Toward a Synergistic Vulnerability Analysis Enhanced with a Multi-Level RAG Model

Min-Joo Yoon
Amgine
SysCore Lab.,
Sejong University
Seoul 05836, South Korea
minjoo1658@gmail.com

Sun-Mo Yoo
Amgine
Seoul 05836, South Korea
smyoo@amgine.co.kr

Jong-Hwa Park
Amgine
Seoul 05836, South Korea
jhpark@amgine.co.kr

Ki-Woong Park*
SysCore Lab.,
Sejong University
Seoul 05006, South Korea
woongbak@sejong.ac.kr

***Abstract*— This paper presents a vulnerability analysis framework that has been enhanced with a multi-level Retrieval-Augmented Generation (RAG) model. The principal objective of this research is to develop a framework that will facilitate a synergistic vulnerability analysis enhanced with a multi-level RAG model. To enhance the analysis of Common Vulnerabilities and Exposures (CVE), the model progressively integrates multiple vulnerability databases, including the National Vulnerability Database (NVD), CERT Vulnerability Notes Database (VNDB), GitHub Advisory Database, and Exploit Database. By employing the RAG framework, the model retrieves pertinent information from these diverse sources and augments the analysis with accurate and up-to-date data, effectively addressing the limitations of existing fine-tuned language models. A comprehensive discussion about challenges and future works based on the model design was provided, thereby contributing valuable insights to the fields of cybersecurity and vulnerability analysis.**

Keywords—Large Language Model, Retrieval-Augmented Generation, Common Vulnerabilities and Exposures

I. INTRODUCTION

In this paper, we introduce a multi-level Retrieval-Augmented Generation (RAG) based model aimed at enhancing the quality of vulnerability analysis through the progressive integration of multiple vulnerability databases. By incorporating data from the National Vulnerability Database (NVD), CERT Vulnerability Notes Database (VNDB), GitHub Advisory Database, and Exploit Database, we anticipate that this model will provide a more comprehensive and up-to-date assessment of vulnerabilities [1], [2], [3], [4]. Besides, we have designed a hybrid search system that utilizes both exact matching and top-k retrieval when a CVE-ID is included in the query. By combining the benefits of text-based search with precise CVE-ID retrieval, our approach is expected to significantly improve the accuracy and depth of vulnerability analysis, facilitating more effective identification and management of security risks.

II. RELATED WORK

Large Language Models (LLMs) have emerged as promising tools for vulnerability detection and analysis in

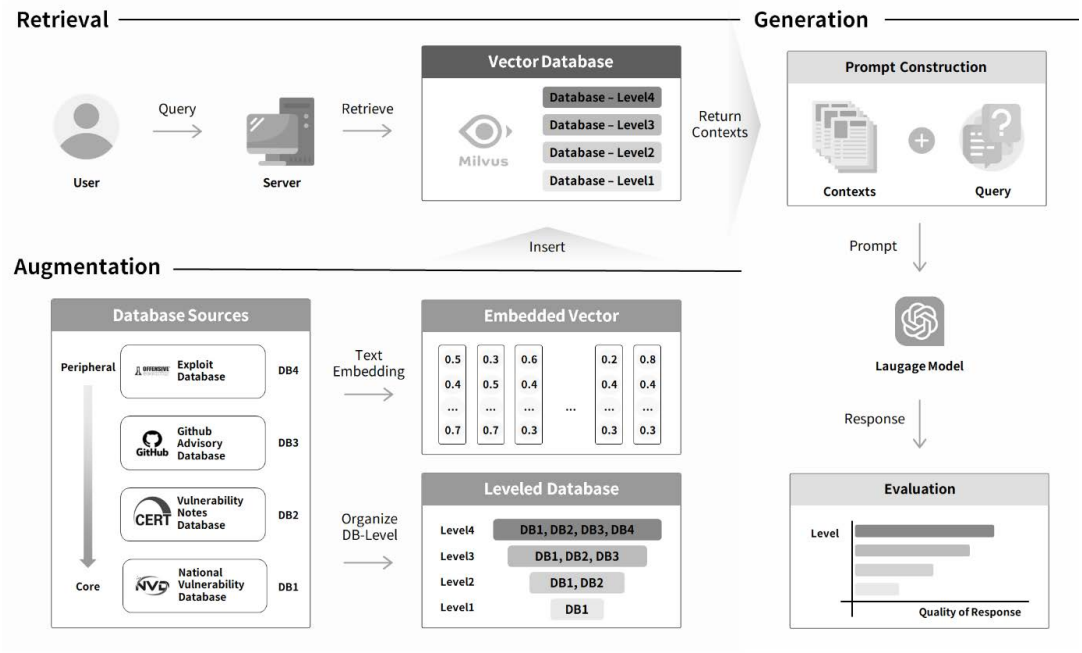

Fig. 1. Multi-Level RAG-based Vulnerability Analysis Model

2831-7203/24/$31.00 ©2024 IEEE
DOI 10.1109/AIxDKE63520.2024.00033

software systems [5]. The IRIS framework integrates LLMs with static analysis to uncover vulnerabilities in complex Java projects that traditional tools miss, while PrimeVul highlights data quality issues in vulnerability datasets that hinder model reliability, underscoring the importance of robust datasets [6], [7]. Vul-RAG improves vulnerability classification and remediation accuracy by leveraging a retrieval-augmented generation (RAG) framework, achieving a 12.96% increase in classification accuracy [8]. RAGAS further enhances this field by offering an evaluation approach that assesses RAG systems without predefined answers, focusing on retrieval quality and response relevance [9].

III. MODEL DESIGN

The architecture of the model is designed in Fig. 1. to perform in-depth vulnerability analysis by organizing a multi-level database approach. It consists of three main components like existing RAG-based models: Retrieval, Augmentation, and Generation. In the retrieval phase, the user initiates a query through the server, which retrieves relevant information from a Vector Database. CVE-ID is extracted from query during the search process to ensure more accurate search results and generation results. Then the server interfaces with Milvus, a vector database for document retrieval, to execute this operation and returns contextual information that is pertinent to the user's query. To validate this approach, we collected 4 types of vulnerability datasets, shown in Table I.

In addition, hybrid approach is employed to combine Exact Match and Top K Match techniques illustrated in Fig. 2. Utilizing the CVE-ID as a unique identifier, the model conducts searches to improve contextual retrieval performance significantly. When a CVE-ID is present, the system prioritizes exact matches, ensuring precise results. Conversely, in scenarios where the CVE-ID is absent, the Top K Match method maintains performance levels by retrieving the most relevant documents based on vector operations. This dual strategy enhances the overall effectiveness of context retrieval, allowing for a more robust response to a diverse range of queries.

TABLE I. COLLECTION OF VULNERABILITY DATA BY SOURCE

Source	Institution	CVE	Non-CVE	Total
NVD	NIST	263,522	0	263,522
VNDB	CERT	2,824	773	3,597
Advisory Database	GitHub	17,823	2,264	20,087
ExploitDB	Offensive Security	26,917	19,654	46,571

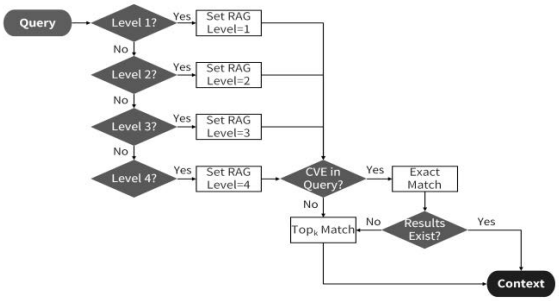

Fig. 2. Hybrid Search process for Optimal Retrival Performance

IV. DISCUSSION

In this section, we highlight the importance of integrating diverse vulnerability data across various channels, a critical step towards developing a comprehensive RAG-based framework. Future efforts will focus on the optimal utilization of multi-level databases to harness the unique strengths of each data source, thereby deepening vulnerability assessments. Additionally, we plan to refine the composition of benchmark datasets, which is vital for accurately evaluating our model's performance. By incorporating diverse data sources and building well-defined datasets, we aim to enhance our framework's ability to address vulnerability scenarios.

V. CONCLUSION

In conclusion, this paper introduces a robust multi-level Retrieval-Augmented Generation (RAG) framework designed to enhance vulnerability analysis by integrating diverse data sources, including the National Vulnerability Database, CERT Vulnerability Notes Database, GitHub Advisory Database, and Exploit Database. The proposed model not only improves the accuracy and depth of vulnerability assessments through a hybrid search system that employs both exact matching and top-k retrieval but also addresses the limitations of existing models by leveraging comprehensive and up-to-date information. Future work will focus on optimizing the utilization of multi-level databases and refining benchmark datasets to further enhance the framework's ability to address complex vulnerability scenarios.

ACKNOWLEDGMENT

This research was supported by the Future Challenge Defense Technology Research and Development Project (9150921) hosted by the Agency for Defense Development Institute in 2023.

REFERENCES

[1] "NVD - Home." Accessed: Sep. 28, 2024. [Online]. Available: https://nvd.nist.gov/

[2] "CERT Coordination Center." Accessed: Sep. 29, 2024. [Online]. Available: https://www.kb.cert.org

[3] "GitHub Advisory Database," GitHub. Accessed: Sep. 29, 2024. [Online]. Available: https://github.com/advisories

[4] "OffSec's Exploit Database Archive." Accessed: Sep. 29, 2024. [Online]. Available: https://www.exploit-db.com/

[5] V. Akuthota, R. Kasula, S. T. Sumona, M. Mohiuddin, M. T. Reza, and M. M. Rahman, "Vulnerability Detection and Monitoring Using LLM," in *2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, Jan. 2023, pp. 309–314. doi: 10.1109/WIECON-ECE60392.2023.10456393.

[6] Z. Li, S. Dutta, and M. Naik, "LLM-Assisted Static Analysis for Detecting Security Vulnerabilities," May 27, 2024, *arXiv*: arXiv:2405.17238. doi: 10.48550/arXiv.2405.17238.

[7] Y. Ding *et al.*, "Vulnerability Detection with Code Language Models: How Far Are We?," Jul. 10, 2024, *arXiv*: arXiv:2403.18624. doi: 10.48550/arXiv.2403.18624.

[8] X. Du *et al.*, "Vul-RAG: Enhancing LLM-based Vulnerability Detection via Knowledge-level RAG," Jun. 19, 2024, *arXiv*: arXiv:2406.11147. doi: 10.48550/arXiv.2406.11147.

[9] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation".

[10] Y. Han, C. Liu, and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," Oct. 18, 2023, arXiv: arXiv:2310.11703. doi: 10.48550/arXiv.2310.11703.