World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:18, No:7, 2024

Retrieval Augmented Generation against the Machine: Merging Human Cyber Security Expertise with Generative AI

Brennan Lodge

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

Abstract**—**Amidst a complex regulatory landscape, Retrieval Augmented Generation (RAG) emerges as a transformative tool for Governance Risk and Compliance (GRC) officers. This paper details the application of RAG in synthesizing Large Language Models (LLMs) with external knowledge bases, offering GRC professionals an advanced means to adapt to rapid changes in compliance requirements. While the development for standalone LLMs is exciting, such models do have their downsides. LLMs cannot easily expand or revise their memory, and they cannot straightforwardly provide insight into their predictions, and may produce “hallucinations.” Leveraging a pre-trained seq2seq transformer and a dense vector index of domain-specific data, this approach integrates real-time data retrieval into the generative process, enabling gap analysis and the dynamic generation of compliance and risk management content. We delve into the mechanics of RAG, focusing on its dual structure that pairs parametric knowledge contained within the transformer model with non-parametric data extracted from an updatable corpus. This hybrid model enhances decision-making through context-rich insights, drawing from the most current and relevant information, thereby enabling GRC officers to maintain a proactive compliance stance. Our methodology aligns with the latest advances in neural network fine-tuning, providing a granular, token-level application of retrieved information to inform and generate compliance narratives. By employing RAG, we exhibit a scalable solution that can adapt to novel regulatory challenges and cybersecurity threats, offering GRC officers a robust, predictive tool that augments their expertise. The granular application of RAG’s dual structure not only improves compliance and risk management protocols but also informs the development of compliance narratives with pinpoint accuracy. It underscores AI’s emerging role in strategic risk mitigation and proactive policy formation, positioning GRC officers to anticipate and navigate the complexities of regulatory evolution confidently.

Keywords**—**Retrieval Augmented Generation, Governance Risk and Compliance, Cybersecurity, AI-driven Compliance, Risk Management, Generative AI.

I. INTRODUCTION

THE industry of GRC (Governance, Risk, and Compliance) faces continual pressure on all fronts of cybersecurity defense. GRC analysts are crucial defenders in the digital battlefront, facing a barrage of challenges that require swift, precise decisions and in depth understanding of the cybersecurity risk and privacy domain. This paper and technology present a solution tailored to uplift and empower both these crucial aspects of GRC through RAG. The Comprehensive AI-based Data Driven Integrated Environment, or CADDIE, represents such an innovation, empowering GRC officers with a robust toolset designed to streamline and elevate compliance processes.

CADDIE employs the RAG framework to harmonize the extensive knowledge of LLMs with the depth of external databases. This fusion creates a potent advisory system, enabling GRC professionals to conduct real-time gap analyses, interpret complex data privacy laws, and dynamically update organizational policies. Through the application of Chroma Database’s (DB) vector database capabilities, CADDIE enhances its retrieval functions with high-speed, in-memory computations, facilitating the efficient management of vast, unstructured data sources essential for comprehensive GRC advisories.

This approach positions CADDIE as an environment where semantic search intersects with strategic regulatory analysis. By leveraging the vector storage system of Chroma DB, CADDIE not only queries and retrieves relevant regulatory content but also engages in semantic analysis to identify nuances across different jurisdictions’ privacy laws. Such capabilities are imperative for organizations aiming to do business internationally, ensuring their adherence to a myriad of evolving global compliance standards.

Navigating the labyrinth of global data privacy regulations is an increasingly daunting task for organizations. The Information Commissioner's Office (ICO) in the UK highlights the essence of starting with data protection, underscoring the need for organizations, especially small ones, to comprehend and comply with the General Data Protection Regulation (GDPR) [10]. They emphasize the importance of recognizing the severity of data breaches and the consequent legal and reputational repercussions. The complexity deepens as new legislation, such as Maryland's pioneering children's privacy laws, presents additional layers of compliance, particularly in the tech industry, indicating a trend towards more stringent data protection standards [11]. The stakes of non-compliance are tangible and escalating, as illustrated by the GDPR's hefty fines imposed on companies for violations. Fines reaching into the millions serve as a stern reminder of the financial risks associated with failing to adhere to these evolving regulations [12].

As 2024 approaches, cybersecurity teams must recognize the

Brennan Lodge is with New York University, USA (e-mail: bl1584@nyu.edu).

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

transformation from theoretical generative AI concepts to their practical application, as 75% of enterprises are anticipated to adopt this shift [1]. This growing trend underscores the industry's move towards crafting tailored, in-house AI models that cater to specific needs. To stay ahead in their defenses, GRC teams should leverage generative AI meaningfully, focusing on solutions tailored to address unique challenges like regulations and threats, rather than relying on broad, generalized models which may not offer the context they need for defense and regulatory changes.

Owning and controlling these models offer adaptability and customization according to specific organizational demands. One of the standout advantages is speed and efficiency. Localized data access provides fast query responses in near real-time. The dynamic cross-referencing of live data ensures deeper, more accurate insights. Contrary to many LLM solutions that jeopardize intellectual property by querying the internet through APIs, CADDIE's design emphasizes security, pace, and efficiency, utilizing local resources and a cutting-edge RAG infrastructure.

In the context of data security and sovereignty, organizations are increasingly recognizing the value of self-hosted, on-premises solutions like the RAG model for managing sensitive GRC-related data. A self-contained RAG system such as CADDIE offers the dual advantages of protecting intellectual property and tailoring AI to an organization's unique regulatory landscape without the risk exposure inherent in external API dependencies. By operationalizing CADDIE within a company's virtual private cloud (VPC), data are processed in a controlled environment, which not only fortifies against breaches but also ensures compliance strategies are proactive and attuned to the organization's specific operational fabric.

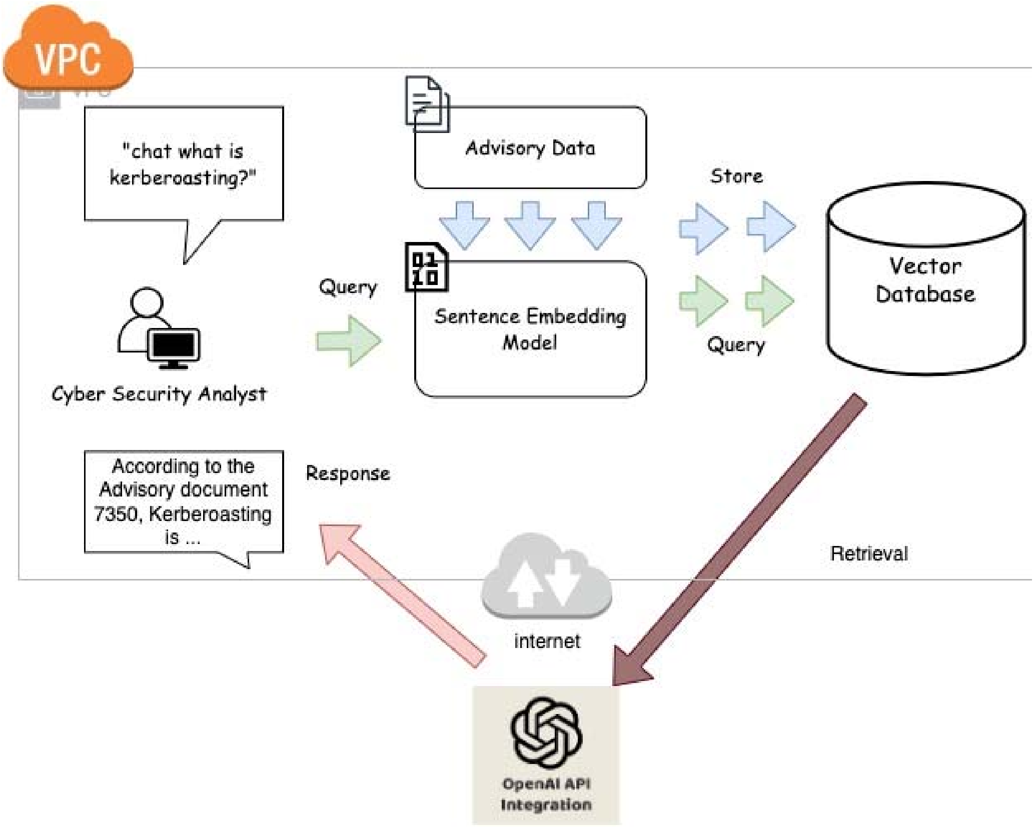


Fig. 1 A RAG Solution Using OpenAI

The RAG architecture detailed in the diagram in Fig. 1 underscores the importance of in-house data management, where CADDIE, empowered with vector databases and sentence embedding models, can swiftly respond to queries with heightened relevance and context-awareness. This encapsulated ecosystem negates the vulnerability of transmitting intellectual property over the internet, as is the case with solutions like OpenAI's external API. It brings to the fore an intelligent system capable of intricate semantic analysis, understanding the subtleties of company policies and regulations, and providing GRC professionals with real-time, insightful advisory—all while upholding the principles of data sovereignty and maximizing cybersecurity hygiene.

The key takeaways from this proposed solution approach include its prime focus on near real-time AI-driven risk advisory and rapid analysis. The system draws upon various sources like the UNCTAD Global Cyberlaw Tracker [2], GDPR, SOX, NIST, and company or government regulatory specific Information Security Policies, ensuring a comprehensive data-driven and cross-referenced response. The advanced AI system offers not just context but also serves up customized, relevant data, links to localized or referenced

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

documents and support. Designed to integrate seamlessly into chat systems like Slack or into document repositories like Confluence, it provides proactive security measures, logging, audit trails and traceability.

This system not only responds and contextualizes threats but also educates analysts about them, giving clear insights into regulatory impacts. Staying true to a cooperative ethos, it remains transparent in its operations and promises to democratize advanced risk tools, ensuring cost-effectiveness, upskilling, speed and awareness.

This paper will further articulate the functions and benefits of CADDIE, detailing how its innovative use of RAG technology stands to redefine the GRC landscape. By integrating CADDIE into their operational framework, organizations can transform their approach to compliance, achieving a higher standard of governance, risk management, and adherence to an ever-changing array of regulations.

II. GOALS AND IMPACT

With a new technological revolution among us, driven by Generative (Gen) Artificial Intelligence (AI), the challenges confronting analysts are escalating at an unprecedented rate. As adversaries harness the power of Gen AI, arming themselves with sophisticated, swift, cost-effective, and automated tools, it becomes vital for our defenses to evolve in tandem. We stand on the brink, faced with a choice: adapt with Gen AI-enhanced strategies or be eclipsed. The call is clear — to combat this new wave for our risk and privacy defenders, we must harness this technology: fighting fire with fire.

In an era where cyber threats persist around the clock, our defense mechanisms are often limited by the conventional 9 to 5 working hours of our cybersecurity staff. These experts, with a jack-of-all trades expertise in multiple computing fields, grapple with a relentless wave of alerts to triage. This can warrant fatigue and anxiety to obtain unrealistic goals like the “SIEM queue to zero.” Compounding factors of these soaring stress levels can lead to burnout, Chief Information Security Officers (CISOs) overwhelmed duties and high turnover within the cybersecurity defense community.

While the demand for risk specialists sees an unprecedented spike—with the United States Bureau of Labor and Statistics forecasting a 35% growth in information security analyst roles by 2031—the industry simultaneously faces a severe talent shortage [2]. An estimation by Cybersecurity Ventures indicates an increase of 350% in unfilled positions from 2013 to 2023, culminating in 3.5 million unoccupied roles for this year. [2].

The strain within this sector plays a pivotal role in discouraging talent retention and acquisition. Corroborating this concern, a Forrester study disclosed that a significant 66% of cybersecurity professionals have confronted severe stress or burnout, with over half of them necessitating medication for their deteriorating mental health [3]. Such prevalent and profound challenges underscore an urgent need for innovative solutions and Gen AI solutions to support and sustain the workforce defending our digital frontiers.

The intensifying barrage of cyber-attacks correlates to the deepening threat of a burnout crisis faced by cybersecurity analysts. In Q1 2023 alone, global cyber-attacks rose by 7% from the previous year, inundating each firm with an average of 1,248 attacks weekly. This surge, as detailed in CheckPoint's recent research, places immense pressure on these professionals, particularly in sectors like education and research, which now grapples with a staggering 2,507 attacks weekly—a 15% jump from Q1 2022 [9].

Amidst this overwhelming alert volume, there is an unmistakable urgency to implement advanced solutions. Leveraging Generative AI technology and RAG can empower analysts with real-time, informed responses. By streamlining access to contextually relevant information, these technologies can significantly alleviate alert pressures, directly addressing and mitigating the burnout that has become endemic in the risk and privacy industry.

Government and private institutions grapple with selecting an LLM model or Gen AI application to fine-tune, especially given the privacy and IP concerns with vendor solutions. They must also evaluate the cost-benefit analysis of establishing a custom, expensive infrastructure for their partially developed use cases. Furthermore, they face the challenge of integrating into a stable production environment that might produce unreliable results.

Knowledge-intensive and critical thinking tasks performed by risk and security analysts are unique and require an innate mindset to investigate and differentiate malicious from benign events. Articulating this institutional gained knowledge and then teaching junior analysts is even a more difficult task. Unfortunately, this is the conundrum that needs to be solved or at least alleviated in the industry.

Brennan, a seasoned GRC officer at an international corporation, is well-versed in navigating the complexities of regulatory compliance. His recent challenge involves ensuring the company's policies align with the stringent demands of the New York State Department of Financial Services' Cybersecurity Regulation (NYCRR 500). This legislation requires detailed protocols for the appointment and function of a Chief Information Security Officer (CISO)—a role central to an organization’s cybersecurity posture.

Within the context of his role, Brennan, objective is to conduct a meticulous gap analysis to compare his company’s current cybersecurity policies against the NYCRR 500 framework. Utilizing CADDIE's RAG-powered capabilities, she inputs queries related to the independence and duties of the CISO as outlined in the regulation. CADDIE, anchored in a scientifically rigorous application of AI, leverages a structured vector database system to dissect the inquiry semantically. It systematically mines the information from a curated dataset that encapsulates not only the NYCRR 500 text but also the interrelated legal precedents and organizational policies.

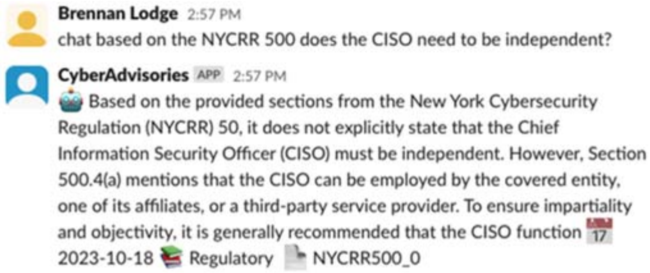
CADDIE employs an advanced algorithmic approach that utilizes a seq2seq transformer model coupled with a top-K retrieval strategy from dense vector spaces to analyze Brennan's query. This technical methodology ensures that the response is

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

precise and evidence-based, pulling from the most relevant sections of the NYCRR 500 and the company's existing policy documentation. The juxtaposition of these datasets allows CADDIE to highlight potential discrepancies with empirical specificity, offering Brennan a scientifically grounded, quantifiable assessment of compliance gaps.

Through this process, Brennan gains insights into aspects of the CISO's role that may require policy revisions. The RAG mechanism within CADDIE not only identifies the sections for review but also provides suggestions for amendments, backed by statistical models that forecast the potential implications of policy changes. This level of analysis underscores CADDIE's commitment to going beyond mere data retrieval, venturing into the predictive analytics that inform Sarah's strategic decision-making.

This use case demonstrates how CADDIE is pivotal in translating complex regulatory content into actionable insights, facilitating compliance not as a static requirement but as a dynamic, ongoing process. The use of CADDIE embodies a move towards an integrated, intelligent approach to GRC, where AI is not just a tool but a partner in the decision-making process, empowering officers to maintain resilience in a demanding regulatory environment.



Brennan Lodge 2:57 PM
chat based on the NYCRR 500 does the CISO need to be independent?

CyberAdvisories APP 2:57 PM
Based on the provided sections from the New York Cybersecurity Regulation (NYCRR) 50, it does not explicitly state that the Chief Information Security Officer (CISO) must be independent. However, Section 500.4(a) mentions that the CISO can be employed by the covered entity, one of its affiliates, or a third-party service provider. To ensure impartiality and objectivity, it is generally recommended that the CISO function 📅 2023-10-18 📚 Regulatory 📄 NYCRR500_0

Fig. 2 Prompt and response to request details about NYCRR 500

III. RELATED WORK

Meta, through its white paper and subsequent blog post titled "Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models," highlights the transformative potential of RAG [5]. This technology, characterized by its unique fusion of intrinsic parametric memory and external non-parametric memory, facilitates the delivery of context-rich insights. RAG's design promotes ease of integration within risk and governance data pipelines, ensuring a streamlined implementation process. As we elaborate in this paper, RAG's capability to combine the adaptability of "closed-book" approaches with the precision of "open-book" methods, transparency and "explainabilty" stand to be a pivotal asset for the risk and compliance community, offering both efficiency, agility, and enhanced threat awareness [5].

RAG Process Explained

The process of RAG stands at the core of CADDIE's functionality, offering a sophisticated solution for GRC officers navigating the complexities of regulatory compliance and policy adherence. This system synthesizes the extensive, pre-trained parametric knowledge embedded within a LLM with the exhaustive non-parametric data from a dense vector index, such as corporate documents, regulations, and policies.

Query Processing

In the context of GRC, a compliance officer's query initiates the RAG workflow. This input is first encoded into a query vector by the Query Encoder, which essentially captures the semantic essence of the query using vector representations. The encoding process is crucial because it transforms the natural language query into a format that the system can work with to find relevant information.

Document Retrieval

The RAG system employs algorithms like Maximum Inner Product Search (MIPS) to retrieve the most relevant documents from a dense vector index. This index is a pre-processed collection of documents transformed into vector space to facilitate quick and efficient retrieval. For a compliance officer, this step would involve finding the most pertinent regulations, standards, and policy documents that align with the query's intent.

Sentence Embedding

Once relevant documents are identified, they are passed through a Sentence Embedding Model. This model further processes the documents, converting them into dense vector representations. These vectors are essential as they capture the deeper semantic significance of the documents, which is important for generating accurate and relevant responses.

Response Generation

Armed with semantically rich embeddings from the previous step, the RAG system utilizes a pre-trained LLM to synthesize this information. The LLM generates a response that not only answers the GRC officer's query but is also factually grounded and comprehensive. The generation is informed by both the parametric knowledge embedded within the LLM and the non-parametric information retrieved from the vector index.

Delivery to Interface

The final step in the RAG loop is the delivery of the generated response back to the user interface. This could be a chat application, an internal GRC platform, or any other system through which the officer interacts with CADDIE. The response provided includes the information required for the officer to make informed decisions regarding compliance, risk management, or policy updates.

The RAG architecture has been effectively utilized for knowledge-intensive NLP tasks by combining the parametric knowledge of a seq2seq model with the non-parametric memory of a dense vector index of information sources. This hybrid approach has been shown to enhance the performance of AI systems in tasks that require a deep understanding of content and context, such as open-domain question answering [4]. By adapting the RAG framework for the GRC domain, CADDIE benefits from the same cutting-edge technology to assist

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

compliance officers in navigating complex regulatory landscapes. RAG is a model that effectively integrates pre-trained parametric and non-parametric memories for language generation, setting the foundation for the technology that CADDIE leverages for a GRC advisor.

Throughout this process, the RAG system maintains an intricate balance between the wealth of static, pre-encoded knowledge and the need for dynamic, real-time data retrieval, ensuring that the generated responses are both accurate and current. This interplay allows the RAG model to construct responses at a granular level, employing different passages from the non-parametric memory as necessary, to formulate the most informed and precise answer.

Technical Insights

The underlying efficiency of the RAG process in CADDIE is rooted in its dual-memory approach, integrating the parametric memory of a seq2seq model with the non-parametric memory of a dense vector index, typically comprising Wikipedia or a tailored corporate database. The RAG models are trained end-to-end, enabling the generator to condition on the input and the retrieved latent documents to produce the output. This optimization is essential for marginalizing the latent documents, where a top-K approximation is employed to account for multiple documents influencing the generated response.

By adopting this dual-memory architecture, CADDIE can provide GRC officers with context-rich insights that are essential for making informed decisions in the face of regulatory changes and compliance requirements. The adaptability and precision of the RAG approach, as elucidated in the paper, position it as a transformative asset for GRC, enhancing the ability to forecast risks, analyze policy gaps, and generate actionable guidelines within the complex landscape of GRC.

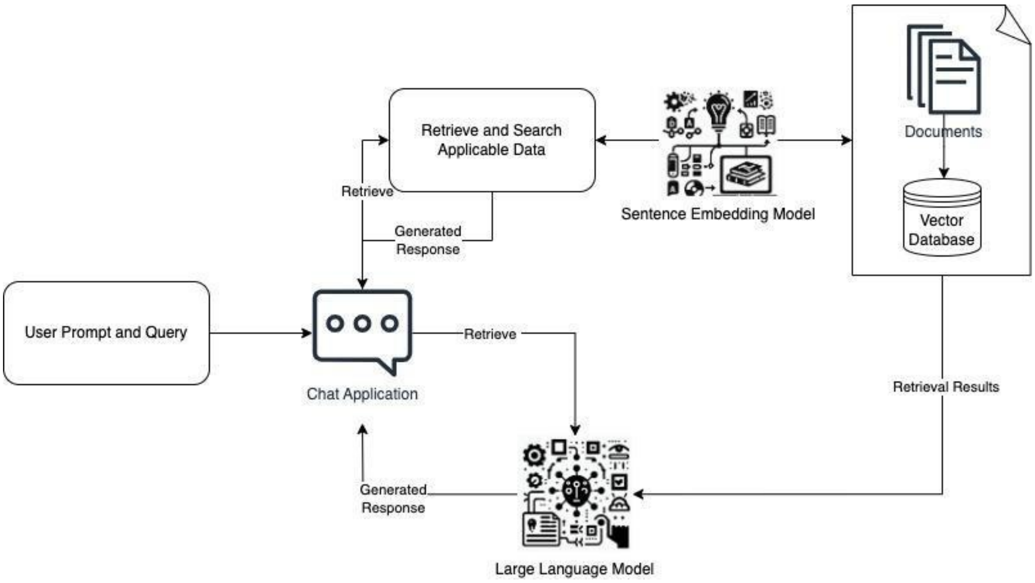


Fig. 3 RAG Explained

a. Sec2sec

Under the hood, sec2sec, a fundamental technology that supports RAGs, has been the linchpin to the proliferation of Gen AI. In the domain of natural language processing (NLP), a sequence-to-sequence (seq2seq) model stands as a cornerstone for various applications, including machine translation, text summarization, and speech recognition.

A seq2seq model comprises two primary components: an encoder and a decoder. The encoder processes an input sequence and compresses this information into a fixed-size context or "state" vector. This vector captures the essence of the input sequence. Subsequently, the decoder takes this vector and produces an output sequence. The strength of seq2seq models lies in its ability to handle sequences of varying lengths, making them adaptable and powerful tools for tasks that involve transforming one sequence into another.

b. Vector Databases

A vector database now comes into play here with the foundation of sec2sec and the RAG delivery. In the CADDIE prototype, the CISA Advisories, SANS Information Security Policies, Cybersecurity Regulatory Requirements and data privacy laws documents are ingested into the vector database in a standard schema to lead to clearer similarity in vectors and a clearer knowledge generation in its referencing and output [8].

Vector databases store data in vectorized forms, where each data point, be it textual, image-based, or any other type of information, is transformed into a high-dimensional vector representation. These vectors, which are mathematical representations of the textual data, are compact and can be compared rapidly for similarity. When integrated with a vector database, these embeddings serve as efficient lookup keys. This structure allows for efficient similarity-based searches, meaning the system can swiftly identify and retrieve

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

information that is contextually or semantically related to a given query.

The rapidly shifting terrain of today's threat and regulatory landscape underscores an urgent necessity: analysts must have access to constantly updated and relevant threat or regulatory intelligence information to effectively combat evolving cyber risks. Where the line between signal and noise, or malicious and benign often blurs, the real challenge for analysts, compliance officers, auditors and engineers goes beyond mere data collection. The real skill set for top notch analysts is their ability to decode the subtle semantic nuances or institutional knowledge embedded within that data and relate it to an actual risk. As threats evolve with increasing sophistication, the need for tools that can navigate these semantic layers becomes ideal. ChromaDB, an open source vector database, fits the bill to meet our demands.

Infused with the prowess of transforming vast textual repositories into dense vector embeddings, ChromaDB ensures that the richness of the data is not lost in translation. ChromaDB's capabilities echo this urgency of offering a dynamic way of constantly ingesting new shards of information. Thus, as cyber threats evolve, our RAG with ChromaDB, ensures that analysts are always armed with the latest and most relevant threat intelligence, making them more adept at navigating the challenging waters of cybersecurity investigations.

One significant integration point or glue between the RAG and of ChromaDB is the use of a GPT4All LLM [6]. This technology is a pioneering advancement in the realm of embedding models, transforming text into dense 384-dimensional vectors [6]. It is designed to encapsulate the semantic essence of the input text. When provided with a string of text, the Sentence Transformer returns a vector that holds profound semantic knowledge, making it an invaluable asset for information retrieval, content clustering, and establishing sentence similarity [7].

By utilizing this framework and the fundamental RAG architecture, analysts have the capability to probe the vector database for intricate details on specific malware signatures or risks. With the integration of LLMs, not only can analysts receive a precise match to their query and a cross-referenced document but also an intelligible explanation. In essence, vector databases underpin the knowledge translation in Gen AI. They bridge the gap between raw data and meaningful insights, ensuring that analysts are equipped with comprehensive and relevant information.

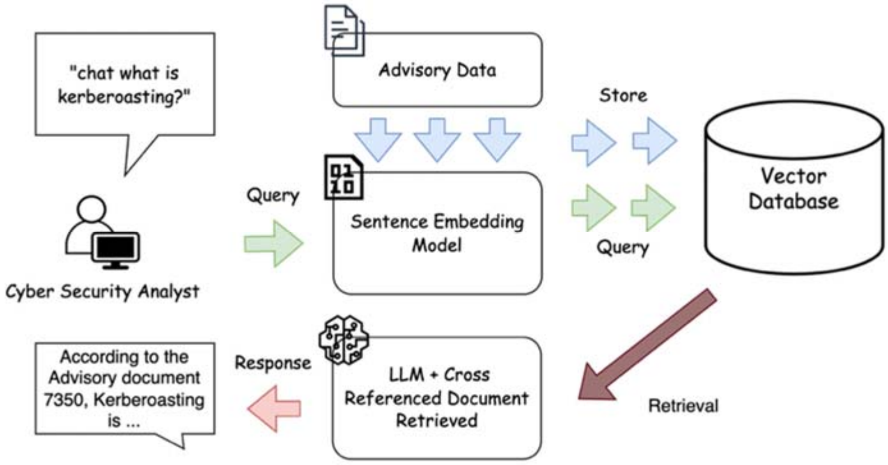

"chat what is kerberoasting?"
Advisory Data
Store
Query
Sentence Embedding Model
Query
Vector Database
Cyber Security Analyst
Response
According to the Advisory document 7350, Kerberoasting is ...
LLM + Cross Referenced Document Retrieved
Retrieval

Fig. 4 RAG + Cyber Advisory Use Case

c. Integrating a LLM to the RAG

Along with the confluence of RAG and vector databases, the third and vital component is the LLM. The success of RAG and the vector databases in enhancing cybersecurity measures hinges on the coherent integration of all their components, and the keystone in this architecture is the LLM. With the dynamism of vector databases illuminating intricate patterns and vulnerabilities, it is important to have a tool that can not only interpret these data but also understand it in ways that are comprehensible and actionable.

The GPT4All LLM is emerging as an indispensable element of our trifecta. Derived from the March 2023 GPT4All release, it stands out as an exemplar of progression in LLM technology, trained on an expansive, curated corpus of various domains ranging from multi-turn dialogues to creative writing [5]. This vast training set empowers GPT4All with a unique blend of precision and creativity, making it a perfect ally for tasks that require both exact matches and intense problem-solving.

d. Open Source

One of the most enticing features of CADDIE is its open-source nature. By openly releasing the training data, curation procedure, and the final model weights, the LLM used (GPT4All) paves the way for research transparency, reproducibility, and community-driven enhancements. GPT4All, with its combination of LLaMA 2, also offers Python bindings, allowing an even broader user base to integrate its capabilities, customize, integrate and scale to cybersecurity engineers’ frameworks.

In terms of evaluation and performance, GPT4All models

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

have demonstrated promising results. Preliminary evaluations, comparing it with other models, have shown its edge, particularly when fine-tuned on custom datasets. In the third paper released by the GPT4All team, titled "GPT4All: An Ecosystem of Open-Source Assistants That Run on Local Hardware," the authors present impressive evaluation results within the non-OpenAI model space [5]. With almost new model releases at a monthly rate, the performance continues to improve and GPT4All is a clear front runner in providing models in the non-Open AI realm.

GPT4All represents a paradigm shift in cost-effective, local, and customizable LLM training. With total experiments costing around $5,000 in GPU expenses and a training cost of a mere $200 for the final model, it democratizes access to state-of-the-art AI capabilities. Moreover, organizations can cost-effectively fine-tune GPT4All with in-house datasets, tailoring its capabilities to meet specific needs. Yet, while its capabilities and commercial viability are enticing, ethical considerations remain at the forefront. While GPT4All is freely available under the Apache 2 license, it offers a user-friendly, and database to query the training data used for its models.

Our decision to incorporate LLMs based on the LLaMA 2 technology and GPT4All models was both strategic and deliberate. These models, particularly GPT4All, represent a pivotable point in open source advancements, offering unparalleled capabilities in data interpretation, creative problem solving, and adaptability. Given its open source nature, it promises a transparent, community-driven evolution, ensuring that its performance remains at the cutting edge. This aligns seamlessly with the demands of our cybersecurity purposeful advisory wrangling with LLMs, or CADDIE’s prototype, where rapid, dynamic threat assessments are critical. The commercial viability offered by GPT4All also ensures that entities, regardless of their scale, can access and benefit from this fusion of RAG, vector databases, and an LLM. Together, they create a trifecta that propels cybersecurity defenses into an era of unprecedented sophistication and efficiency.

Selecting the right infrastructure was paramount, emphasizing high performance, reliability, and scalability. We chose Amazon Web Services (AWS) Elastic Compute (EC2) for its familiarity, scalability, flexibility for experimentation, and cost-effectiveness. Through extensive testing—addressing driver compatibility, integrating models, and refining development efforts—we aimed for a benchmark of a 10-second response from the RAG within a Slack channel. Ultimately, a g4dn.4xlarge instance equipped with a T4 NVIDIA GPU proved to be the perfect fit.

In addition to hitting the acceptance criteria for response to the Slack chat prompt, we hit the cost-effective prowess of another benchmark. We achieved a cost-effective run-rate of less than ~$50 per day for the prototype.

Beyond the foundational infrastructure, it was essential to ensure the optimal utilization of computational resources for the RAG to be integrated in a cybersecurity analyst workflow. This objective led us to the strategy of leveraging GPU offloading for fast processing and a quick delivery of results. This process, by which specific computational layers are directly allocated to the GPU's VRAM, not only optimizes RAM usage but also significantly accelerates processing speed.

In our pursuit of an efficient AI text-generation system, we integrated Kobold Lite, another open source tool, into our architecture. KoboldCpp, building upon the foundational capabilities of llama.cpp, offers a streamlined interface for text-generation. This integration confirms that the complexities of AI-driven text generation are abstracted, enabling a more efficient interaction with the underlying LLM, and thus optimizing system performance.

To demonstrate the efficacy of our prototype, we selected Slack as the primary user interface due to its ubiquity and straightforward API framework. This choice facilitated seamless interactions between users, analysts, and the RAG.

To establish a robust connection between Slack, our GPT4All LLM, and RAG, we employed Flask on our internal EC2 infrastructure. For network endpoint integration, NGROK was installed to ensure endpoint connectivity from Slack. This architecture and underlining infrastructure and services empowers analysts to submit prompts via Slack, enable the RAG to conduct similarity searches within the vector database and subsequently relay the combined insights from the LLM back to the Slack channel.

IV. FEASIBILITY

The CADDIE prototype represents a meticulous orchestration of scripts, services, APIs, and intricate coding to serve its singular mission: to address the queries of cybersecurity analysts via the #cyberadvisor_bot Slack Channel and deliver concise, insightful, cross-referenced, and accurate results.

This entire pipeline and infrastructure were meticulously designed to ensure timely responses, with a delivery target of under 10 seconds. Hosting the workflows, APIs (Application Programming Interface), and codebase on an AWS (Amazon Web Services) EC2 (Elastic Compute) g4dn.4xlarge instance was a critical choice, contributing significantly to system robustness. The NVIDIA T4 GPU was pivotal, managing most of the computational load efficiently. Achieving the benchmark for swift delivery while maintaining stability represented a significant accomplishment. This delicate equilibrium of speed and reliability was made possible through precise system configurations, regular driver updates, strategic alignment of software packages, and a bit of fortuitous success in developing the working prototype.

With the end user for CADDIE in focus, we center the entire experience on the needs and convenience of the cybersecurity analyst. Starting with an intuitive interaction, where the analyst queries the chatbot, the system seamlessly bridges to the Slack API, ensuring real-time responsiveness. The subsequent integration with the RAG and the Kobold API reinforces the system's commitment to delivering precision and accuracy in its insights.

The heart of this system, the GPT4All LLM Model, exemplifies cutting-edge AI capabilities, designed to

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

comprehend and process intricate cybersecurity nuances. Finally, results are coherently communicated back to the analyst via the Slack bot, ensuring clarity and actionable insights.

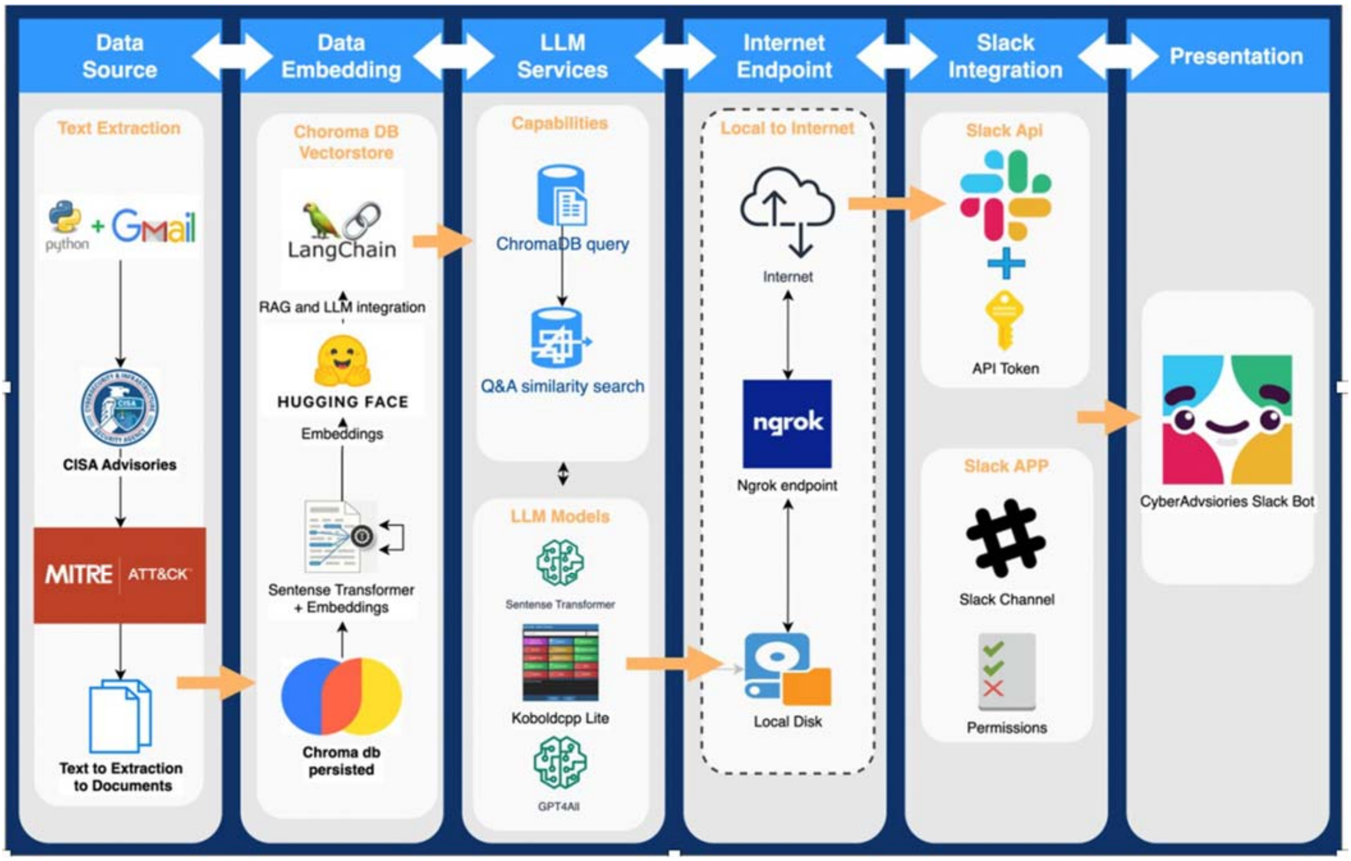


Fig. 5 CADDIE Architecture

This design philosophy illustrates our unwavering focus on enhancing the analyst's efficiency and decision-making capabilities. By embedding advanced AI technology within familiar platforms, we have ensured that CADDIE is not just a tool but an indispensable ally for every cybersecurity analyst, perfectly encapsulating our pursuit of technological excellence.

Our architecture is designed for both on-premises and cloud deployments, ensuring flexibility in infrastructure choices. Central to our technical strategy is the ability to customize configurations based on specific client parameters, particularly for enhancing cybersecurity responsiveness. Detailed descriptions of the RAG's Application Framework and Flask Application in the subsequent sections provide a more granular understanding of the system's integrative and functional aspects.

A. RAG Application Framework

Sentence Embedding with SentenceTransformer

At the foundation of CADDIE's data processing pipeline lies the SentenceTransformer. This component is responsible for converting textual data into dense vector representations. Using the model 'all-MiniLM-L12-v2', this SentenceTransformer ingests sentences and outputs their corresponding embeddings. These embeddings capture the semantic essence of the input, condensing complex data structures into manageable vector forms suitable for rapid retrieval and analysis.

B. ChromaDB – Persistent Vector Storage

Once embeddings are generated, they need efficient storage and retrieval mechanisms. This is where ChromaDB, a vector database, comes into play. It offers a persistent environment for the storage of sentence embeddings, facilitating their quick retrieval. The Chroma database is initialized and linked with the embeddings generated by the SentenceTransformer. This ensures that embeddings are immediately available for subsequent operations. Persistent storage is enabled here for real-time ingestion and updates to the vector database.

C. Data Ingestion & Transformation

Data from sources like UNCTAD Global Cyberlaw Tracker, CISA advisory, Regulatory Requirements, Information Security Policies, and MITRE ATT&CK are ingested using the Python Pandas library. Once ingested, the data schema is standardized, ensuring compatibility with Chroma's vector storage structure. This normalization process involves assigning fields like Subject, Date, Body, File path, and Source as the schema to the ChromaDB.

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

D. Embedding with HuggingFace

In addition to SentenceTransformer, CADDIE leverages the power of HuggingFace embeddings. Using the transformers library, tokenizers and models are instantiated for creating embeddings. The embeddings generated via this method are also stored in ChromaDB, expanding the repository of knowledge that can be accessed during the retrieval phase.

E. Retrieval Augmented Generation

Central to CADDIE's response mechanism is RetrievalQA, a LangChain function. Integrated with Chroma's retrieval capabilities, RetrievalQA ensures that any query posed to the system is matched semantically with the stored embeddings, leading to contextually accurate responses. This component uses the stored vectors to pinpoint relevant information, which is then used to craft detailed responses.

F. Integration with KoboldApiLLM and GPT4all Model

CADDIE further augments its response generation mechanism through the integration of the KoboldApiLLM, which interfaces with the GPT4All large language model. When a query is posed, relevant embeddings are first retrieved from Chroma, and then, with the combined power of RetrievalQA, KoboldApiLLM, and GPT4all, a comprehensive response is generated.

G. Prompt Engineering with Speed in Mind

Embedded inherently within the Kobold API, Smart Context—activated via the '—smartcontext' flag—innovatively enhances prompt engineering by allocating about 50% of the total context space as a buffer. This design not only refines the precision of results but also significantly expedites response and processing times by minimizing the need for frequent prompt recalculations. When nearing its max context, if Smart Context identifies a substantial overlap between two successive prompts, it intelligently restructures the context: the first half is truncated and the second half shifted upward, leaving ample room for appending new text. This streamlining ensures efficient use of context space, ensuring quicker and more accurate responses.

H. Platform Adaptability

The system's modular design ensures that it can be integrated into platforms such as Slack without compromising functionality. Depending on user requirements, CADDIE can operate in diverse environments, from on-premises installations to cloud-based solutions. This application, combined with its API and services, is a well-integrated mix of advanced technologies, with each component fulfilling a unique role in the RAG architecture. The strategic utilization of SentenceTransformer, Chroma, HuggingFace embeddings, RetrievalQA, and a GPT4all model ensures a robust and responsive system capable of addressing intricate GRC challenges.

I. Slack Integration and Event Handling

By integrating the Slack API through the Slack and slackeventsapi libraries, CADDIE seamlessly connects with the popular messaging platform. Utilizing Slack's token authentication (indicated by SLACK_TOKEN and SIGNING_ SECRET), the application establishes a secure channel to listen and respond to incoming messages. The SlackEventAdapter ensures that when messages are received, they trigger specific functions, enabling dynamic interactions between the system and users.

Leveraging state-of-the-art embeddings storage, advanced question-answering mechanisms, and dynamic event handling, this RAG system exemplifies cutting-edge integration in the cybersecurity domain. CADDIE embodies the synergy of contemporary technologies, ensuring unmatched performance and adaptability in addressing cybersecurity challenges. This integration, underpinned by sophisticated embeddings retrieval, robust question-answering frameworks, and dynamic event processing, underscores CADDIE's dedication to delivering real-time, pertinent cybersecurity solutions. This commitment positions CADDIE at the forefront of innovative responses in the ever-evolving landscape of cybersecurity.

V. USE CASES

The CADDIE system represents an interaction of strategic technology integrations, each chosen for its capability in its domain. Together, these elements form a cohesive, and robust solution, and position CADDIE as a formidable contender for servicing and informing cybersecurity analysts. The use cases below are ordered by prototype phase implementations. In its currently development, CADDIE is phase one of a delivery with its prototype buildout.

A. Advisories on Emerging Threats

The dynamic nature of the cyber domain necessitates real-time, updated threat intelligence for analysts. With CADDIE's capability to tap into CISA advisories, Information Security Policies, Regulatory Requirements, and the MITRE ATT&CK framework, analysts receive not only information but actionable intelligence. The quantitative advantage lies in the rapid reduction of unaddressed vulnerabilities and speed to resolution with accuracy. It ensures that responses are well-informed, precise, and aligned with the most current threat landscapes, bolstering overall security postures.

B. Navigating the Complex Landscape of Information Security Policies

For global Fortune 500 entities, information security policy adherence is a labyrinthine endeavor. By integrating a company's policy library, which includes critical documents like the Information Security Policy and Acceptable Use Policies, into CADDIE's comprehensive knowledge base, organizations benefit from a significant reduction in risk uncertainty. These insights can range from identifying potential policy deviations, offering corrective action recommendations, to providing context-aware explanations for compliance-related queries, thereby enabling organizations to maintain a robust security posture.

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

VI. PILOT STUDY

On September 21st, 2023, a risk and compliance Commanding Officer of NYPD's Legal Bureau and Executive Officer of the Intelligence Operations and Analysis Bureau, was graciously elected to assess the credibility and value proposition of the CADDIE application. Given his background in intelligence and operations, our test candidate provided qualified insights into the system's functionalities.

A. Test Scenario

After a briefing on the capabilities and data sources of CADDIE, the tester was encouraged to explore the system by querying themes relevant to his experience: Threat Intelligence Awareness, Upskilling, and Vulnerability Awareness. For real-time feedback, the interaction between the queries and CADDIE's responses it showcased the Slack integration as an advisor bot. The subsequent prompts and results acted as significant indicators of the application's proficiency in addressing these crucial cybersecurity themes. Please note that some bot responses were truncated due to the token threshold restriction to deliver results.

1. Upskilling

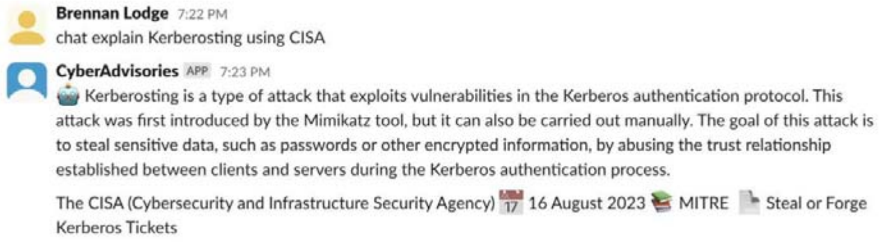
To ensure CADDIE's efficacy in both standard and edge-case scenarios, we subjected it to tests, one of which focused on the term "Kerberosting." This approach not only gauges the tool's capability to upskill junior cybersecurity analysts on less conventional threats but also its proficiency in delivering quality, cross-referenced results even for less common inquiries.

"Chat, explain Kerberoasting using MITRE"

Bot Response:

Kerberosting is a type of attack that exploits vulnerabilities in the Kerberos authentication protocol. This attack was first introduced by the Mimikatz tool, but it can also be carried out manually. The goal of this attack is to steal sensitive data, such as passwords or other encrypted information, by abusing the trust relationship established between clients and servers during the Kerberos authentication process.

- RAG References:
- Date: 16 August 2023 Source: MITRE
- Title: Steal or Forge Kerberos Tickets

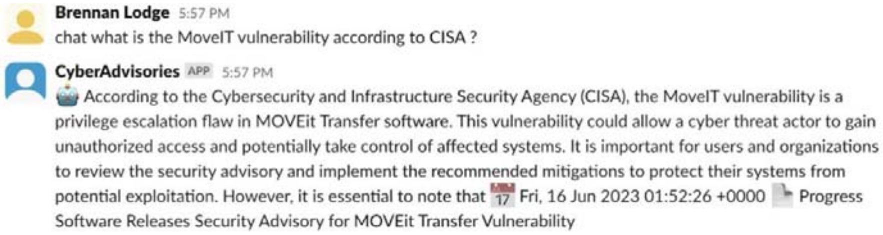


Brennan Lodge 7:22 PM
chat explain Kerberosting using CISA

CyberAdvisories APP 7:23 PM
Kerberosting is a type of attack that exploits vulnerabilities in the Kerberos authentication protocol. This attack was first introduced by the Mimikatz tool, but it can also be carried out manually. The goal of this attack is to steal sensitive data, such as passwords or other encrypted information, by abusing the trust relationship established between clients and servers during the Kerberos authentication process.

The CISA (Cybersecurity and Infrastructure Security Agency) 16 August 2023 MITRE Steal or Forge Kerberos Tickets

Fig. 6 Slack Application for Upskilling SOC Analysts



Brennan Lodge 5:57 PM
chat what is the MoveIT vulnerability according to CISA ?

CyberAdvisories APP 5:57 PM
According to the Cybersecurity and Infrastructure Security Agency (CISA), the MoveIT vulnerability is a privilege escalation flaw in MOVEit Transfer software. This vulnerability could allow a cyber threat actor to gain unauthorized access and potentially take control of affected systems. It is important for users and organizations to review the security advisory and implement the recommended mitigations to protect their systems from potential exploitation. However, it is essential to note that Fri, 16 Jun 2023 01:52:26 +0000 Progress Software Releases Security Advisory for MOVEit Transfer Vulnerability

Fig. 7 Slack Application for Upskilling Threat Intelligence Analysts

2. Vulnerability Awareness

In this final test case scenario, we probed the chat bot's capability with regard to vulnerability awareness, specifically referencing the "MoveIT vulnerability." This testing not only evaluated the tool's ability to inform cybersecurity analysts about current software vulnerabilities but also assessed its precision in pulling detailed, cross-referenced insights even for specific security advisories.

Prompt 1C

"Chat what is the MoveIT vulnerability according to CISA?"

Bot Response:

According to the Cybersecurity and Infrastructure Security Agency (CISA), the MoveIT vulnerability is a privilege escalation flaw in MOVEit Transfer software. This vulnerability could allow a cyber threat actor to gain unauthorized access and potentially take control of affected systems. It is important for users and organizations to review the security advisory and implement the recommended mitigations to protect their systems from potential exploitation. However, it is essential to note that …

RAG References:

- Date: Fri, 16 Jun 2023 01:52:26 +0000
- Source: CISA Advisory
- Title: Progress Software Releases Security Advisory for MOVEit Transfer Vulnerability

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf

3. GRC Use Case

GRC practitioners continually face the challenge of aligning company policies with evolving regulatory frameworks. The CADDIE framework presents a solution in this domain. In the example below, we provide an inquiry about the alignment of a company's hypothetical SANS Policies with the New York State or NYCRR 500 cybersecurity requirements for financial services companies. CADDIE provides a precise breakdown of how the company's policies align with the specified regulations, pinpointing areas of compliance and highlighting where updates may be needed.

In the example below, the platform efficiently cross-references company policies with the nuances of a regulatory document, providing actionable insights. This example reinforces the ability of CADDIE to merge regulatory documentation with company policies, ensuring efficient compliance mapping.

prompt=("User: "If the SANS Policies are our hypothetical company, are any of our companies policies mentioned in the New York State or NYCRR 500 CYBERSECURITY REQUIREMENTS FOR FINANCIAL SERVICES COMPANIES?")

response: Yes, some of your company's policies might be mentioned in the New York State or NYCRR 500 CYBERSECURITY REQUIREMENTS FOR FINANCIAL SERVICES COMPANIES. The regulations address various aspects of cybersecurity, including risk assessment, access control, data protection, incident response, and more. Your company might already have policies in place that align with these requirements, such as:

Risk Assessment: If your company has a policy to regularly assess and identify potential risks to its information technology systems and customer information, it may be compliant with the NYCRR 500 requirements.

Access Control: If your company has a policy to ensure that only authorized individuals have access to sensitive data and IT systems, this aligns with the access control requirements in the regulations.

Data Protection: If your company has policies in place to protect sensitive data through encryption, secure storage, and proper disposal methods, these practices may already meet some of the data protection requirements mentioned in NYCRR 500.

Incident Response: If your company has a policy to promptly respond to and investigate any security incidents or unauthorized access attempts, this aligns with the incident response requirements outlined in the regulations.

However, it is important to note that while some of your company's policies might already be compliant with NYCRR 500, the regulations also introduce new minimum standards for cybersecurity programs. Your company may need to review and potentially update its policies and procedures to fully meet these new requirements.

metadata={'Date': '2023-10-18', 'Source': 'Regulatory', 'Subject': 'new york state department of financial services 23 nycrr 500 cybersecurity requirements for financial services companies* section 500.0 introduction.', filepath': './Regulatory_documents/NYCRR500_0.pdf'})]}

VII. CONCLUSION

This exploration of RAG technology has distilled its application into a system designed to navigate the intricacies of regulatory landscapes with precision and insight. By integrating sophisticated algorithms and models, this approach redefines the conventional methodologies used in policy analysis and compliance strategy development.

Throughout the paper, we have unveiled how leveraging such an intelligent system transcends mere automation. It offers an intuitive framework for analyzing regulatory texts, revealing gaps, and crafting policies aligned with the latest international standards. This is paramount as organizations grapple with the complexities of various data privacy laws, aiming to enhance their overall risk posture and maintain robust compliance practices.

Our discussions and case studies have outlined the system's capability to translate vast volumes of regulatory information into actionable intelligence. This enables a deeper understanding of compliance obligations across different jurisdictions, reflecting a forward-thinking approach to global regulatory challenges.

CADDIE serves as an ally for professionals seeking to ensure that organizational policies are not only current but also anticipate future legal trends. This proactivity is vital in an era where the stakes of regulatory non-compliance are ever-increasing, and the need for a proactive defense against potential breaches is critical. It stands as a testament to the promise of such technology in supporting a proactive and strategic vision for risk and compliance management in a rapidly evolving digital landscape.

REFERENCES

[1] Bureau of Labor Statistics, U.S. Department of Labor, Occupational Outlook Handbook, Information Security Analysts, at https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm (visited September 06, 2023).
[2] UNCTAD (United Nations Conference on Trade and Development). (n.d.). Summary of Adoption of E-commerce Legislation Worldwide. UNCTAD Global Cyberlaw Tracker. Retrieved from https://unctad.org/topic/ecommerce-and-digital-economy/ecommerce-law-reform/summary-adoption-e-commerce-legislation-worldwide
[3] Morgan, S. (2023, April 14). Cybersecurity Jobs Report: 3.5 Million Unfilled Positions in 2025. Cyber Crime Magazine. https://cybersecurityventures.com/jobs/
[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Facebook AI Research; University College London; New York University. Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (v4). Available at: https://arxiv.org/abs/2005.11401
[5] Riedel, S., Kiela, D., Lewis, P., & Piktus, A. (2020). Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models. ML Applications, Open Source. https://ai.meta.com/blog/retrieval- augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/
[6] Anand, Y., Nussbaum, Z., Duderstadt, B., Schmidt, B. M., Treat, A., & Mulyar, A. (2023). GPT4All: An ecosystem of open source assistants that

run on local hardware. Nomic AI.
[7] Pinecone Company. (2023). Sentence Transformers: Meanings in Disguise. https://www.pinecone.io/learn/series/nlp/sentence-embeddings/
[8] SANS Organization. (n.d.). Security Policy Templates. Retrieved from https://www.sans.org/information-security- policy/
[9] Checkpoint Research. (2023). Global Cyberattacks Continue to Rise with Africa and APAC Suffering Most. Retrieved from https://blog.checkpoint.com/research/global-cyberattacks-continue-to-rise/
[10] "Getting Started with Data Protection." Information Commissioner's Office, n.d., https://ico.org.uk/for-organisations/advice-for-small-organisations/frequently-asked-questions/getting-started-with-data-protection/#receivescomplaint.
[11] "Maryland's New Children's Privacy Laws and Their Impact on Technology." The New York Times, 7 Apr. 2024, https://www.nytimes.com/2024/04/07/technology/maryland-children-privacy-laws-technology.html.
[12] "The Biggest GDPR Fines So Far." Termly, n.d., https://termly.io/resources/articles/biggest-gdpr-fines/

Open Science Index, Information and Communication Engineering Vol:18, No:7, 2024 publications.waset.org/10013734.pdf