation-aware RAG framework integrates attack classification, prompt generation, and response generation.

numerical vector in semantic space. In our implementation, $d = 1536$, as we use the `text-embedding-ada-002` model. Formally, for a given text chunk d_i, we obtain its vector embedding $\mathbf{z}_i = E(d_i) \in \mathbb{R}^d$. All such vectors $\mathbf{z}_i$ are stored in a vector database built with Facebook AI Similarity Search (FAISS), an open-source library for efficient similarity search on large collections of vectors. The vector database allows semantic retrieval: given any query vector, it can quickly find the stored vectors most similar to it. By using vector embeddings of domain-specific text, our system can retrieve relevant expert knowledge based on semantic content, rather than simple keyword matching.

C. Query Construction

When the classifier flags a record as malicious, we construct a multi-step, explanation-aware prompt for the LLM using abnormal features and retrieved context. For each record x, we select the top-k influential features $\mathcal{F}\text{top}(x)$ (with $k = 5$) based on feature importance or domain knowledge. For each feature $f \in \mathcal{F}\text{top}(x)$ with value v_f, we generate a one-line natural-language explanation L_f, combining its description $D(f)$ and interpretation $I(f)$ (e.g., $D(sload)$ = "Source bytes per second"; $I(sload)$ = "unusually high traffic rate, possibly DoS"). These explanations are embedded into a structured LangChain *PromptTemplate*, which introduces the scenario, lists the k feature-based observations, and ends with a direct question prompting the LLM to suggest mitigation strategies.

D. Semantic Retrieval and Response Generation using RAG

The explanation-rich prompt P from subsec:promptgen is fed into the RAG process. P is embedded as $\mathbf{q} = E(P)$ and used to query the FAISS store, which returns the top-K chunks $d_{q,1}, \ldots, d_{q,K}$ with highest cosine similarity. These form the context set C_q. We then employ a RetrievalQA chain (via LangChain) to generate the answer using an LLM, conditioned on both the prompt and the retrieved context C_q. In our setup, the chain constructs a new query to the LLM that includes the original question (mitigation advice) along with the additional context of the retrieved documents (typically by prepending the context passages or injecting them into a system prompt). We use OpenAI's GPT series LLM to produce the final response. Essentially, the LLM is asked to answer the question: "What specific mitigation strategies should be applied…?" while being provided with relevant excerpts from ENISA/NIST/CISA guidelines that were fetched in the previous step. The language model then generates a grounded answer, i.e., an answer that directly incorporates and references the domain knowledge from the context. This RAG approach ensures that the mitigation strategies suggested by the LLM are not just plausible-sounding, but are aligned with vetted cybersecurity best practices. The output is an explanation-aware, context-specific mitigation plan for the detected threat.

E. Multi-Query Retrieval Augmentation

To improve the semantic coverage of retrieved context documents, we incorporate a multi-perspective query expansion strategy using LangChain's MultiQueryRetriever. Rather than relying on a single query derived from the explanation-aware prompt, we generate multiple paraphrased versions of the input question to diversify retrieval. This is particularly helpful in overcoming limitations associated with vector similarity search, such as lexical sparsity and rigid embedding locality.

Formally, for a given user query q (in our case, a dynamically generated prompt describing abnormal network behavior

and attack type), we generate five semantically distinct reformulations $\{q_1, q_2, \ldots, q_5\}$ using a prompt template applied to a ChatGPT model. This prompt explicitly asks the LLM to rephrase the original question from different perspectives to aid document retrieval. Each reformulated query q_i is independently embedded and sent to the FAISS vector store for retrieval. The corresponding top-k documents $\mathcal{D}_i$ are collected for each query. We define the complete context set as the unique union of all retrieved chunks with duplicates removed via serialization and hashing to ensure consistency. The union set $\mathcal{D}_{\text{union}}$ is used as context for the final RAG chain, which performs multi-query retrieval, explanation-aware question construction, and LLM-based response generation. This approach significantly improves diversity and coverage of mitigation strategies.

F. Proposed Weighted Reciprocal Rank Fusion

To address the limitations of equal-weight reciprocal rank fusion in multi-query RAG pipelines, we propose a *Confidence-Weighted Reciprocal Rank Fusion (WRRF)* mechanism that enhances retrieval quality by integrating document-level confidence scores into the fusion process. In standard RAG-Fusion, multiple sub-queries are generated from the original question using a large language model (LLM), and for each query q_i, a set of top-k documents is retrieved and ranked based on vector similarity. These documents are then fused using *Reciprocal Rank Fusion (RRF)*, which assigns each document d a cumulative score defined as:

$$\text{RRF}(d) = \sum_{i=1}^{N} \frac{1}{r_{i,d} + k}, \tag{1}$$

where $r_{i,d}$ denotes the rank of document d in the retrieval result for query q_i, N is the total number of generated queries, and k is a smoothing constant (typically $k = 60$) that diminishes the influence of lower-ranked documents. However, this formulation treats all retrieved documents with equal weight, regardless of how strongly the retriever believes a document is relevant to its respective query. This can lead to degraded performance in real-world settings where query quality or document relevance varies significantly. In contrast, the proposed WRRF method augments the RRF formula by incorporating a *confidence score* $c_{i,d} \in [0, 1]$ for each document d retrieved under query q_i, resulting in the following scoring function:

$$\text{WRRF}(d) = \sum_{i=1}^{N} \frac{c_{i,d}}{r_{i,d} + k}. \tag{2}$$

Here, $c_{i,d}$ reflects the retriever's confidence in the semantic relevance of document d to query q_i. These confidence scores are computed using a *min-max normalization* of raw retrieval similarity scores (e.g., cosine similarity for dense retrievers or BM25 scores for lexical retrievers). Specifically, for a given query q_i, if the set of raw similarity scores for the top-k documents is $\mathbf{s}_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,k}\}$, then we normalize

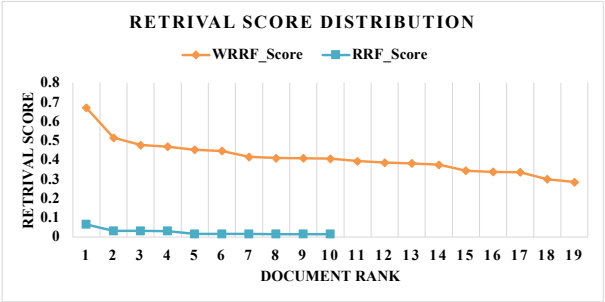

Fig. 2. Retrieval score distribution for WRRF and RRF.

retriever scores to [0,1]. This ensures that the most confidently retrieved document receives a confidence score close to 1, while less relevant documents receive proportionally lower weights. By multiplying the reciprocal rank term with this normalized confidence, WRRF naturally prioritizes documents that are both highly ranked and confidently retrieved. The full pipeline proceeds as follows: an LLM generates multiple diverse sub-queries from the input question; each query is independently passed through a retriever that outputs ranked documents and similarity scores; scores are normalized to produce per-query confidence values; the WRRF score is computed for all retrieved documents, which are then merged, deduplicated, and sorted by their final score. The top-m ranked documents form the context passed to the LLM, along with the original question, to generate the final answer. This modification leads to a more robust and adaptive reranking mechanism that outperforms standard RAG-Fusion in domains where document relevance is highly variable, such as cybersecurity or technical compliance.

IV. Experiment and Result Analysis

This section evaluates five response-generation methods—OpenAI (OAI), RAG, Multi-Query (MQ), RRF, and our proposed WRRF—on their ability to generate context-aware mitigation strategies for suspicious network traffic.

A. Retrieval Score Distribution and Ranking Resolution

We begin by analyzing the retrieval quality of WRRF compared to RRF through score distribution and rank separation. Figure 2 illustrates the distribution of retrieval scores across ranked documents generated by two methods: the standard Reciprocal Rank Fusion (RRF) and the proposed Weighted Reciprocal Rank Fusion (WRRF). The x-axis represents the document rank, while the y-axis denotes the corresponding retrieval score assigned by each method. The WRRF curve exhibits a significantly higher starting score (0.6712 at rank 1) and a gradual decline across ranks, demonstrating its strong ability to discriminate between highly relevant and less relevant documents. In contrast, the RRF scores are tightly clustered within a narrow range (0.0159 to 0.0667), producing a nearly flat curve that suggests limited differentiation among top-ranked documents. This compression in RRF indicates weak rank confidence and hampers its ability to prioritize truly

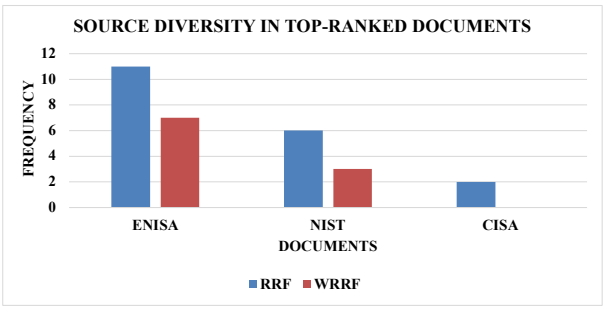

Fig. 3. Source diversity

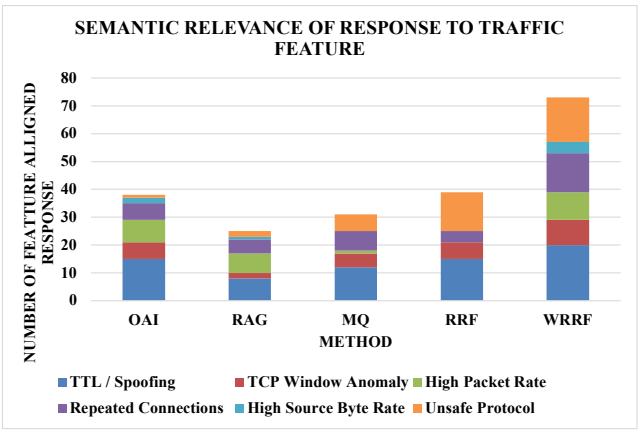

Fig. 4. Semantic relevance of responses.

relevant content. WRRF's broader score spread and steeper initial drop are indicative of a robust ranking strategy that leverages confidence-weighted fusion to elevate documents with stronger semantic alignment and retrieval signals. Consequently, WRRF provides better separation between high-value and low-value documents, which is critical for evidence selection in retrieval-augmented systems. The plot supports the claim that WRRF offers superior ranking resolution and relevance awareness compared to RRF.

B. Document Source Diversity Analysis

To further evaluate the effectiveness of the WRRF ranking method, we analyzed the distribution of authoritative document sources—*ENISA*, *NIST*, and *CISA*—across the top-ranked documents retrieved by WRRF and RRF. As shown in Figure 3, WRRF achieves significantly broader source diversity compared to RRF. While both methods rank *ENISA.pdf* highly, RRF exhibits redundancy—repeating the same document multiple times in its top-10 list. In contrast, WRRF not only identifies high-confidence documents from ENISA, but also integrates content from *NIST.pdf* and *CISA.pdf*, thereby enhancing evidence variety. This increased heterogeneity is crucial for robust cyber-threat mitigation, as it provides a more comprehensive and multi-perspective understanding of best practices. Moreover, the presence of *CISA.pdf* in WRRF's ranking—absent entirely in RRF—demonstrates WRRF's ability to surface authoritative but non-obvious sources. These findings highlight that WRRF does not rely solely on term overlap or lexical similarity but leverages semantic relevance and confidence weighting to promote expert-validated content across multiple institutional domains.

C. Semantic Relevance to Observed Traffic Features

To assess the contextual quality of the mitigation responses, we evaluated the degree to which each model's output aligns with key indicators extracted from the suspicious network traffic. The observed traffic characteristics—such as abnormal TTL values, no payload, protocol misuse, repeated connections, and high packet rates—are highly indicative of spoofed DoS or reconnaissance behavior. Accordingly, we defined six traffic-feature themes and identified a set of relevant mitigation keywords for each, including terms such as *spoofing*, *packet inspection*, *rate limiting*, and *protocol filtering*. We then performed a semantic matching analysis on each response, counting how many times the model addressed these traffic-aligned features across all scenarios. The results are visualized in Figure 4, using a stacked bar chart to illustrate per-method alignment with each feature.

TABLE I
MODEL RESPONSIVENESS TO TRAFFIC FEATURES. ✓= DIRECT MATCH, != PARTIAL MATCH, ✗= NONE.

Traffic Feature	OAI	RAG	MQ	RRF	WRRF
High TTL / Spoofing	✗	✗	!	!	✓
Zero-byte Traffic	✗	✗	!	!	✓
Unsafe Protocol	✗	✗	✓	!	✓
Repeated Connections	✗	✗	✓	✓	✓
High Packet Rate	!	✗	✓	✓	✓
TCP Window Size Anomaly	✗	✗	!	✗	✓
High Source Byte Rate	✗	✗	!	!	✓
Total Responsive	1/7	0/7	5/7	4/7	**7/7**

As shown in the figure, the proposed WRRF model consistently demonstrates the highest overall semantic alignment to the underlying traffic characteristics. WRRF outperforms all baseline models in categories such as *TCP Window Anomaly* (9 mentions), *Repeated Connections* (7 mentions), and *High Packet Rate* (6 mentions), while maintaining strong coverage in *TTL/Spoofing* and *Protocol Filtering*. This indicates that WRRF not only captures relevant mitigation categories but also tailors them to the specific threat indicators observed. In contrast, while the Multi-Query and RRF baselines show high relevance in select categories (e.g., *TTL/Spoofing*), they suffer from imbalanced coverage, failing to address anomalies such as *Source Byte Rate* or *TCP anomalies*. OpenAI and RAG models exhibit broad but shallow coverage, frequently defaulting to general advice rather than addressing packet-level or session-specific anomalies. This analysis highlights the core strength of the WRRF model: its ability to synthesize context-aware, actionable strategies that directly correspond to the observed network anomalies. By tightly mapping response content to threat indicators, WRRF enhances both the precision and operational utility of automated cyber-threat mitigation.

D. Mitigation Strategy Adaptation

We evaluate how five different response-generation methods—OpenAI, RAG, Multi-Query, RRF, and the proposed WRRF—adapt their mitigation strategies in response to a complex and potentially malicious network traffic scenario. We analyzed a sample where network features suggest the presence of a spoofed Denial-of-Service (DoS) or scanning attack, characterized by (i) abnormal Time-To-Live (TTL) values (destination TTL = 0, source TTL = 254), (ii) unsafe protocol usage ("unas"), (iii) zero-byte data transfers, (iv) source TCP window size of 0, (v) high packet rate exceeding 125,000 packets per second, (vi) repeated connection attempts to the same service, and (vii) a very high source byte transmission rate (100MBps). Collectively, these indicatorsss imply malicious intent with possible spoofing, resource exhaustion, and reconnaissance objectives. We analyze all responses where OpenAI and RAG responses were largely generic, consisting of static recommendations such as IP blocking or procedural steps such as contacting ISPs. These methods did not contextualize the unusual TTL values or recognize indicators of spoofed traffic or volumetric abuse. As a result, they lacked specific countermeasures for time-sensitive and packet-level anomalies. In Multi-Query, by leveraging diverse sub-prompts, demonstrated partial contextual alignment. It identified issues such as repeated service access and suggested practical defenses like rate limiting and firewall tuning. However, it did not tightly associate protocol-level characteristics (e.g., "unas" protocol, TTL values) with actionable mitigation.

In RRF, responses showed broader strategy coverage due to multi-query fusion, but the approach suffered from equal weighting of all queries, causing dilution of high-relevance cues. For example, high TTL values and zero-byte traffic were occasionally mentioned but not consistently linked to spoofing or scanning. In contrast, the proposed WRRF method achieved the highest degree of traffic-aware mitigation. By incorporating confidence-weighted reciprocal rank fusion, the WRRF prioritized and emphasized key anomalous indicators in the final response. It accurately associated a TTL of 254 with potential spoofing behavior and recommended packet inspection or anti-spoofing filters. Similarly, it identified zero-byte transfers and excessive packet rate as signs of DoS and advised layered defenses including Intrusion Detection/Prevention Systems (IDPS), DDoS mitigation services, and traffic shaping. In addition, WRRF uniquely flagged protocol anomalies and suggested protocol-based filtering strategies. Table I shows how methods respond to traffic features. WRRF covers all seven indicators, while Multi-Query and RRF address five and four. OpenAI and RAG perform worst with minimal contextual awareness. This highlights WRRF's broader coverage and nuanced feature-aware responses, making it well-suited for adaptive cybersecurity.

V. Conclusion

We propose an explanation-aware RAG framework for cyber-threat mitigation that integrates attack classification, semantic prompt construction, and expert-informed retrieval. Our key contribution is Weighted Reciprocal Rank Fusion (WRRF), which incorporates retriever confidence into document ranking for improved relevance. Experiments on UNSW-NB15 and cybersecurity corpora show WRRF outperforms OpenAI, standard RAG, multi-query, and RRF in semantic relevance, strategy diversity, source heterogeneity, and precision. Future work includes fine-tuning lightweight domain LLMs and adding human-in-the-loop feedback for more precise and trustworthy cyber defense.

References

[1] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Najafi, F. Cheng, and C. Meinel, "Large language models in cybersecurity: State-of-the-art," *arXiv preprint arXiv:2402.00891*, 2024.

[2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[3] S. M. Xie and S. Min, "How does in-context learning work? a framework for understanding the differences from traditional supervised learning," *A framework for understanding the differences from traditional supervised learning*, 2022.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[5] Z. Rackauckas, "Rag-fusion: a new take on retrieval-augmented generation," *arXiv preprint arXiv:2402.03367*, 2024.

[6] M. A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, and N. Tihanyi, "Generative ai and large language models for cyber security: All insights you need," *Available at SSRN 4853709*, 2024.

[7] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: A systematic literature review," *arXiv preprint arXiv:2405.04760*, 2024.

[8] Y. Yang, B. Xu, X. Gao, and H. Sun, "Context-enhanced vulnerability detection based on large language model," *arXiv preprint arXiv:2504.16877*, 2025.

[9] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, and T. Lestable, "Revolutionizing cyber threat detection with large language models," *arXiv preprint arXiv:2306.14263*, pp. 195–202, 2023.

[10] S. Paul, F. Alemi, and R. Macwan, "Llm-assisted proactive threat intelligence for automated reasoning," *arXiv preprint arXiv:2504.00428*, 2025.

[11] P. A. Gandhi, P. N. Wudali, Y. Amaru, Y. Elovici, and A. Shabtai, "Shield: Apt detection and intelligent explanation using llm," *arXiv preprint arXiv:2502.02342*, 2025.

[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-augmented generation for large models: A survey," *ArXiv*, vol. abs/2312.10997, 2023.

[13] M. Rahman, K. O. Piryani, A. M. Sanchez, S. Munikoti, L. De La Torre, M. S. Levin, M. Akbar, M. Hossain, M. Hasan, and M. Halappanavar, "Retrieval augmented generation for robust cyber defense," Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), Tech. Rep., 2024.

[14] X. Du, G. Zheng, K. Wang, J. Feng, W. Deng, M. Liu, B. Chen, X. Peng, T. Ma, and Y. Lou, "Vul-rag: Enhancing llm-based vulnerability detection via knowledge-level rag," *arXiv preprint arXiv:2406.11147*, 2024.

[15] S. Rajapaksha, R. Rani, and E. Karafili, "A rag-based question-answering solution for cyber-attack investigation and attribution," in *European Symposium on Research in Computer Security*. Springer, 2024, pp. 238–256.

[16] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.