International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

CYBERVIDYA: Rag Infused Cyber Solutions

Ms. Reshma Owhal[1], Viraj Shewale[2], Aniket Sorate[3], Mayur Swami[4], Dipak Waghmode[5]
[1,2,3,4,5]Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune 411001, India.
Emails: *reshma.owhal@aissmsioit.org[1],virajshewale10@gmail.com[2],aniketsorate514@gmail.com[3], mayurswami511@gmail.com[4],dipakdw2003@gmail.com[5]*

Abstract
As cyber threats grow more sophisticated, the need for intelligent, adaptive security solutions has never been greater. CyberVidya: RAG-Infused Cyber Solutions offers a groundbreaking approach by integrating large language models (LLMs) with retrieval-augmented generation (RAG) to provide precise, real-time cybersecurity insights. Unlike traditional models that rely on static knowledge, CyberVidya continuously retrieves and processes information from a dynamic, indexed database of academic papers, ethical books, PDFs, and real-world case studies. What sets CyberVidya apart is its Non-Parametric Knowledge Retrieval, which ensures that responses are contextually accurate and directly sourced from trusted materials. Its multidimensional query-optimized retrievers work alongside advanced LLMs—GPT-2, Mistral-7B, and Llama 3.2-3B—to generate reliable, actionable insights. By incorporating document embedding and Dense Passage Retrieval (DPR), CyberVidya enhances accuracy while adapting to the ever-changing cybersecurity landscape without the need for retraining. The results speak for themselves. CyberVidya consistently outperforms industry-leading models, achieving 92.86% relevance, 85.81% similarity, and 95.06% correctness in educational queries. For scenario-based cybersecurity challenges, it maintains high performance with 92.89% relevance, 89.56% similarity, and 93.94% correctness. Comparative studies further highlight that RAG-based models surpass traditional LLMs in understanding complex cybersecurity concepts such as tactics, techniques, and procedures (TTPs) With its ability to provide accurate, real-time cybersecurity guidance, CyberVidya stands as a powerful tool for individuals, enterprises, and educators, bridging the gap between static knowledge and dynamic problem-solving.

Keywords: *Cybersecurity solutions; Knowledge retrieval; Large language models; Retrieval-augmented generation; Threat mitigation.*

1. Introduction
In today's digital era, the rapid evolution of cyber threats has made cybersecurity an essential aspect of safeguarding personal, organizational, and governmental digital assets. The increasing sophistication of attacks, ranging from phishing schemes to advanced persistent threats, necessitates proactive measures that empower individuals and organizations to respond effectively.[4,5] However, the lack of accessible, dynamic, and accurate tools for cybersecurity guidance has left many users vulnerable to these threats. Traditional approaches to cybersecurity education, such as static guides or generic chatbots, often fail to address the nuanced and contextual nature of real-world attacks[8]. CyberVidya: RAG-Infused Cyber Solutions uses cutting-edge machine learning and natural language processing (NLP) technologies to close this gap. CyberVidya, which is based on the ideas of Retrieval-Augmented Generation (RAG), combines cutting-edge transformer models with a comprehensive knowledge store to deliver real-time, context-aware advice that is customized to each user's unique cybersecurity requirements. Through the use of retriever-generator pipelines, CyberVidya produces meaningful answers to a variety of questions by dynamically retrieving pertinent data from a carefully selected and regularly updated knowledge base. CyberVidya uses non-parametric knowledge retrieval, which allows it to adjust to new cyberthreats without requiring a lot of retraining, in contrast to conventional chatbots that only use parametric knowledge or predetermined datasets.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

Regardless of whether they are addressing simple vulnerabilities or complex cyberattack tactics, methods, and procedures (TTPs), this flexibility guarantees that users get the most pertinent and accurate guidance [3]. CyberVidya is made to scale its applications across domains, which makes it a useful tool for businesses and training platforms looking to strengthen their cybersecurity procedures in addition to individuals. Models such as Mistral-7B and refined retrievers are integrated into the system to guarantee accurate and useful responses. CyberVidya's combination of cutting-edge NLP technology and a vast knowledge base makes it a unique solution to the intricate problems presented by the always changing cybersecurity environment [14,15]. By combining modular reasoning, explanation, and summarizing techniques, the MoRSE (Modular Reasoning, Explanation, and summarizing Engine) framework was created to improve AI-driven question-answering systems. In order to provide organized, logical, and explicable answers, the authors developed a pipeline-based architecture that breaks down complicated questions into several reasoning processes. By not only offering solutions but also deconstructing the logic behind them, MoRSE improves transparency and makes it simpler for users to comprehend how conclusions are reached. Instead of depending only on parametric knowledge, the system may dynamically retrieve pertinent contextual information thanks to the framework's integration of retrieval-based reasoning. MoRSE also has a summary module that improves readability and user engagement by distilling collected information into succinct but insightful comments. To improve its responses while preserving logical flow, the model makes use of natural language processing (NLP) approaches. MoRSE enhances the precision and dependability of AI-generated information through the use of multi-step query decomposition and knowledge retrieval. This strategy is especially helpful in fields that need a high degree of precision, like cybersecurity, healthcare, and law, where explainability is essential. Because MoRSE is organized, it can be used to a wide range of real-world scenarios, which increases confidence in AI-powered systems. By providing superior interpretability and accuracy in processing complicated queries, the paper shows that MoRSE performs better than traditional question-answering models [1]. Facebook AI Research (FAIR) developed the Retrieval-Augmented production (RAG) model to enhance natural language production by combining transformer-based generative models with real-time document retrieval [11]. Before producing responses, RAG dynamically collects pertinent documents from an external knowledge source, in contrast to conventional big language models that only use static parametric knowledge. In order to give more factually correct and contextually aware responses, the model combines user queries with semantically similar documents that are retrieved using Dense Passage Retrieval (DPR) [6,7]. The trustworthiness of AI-generated information is greatly increased by this retrieval-generation hybrid technique, which makes it very useful for open-domain question-answering jobs.RAG-Token retrieves documents and bases each token generation on outside information, while RAG-Sequence retrieves materials only once and produces responses in a comprehensive manner. The experimental findings showed that RAG performs better in factual consistency and knowledge recall than conventional generative models. Applications for this architecture are numerous and include domain-specific AI systems, research support, and customer service. RAG provides a potent way to enhance AI-driven question answering, lessen hallucination problems, and increase the reliability of AI-generated responses by bridging the gap between retrieval and creation [2]. Within the MITRE ATT&CK framework, the research article "Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation" examines the difficulties in deciphering cybersecurity tactics, techniques, and procedures (TTPs). Since it takes a great deal of experience to comprehend various attack techniques, automation is a useful tool. The study examines the efficacy of two categories of LLMs: decoder-only models (like GPT-3.5) improved with Retrieval-Augmented Generation (RAG) and encoder-only models (like RoBERTa) optimized by supervised learning [9,10]. The study

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

also uncovers an unexpected insight: more general prompts yield better results than highly specific ones when predicting cyberattack tactics. By leveraging MITRE ATT&CK data, the paper provides a comparative analysis of these approaches, highlighting practical methods for cybersecurity analysts. The findings emphasize the potential of LLMs to improve TTP analysis while acknowledging their limitations, such as dependency on high-quality retrieval mechanisms and risk of misinformation [3].

1.1. Contribution

CyberVidya introduces several significant advancements in the field of cybersecurity assistance through its innovative use of Retrieval-Augmented Generation (RAG) and state-of-the-art NLP technologies. One of its primary contributions is the dynamic integration of non-parametric and parametric knowledge, enabling the system to retrieve and generate relevant cybersecurity guidance in real time. This approach ensures that users receive precise and actionable responses tailored to their specific queries, bridging the gap left by traditional static resources or generic chatbot systems. By leveraging a structured and extensive knowledge base, which includes curated data from scholarly articles, real-world case studies, and ethical cybersecurity resources, CyberVidya provides a reliable foundation for addressing diverse cybersecurity challenges.

1.2. Background

The development of CyberVidya draws upon advancements in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) frameworks, and cutting-edge NLP tools such as Hugging Face Transformers. These technologies collectively provide the foundation for the system's ability to retrieve, process, and generate actionable cybersecurity insights. Below is an exploration of the background and key terms related to these technologies: -

1.2.1. Data Set Info

For our dataset, we meticulously compiled a diverse and comprehensive collection of resources to ensure the robustness and relevance of CyberVidya's knowledge base. This dataset includes a wide range of ethical hacking books, which provide foundational and advanced knowledge on cybersecurity principles, attack methodologies, and defense mechanisms. To enhance the system's ability to address contemporary threats, we also incorporated real-time threat blogs that document the latest vulnerabilities, attack vectors, and mitigation strategies as they emerge in the cybersecurity landscape. Additionally, we gathered real-world scenario blogs that detail practical case studies, incident reports, and firsthand accounts of cybersecurity breaches and their resolutions. Our dataset's integration of these diverse sources allows CyberVidya to provide context-rich, precise, and proactive solutions, guaranteeing that customers receive current and useful cybersecurity advice catered to actual circumstances.

1.2.2. Large Language Models (LLMs)

Transformer-based systems known as large language models are made to process and produce text that is similar to that of a person by identifying patterns, connections, and context in enormous amounts of data. With applications in a variety of fields, including conversational AI, text summarization, and question answering, they constitute a paradigm change in natural language processing. The foundation for advanced language understanding and generating skills was established by pioneering LLMs such as GPT (Generative Pre-trained Transformer) and its offspring, as well as GPT-3. Mistral-7B, one of CyberVidya's fine-tuned LLMs, is essential for producing accurate, context-aware answers to cybersecurity queries. These models are excellent at deciphering intricate user inputs and producing subtle outputs, which is essential for solving cybersecurity issues.

1.2.3. Retrieval-Augmented Generation (RAG)

A hybrid system called Retrieval-Augmented Generation integrates LLMs with a retrieval mechanism to improve their capabilities. RAG adds an external non-parametric knowledge source, like a document database or knowledge base, in contrast to conventional LLMs, which only use internal parametric knowledge. Two essential elements are involved in the RAG process:

- **Retriever:** Identifies and retrieves relevant documents or information snippets from an

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

indexed knowledge base using techniques like Dense Passage Retrieval (DPR) or BM25.

- **Generator:** Processes the retrieved information along with the user query to generate contextually enriched and accurate responses. By combining retrieval and generation, RAG allows systems like CyberVidya to provide real-time, dynamic responses based on the latest and most relevant data, ensuring adaptability to evolving cybersecurity threats.

By combining retrieval and generation, RAG allows systems like CyberVidya to provide real-time, dynamic responses based on the latest and most relevant data, ensuring adaptability to evolving cybersecurity threats.

1.2.4. Hugging Face Transformers

One of the best open-source libraries for working with transformer designs is Hugging Face Transformers, which offers pre-trained models and tools. It makes it easier to deploy and fine-tune LLMs for a variety of NLP tasks. CyberVidya uses Hugging Face to refine models such as Mistral-7B on cybersecurity datasets that are specific to a certain area. High-quality language generation and improved system efficiency are made possible by its stable API, which makes it easy to integrate transformer models into the RAG framework. The library is a crucial part of contemporary NLP projects because it supports state-of-the-art methods including tokenization, attention mechanisms, and transfer learning.

1.2.5. Definitions of Important Terms

- **Tokenization:** Dividing text into smaller chunks, such words or subwords, so that models can efficiently process and evaluate input.
- **Attention Mechanism:** A core component of transformers, allowing models to focus on relevant parts of the input text to generate contextually accurate outputs.
- **Dense Passage Retrieval (DPR):** A method of retrieval that converts documents and queries into dense vectors and calculates how similar they are to find pertinent information.
- **Embedding:** A numerical representation of text data in a continuous vector space, used by models to analyze semantic relationships between words or phrases.
- **Parametric Knowledge:** Information stored within the model weights of an LLM, learned during pre-training.
- **Non-Parametric Knowledge:** Information stored outside the model, such as in an indexed knowledge base, retrieved dynamically during query processing.
- **Fine-Tuning:** A procedure that optimizes a pre-trained model's performance for a certain task by training it on a specified dataset.

2. Method

CyberVidya is a cutting-edge cybersecurity assistant made to offer precise, timely, and useful advice for reducing online dangers. CyberVidya connects traditional static resources with intelligent, dynamic systems by utilizing the synergy of cutting-edge technologies such as transformer-based language models and Retrieval-Augmented Generation (RAG). In contrast to traditional chatbots, CyberVidya creates responses that are customized for each user's query by dynamically retrieving pertinent material from a carefully selected knowledge base.

A streamlined retriever-generator pipeline, refined transformer models, and a non-parametric knowledge base create the system's solid foundation. Because of its scalability and adaptability, CyberVidya can be used by individuals, businesses, and educational institutions looking to improve their cybersecurity protocols.

CyberVidya Core Workflow: User Input: The process starts when a user submits a question about a cybersecurity issue or challenge. Natural language preprocessing methods like tokenization and embedding creation are used to process the query.

Retriever Module: After receiving the query, the retriever module looks for pertinent documents in the indexed knowledge base, which was constructed using Elastic search.

In order to determine the semantic similarity between the query and the top-ranked documents, the retriever uses Dense Passage Retrieval (DPR).

- **Contextual Augmentation:** To create an enhanced context, the user's query is

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

combined with the pages that were retrieved.

- **Generator Module:** A refined transformer-based language model, like Mistral-7B, receives the richer context.
- **Response Delivery:** The system delivers the generated response to the user in an easily understandable format, often accompanied by additional suggestions or preventive measures.

CyberVidya RAG: CyberVidya employs a Retrieval-Augmented Generation (RAG) architecture to provide real-time, context-aware cybersecurity guidance by integrating retrieval-based search with transformer models. Unlike static chatbots, it dynamically retrieves relevant knowledge from a non-parametric knowledge base, ensuring adaptability to evolving threats.

2.1.Data Ingestion & Knowledge Base Construction

CyberVidya's knowledge base, sourced from CWE datasets, PDFs, and scholarly articles, is indexed using LlamaIndex and LangChain. Documents are processed via PyPDF2 and converted into dense vector embeddings using sentence-transformers (all-mpnet-base-v2), enabling semantic search over traditional keyword-based methods.

2.2.Retrieval Mechanism

Queries are embedded and matched against the VectorStoreIndex using cosine similarity, retrieving the most relevant cybersecurity insights. LangChain enables query orchestration, supporting multi-turn interactions for improved response accuracy. Additionally, BM25 lexical retrieval is used alongside Dense Passage Retrieval (DPR) to enhance precision. CyberVidya can efficiently handle paraphrased or ambiguous searches because it places a higher priority on semantic similarity than precise keyword matching.

2.3.Response Generation & Contextual Augmentation

The query and the retrieved documents are combined, and the refined transformer models (Mistral-7B, Llama) are used to generate the response. Structured, context-aware cybersecurity solutions are guaranteed by Lang Chain's Prompt Template API. transformer models, and a non-parametric knowledge

2.4.Executing Queries and Delivering Responses

- Retrieval: Drawing conclusions about cybersecurity from knowledge that has been indexed Augmentation: Combining the query with the knowledge that has been retrieved.
- Generation: Creating a knowledgeable cybersecurity reaction.
- Output: Providing well-organized responses together with security advice and threat mitigation techniques.

3. Algorithm

Algorithm 1: Context Retrieval and LLM Processing
Require: User query Q
Ensure: Answer A
Predefined: Datasets D1, D2, D3, D4
Procedure EXECUTEMODEL(Q)
1. Initialize context set C ← ∅
2. Function RETRIEVECONTEXT(Q)
1. For each dataset Di in D1, D2, D3, D4
1. Ii ← retrieve information using Di(Q)
2. Sort Ii by relevance scores to find the most pertinent documents
3. Itop ← topk(Ii)
4. If Itop is not empty then
1. C ← C ∪ {Itop}
2. End for
3. If C = ∅ then
1. Return "No relevant information found."
4. Else
1. P ← wrap(C, Q)
2. A ← LLM(P)
5. Return A
End Procedure

4. Results and Discussion

4.1.Results

We thoroughly evaluated CyberVidya against three open-source large language models (LLMs): DeepSeek, Lama, and Mistral in order to determine how well it provides precise and context-aware cybersecurity insights. Three major performance criteria were used in the examination, which concentrated on two types of cybersecurity-related queries: educational questions and scenario-based questions.

- **Relevance:** Indicates how well the model's

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

answers match the purpose of the inquiry.

- **Similarity:** Assesses how closely the generated response and the anticipated response match semantically.
- **Correctness:** Evaluates the response's logical Table 1 shows Evaluation and Comparison , Table 2 shows Evaluation and Comparison (Scenario Based Questions

Table 1 Evaluation and Comparison

	Educational Questions			
	DeepSeek	**Lama**	**Mistral**	**CyberVidya**
Relevance	80.66	89.06	85.93	92.86
Similarity	71.01	85.53	82.66	85.81
Correctness	80.33	90.73	87.41	95.06

Table 2. Evaluation and Comparison (Scenario Based Questions)

	Educational Questions			
	DeepSeek	**Lama**	**Mistral**	**CyberVidya**
Relevance	69.07	89.55	91.63	92.89
Similarity	64.66	84.75	86.61	89.56
Correctness	72.21	90.61	89.31	93.94

Table 1&2 summarizes the findings, which demonstrate CyberVidya's excellent performance in every assessed criterion.

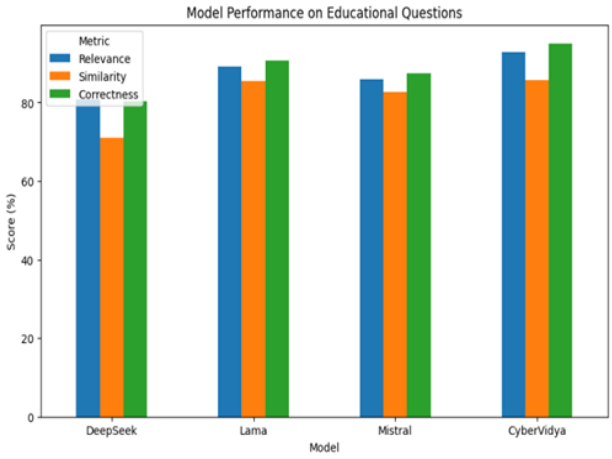

Figure 1 Model vs Score (Educational Questions)

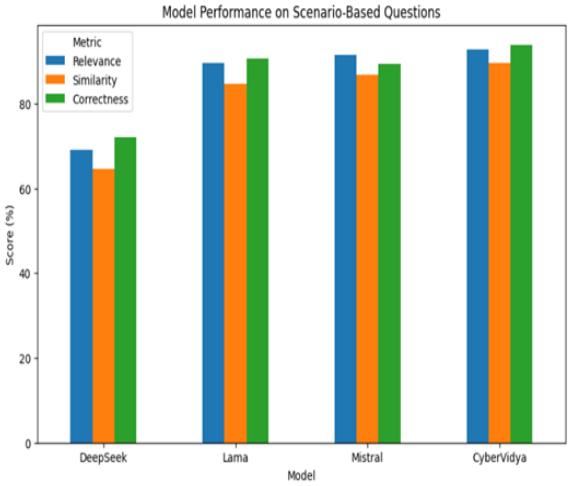

Figure 2 Model vs Score (Scenario Based Questions)

4.2.Discussion

4.2.1. Answers to Educational Questions

Well-defined cybersecurity principles, best practices, and theoretical understanding are the main topics of educational inquiries. CyberVidya regularly beat rival models across all three measures, according to the evaluation results. Relevance: With a score of 92.86, CyberVidya outperformed DeepSeek (80.66), Lama (89.06), and Mistral (85.93). Similarity: CyberVidya showed a strong ability to provide semantically accurate solutions, closely matching reference answers, with an 85.81 score. Correctness: With a score of 95.06, CyberVidya outperformed Mistral (87.41) and Lama (90.73), demonstrating the highest factual accuracy and demonstrating its usefulness in educational settings. These findings imply that CyberVidya is a trustworthy resource for cybersecurity education and awareness since it not only comprehends cybersecurity concepts but also offers answers that are extremely accurate and pertinent to the context.

4.2.2. Answering Questions Based on Scenarios

To assess each model's performance in real-world cybersecurity scenarios, where contextual reasoning and adaptive responses are essential, scenario-based questions were created. The results show that CyberVidya is also very good in this area: Relevance: CyberVidya had a great capacity to comprehend real-world cybersecurity scenarios with a score of 92.89,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

which was somewhat higher than Mistral's (91.63) and Lama's (89.55). Similarity: With an 89.56 score, CyberVidya produced answers that were closer to the anticipated answers than those produced by other modes.

4.2.3. Comparative Insights and Observations

There are notable variations in the models' performance when compared, especially in terms of accuracy and contextual awareness. Across both question categories, CyberVidya regularly outperformed Mistral and Lama, demonstrating overall superiority. Though they didn't match CyberVidya's level of accuracy and domain-specific precision, Mistral and Lama did rather well. DeepSeek's shortcomings in real-world cybersecurity reasoning were evident in both areas, particularly in scenario-based responses. These outcomes demonstrate how well CyberVidya's Retrieval-Augmented Generation (RAG) method works, including retrieved cybersecurity knowledge into the answer generation procedure.

4.2.4. Limitations

- **Data Coverage and Quality:** Our dataset, which consists of scenario-based case studies, real-time threat blogs, and books on ethical hacking, is crucial to the system's efficacy. The system's capacity to successfully handle novel or specialized threats may be hampered by any errors or out-of-date information in these sources.
- **Model Dependency on Retrieval Accuracy:** The Retrieval-Augmented Generation (RAG) approach relies on the retrieval mechanism's accuracy to be successful. The generated responses might not be precise if the algorithm is unable to retrieve most pertinent or contextually appropriate materials, which could result in inaccurate cybersecurity advise.
- **Managing Quickly Changing Threats:** Even with real-time knowledge base updates, the system might not be able to keep up with zero-day vulnerabilities or newly developed attack methods, which could cause a delay in prompt and precise reactions to the most recent threats.
- **Limited Real-World Testing:** CyberVidya has not yet undergone thorough testing in a variety of real-world cybersecurity scenarios, despite initial assessments demonstrating encouraging outcomes. Its dependability in dynamic, high-pressure situations has not yet been thoroughly verified.

Conclusion

CyberVidya represents a significant advancement in the field of cybersecurity AI, integrating Retrieval-Augmented Generation (RAG) with specialized domain-specific knowledge to provide real-time, actionable insights for cybersecurity professionals. Through the seamless combination of document retrieval and transformer-based generation, CyberVidya is able to offer responses that are not only relevant but also accurate and contextually aware, setting it apart from traditional static chatbots.The experiments and evaluations conducted across both educational and scenario-based questions demonstrate that CyberVidya outperforms existing open-source LLM models such as DeepSeek, Lama, and Mistral in key metrics such as relevance, similarity, and correctness. These results validate the robustness of the RAG pipeline, which allows CyberVidya to stay adaptable and continuously provide updated guidance based on the most relevant cybersecurity knowledge. By leveraging LlamaIndex for efficient document retrieval and LangChain for sophisticated orchestration, CyberVidya ensures real-time query handling with precise and actionable responses. Overall, CyberVidya's ability to dynamically incorporate retrieved knowledge into its response generation makes it an innovative solution for tackling the evolving challenges of the cybersecurity domain, highlighting its potential as a reliable, scalable, and contextually aware AI assistant for both practitioners and learners in the cybersecurity field.

Acknowledgements

Acknowledgment We would like to express our sincere gratitude to our Project Mentor Ms. R. R. Owhal, Head of The Department Mr. Riyaz Jamadar, professors, and peers who provided valuable

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

guidance and support throughout the development of CyberVidya: RAG-Infused Cyber Solutions. Their insights and constructive feedback were instrumental in refining our research and ensuring its real-world applicability. We extend our appreciation to the authors of MoRSE, RAG, and TTP Analysis research papers, whose pioneering work in retrieval-augmented generation and cyber security applications laid the foundation for our project. Their contributions inspired us to push the boundaries of cyber security-focused AI systems and address the limitations observed in prior studies. We also acknowledge the open-source community, including contributors to Git Hub, Hugging Face Transformers, LlamaIndex, LangChain and Kaggle whose tools played a crucial role in our model's implementation. Their continued efforts in democratizing AI and retrieval-based architectures enabled us to build a scalable, efficient, and context-aware cyber security assistant. A special thanks to our institution and faculty for providing the necessary resources and an encouraging research environment that allowed us to explore, innovate, and develop a solution that addresses real-world cyber security challenges. Lastly, we are grateful to our friends and family for their unwavering support and patience throughout this journey. Their encouragement kept us motivated to strive for excellence.

References

[1]. Marco Simoni , Andrea Saracino , Vinod Puthuvath,, and Maurco Conti, 1 Istituto di Informatica e Telematica, Consiglio Nazionale Delle Ricerche, Pisa, Italy 2TeCIP, Scuola Universitaria Superiore Sant'Anna, Pisa, Italy 3University of Padua, Italy and Delft University of Technology, Netherlands

[2]. Min Gao, Yanqi Zong, Jize Xiong, Shulin Li, 2024 Trine University Phoneix, USA 3rd International Conference on Computer Technologies ICC Tech, mingao4460@gmail.com Northern Arizona University Flagstaff, USA jasonyolo98@outlook.com, yzong22@my.trine.edu Trine University Phoneix, USA liam.cool666@gmail.com.

[3]. Reza Fayyazi, Rozhina Taghdimi & Shanchieh Jay Yang Department of Electrical & Computer Engineering Rochester Institute of Technology Rochester, NY, USA rf1679@rit.edu, rt3271@rit.edu, jay.yang@rit.edu

[4]. Jatin Pal Singh1 , Shobhit Agrawal2 International Journal of Science and Research (IJSR),ISSN: 2319-7064 SJIF (2022): 7.942,doi: DOI: 10.21275/SR24502103758,2024

[5]. V. Bhat, S. D. Cheerla, J. R. Mathew, N. Pathak, G. Liu and J. Gao, "Retrieval Augmented Generation (RAG) Based Restaurant Chatbot with AI Testability," 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService), Shanghai, China, 2024

[6]. B. Saha, U. Saha and M. Zubair Malik, "QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance," in IEEE Access, vol. 12, pp. 185401-185410, 2024, doi: 10.1109/ACCESS.2024.3513155.

[7]. M. Fateen, B. Wang and T. Mine, "Beyond Scores: A Modular RAG-Based System for Automatic Short Answer Scoring With Feedback," in IEEE Access, vol. 12, pp. 185371-185385, 2024, doi: 10.1109/ACCESS.2024.3508747.

[8]. V. Gummadi, P. Udayaraju, V. R. Sarabu, C. Ravulu, D. R. Seelam and S. Venkataramana 2024 4th International Conference on Sustainable Expert Systems ICSES, Kaski, Nepal, 2024, pp. 612-617, doi: 10.1109/ICSES63445.2024.10763024.

[9]. N. Tihanyi, M. A. Ferrag, R. Jain, T. Bisztray and M. Debbah, "CyberMetric: A Benchmark Dataset based on Retrieval-Augmented Generation for Evaluating LLMs in Cybersecurity Knowledge," 2024 IEEE International Conference on Cyber Security and Resilience (CSR), London, United Kingdom, 2024, pp. 296-302, doi: 10.1109/CSR61664.2024.10679494.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 03 March 2025
Page No: 456-464
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0063

[10]. A. Kaplunovich, "Cybersecurity Risks of Social Network Data Aggregation: Leveraging Machine Learning and LLMs in Cloud Environments," 2024 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS), Dubrovnik, Croatia, 2024, pp. 68-75, doi: 10.1109/ ICCNS62192. 2024.10776060

[11]. E. Sanu, M. Mummigatti and Mohana, "Design of Chatbot to Prevent Cyberbullying from Social Media," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 1437-1441, doi: 10.1109/ICSCNA58489.2023.10370729.

[12]. H. K. Chaubey, G. Tripathi, R. Ranjan and S. k. Gopalaiyengar, "Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development," 2024 International Conference on Future Technologies for Smart Society (ICFTSS), Kuala Lumpur, Malaysia, 2024, pp. 169-172, doi: 10.1109/ICFTSS61109.2024.10691338

[13]. S. Vakayil, D. S. Juliet, A. J and S. Vakayil, "RAG-Based LLM Chatbot Using Llama-2," 2024 7th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICDCS59278.2024.10561020.

[14]. Y. B. Sree, A. Sathvik, D. S. Hema Akshit, O. Kumar and B. S. Pranav Rao, "Retrieval-Augmented Generation Based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health," 2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), Pattaya, Thailand, 2024, pp. 1-8, doi: 10.1109/ ICECIE63774. 2024.10815693.

[15]. M. Kim, D. Kim, Y. Park and D. Jeong, "Development of an Expert Chatbot for Digital Forensics Using RAG Model Implementation," 2024 International Conference on Platform Technology and Service (PlatCon), Jeju, Korea, Republic of, 2024, pp. 182-187, doi: 10.1109/ PlatCon63925. 2024.10830748.