



AI in Finance Summit New York 2024

RAG against the machine

Brennan Lodge
BLodgic Inc.





Agenda

- . Intro
- . AI as a guide not a guard
- . The good, the bad, the ugly of AI in cybersecurity
- . Briefer on AI
- . RAG against the machine
- . Roll your own
- . Demos
- . Next steps



Brennan.Lodge Whois Lookup

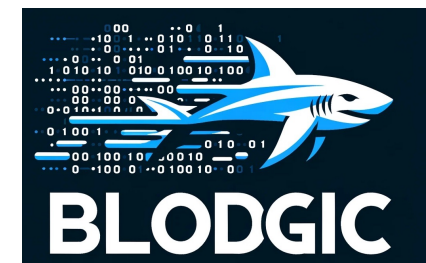
- **Role:** Co-founder and CTO
- **Experience:**
 - Over 15 years in financial services focusing on cybersecurity, data science, and leadership.
 - Previous affiliations include **JP Morgan Chase**, **Federal Reserve Bank of New York**,
 - **Bloomberg**, **Goldman Sachs**, and **HSBC**.
 - Led AI-centric cyber solutions at HSBC, streamlining cybersecurity processes.
 - **Professor at NYU**, Information Tech Management & Data Analytics.
 - Technical advisor roles at CounterFlow AI, Inc. and DataKind.
 - Published a peer-reviewed paper on machine learning for security incident response teams.
- **Expertise:**
 - Seasoned veterans in cyber defense strategies, leveraging AI and machine learning solutions.
 - Proficient in threat analysis, incident detection, and response.
 - Proven track records in conceptualizing, designing, and executing cybersecurity solutions.



ML x Cases Finance/ESG
competition winner 2023



US Cyber Command
AI RPE winner 2022



The GOOD of... AI in Cybersecurity?

- **Information Overload**

- Cybersecurity pros face a deluge of threat data; AI can filter and prioritize this to manageable levels.

- **24/7 Defense Needs**

- AI operates continuously, supporting human teams beyond regular work hours.

- **Talent Gap & Burnout**

- AI mitigates the industry's talent shortage and reduces professional burnout by automating routine tasks.

- **Rising Attack Volume**

- While the demand for cybersecurity 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
- 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.

- **Empowering Analysts**

- AI provides rapid context, enhancing decision-making and easing the strain on analysts.

- **Complex Integration Challenges**

- Tailoring AI solutions to address privacy, cost, and operational concerns in cybersecurity infrastructures.



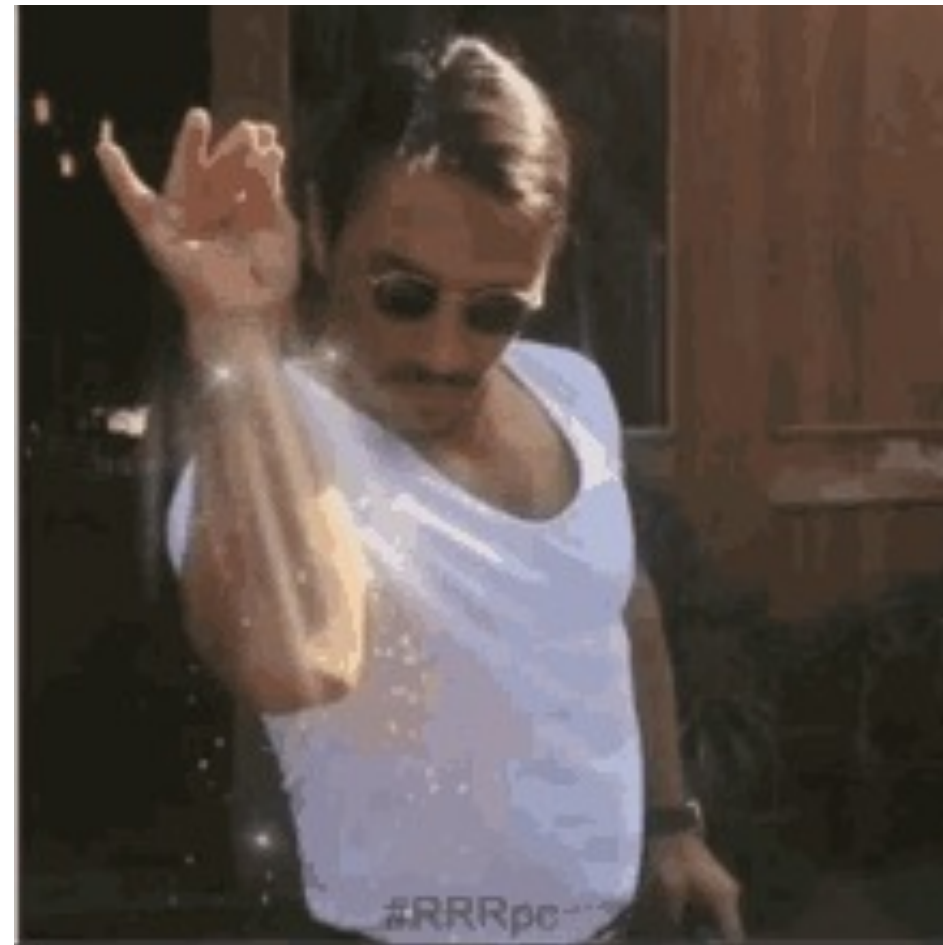
The Bad

...sprinkle a little AI on it

- Shift-left
 - Vendor review
 - Security design review
 - Threat modeling
 - Code analysis
- Production surface
 - Firewalls, load balancers, email, DNS
 - Identity, access management, zero-trust
 - Endpoint, network, cloud, shadow IT, SaaS
 - Public site & intranet
 - Applications and their behavior
 - Compliance & audits
- Security operations
 - Help desk
 - Digital forensics & incident response
 - Patch management
 - Threat hunting & threat intel
 - Detection engineering
 - Security awareness programs

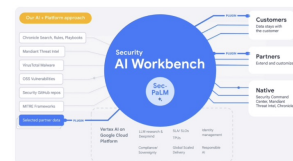
To Do What?

- SOAR 2.0
- Data co-pilot; smart automation; autonomous investigation; threat intel handling
- Detection engineering
Easier, cover text columns, add smarter decision making; questions of scale, fidelity, model
- Answer Third party risk reviews
- Upskilling: Junior analysts, coding,



Siloed platforms

Cross-silo platforms



Google Launches Security AI Workbench to Rival Microsoft's Security Copilot

Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity

May 30, 2023 Michael Sentonas Endpoint & Cloud Security

Announcing

Microsoft Security Copilot

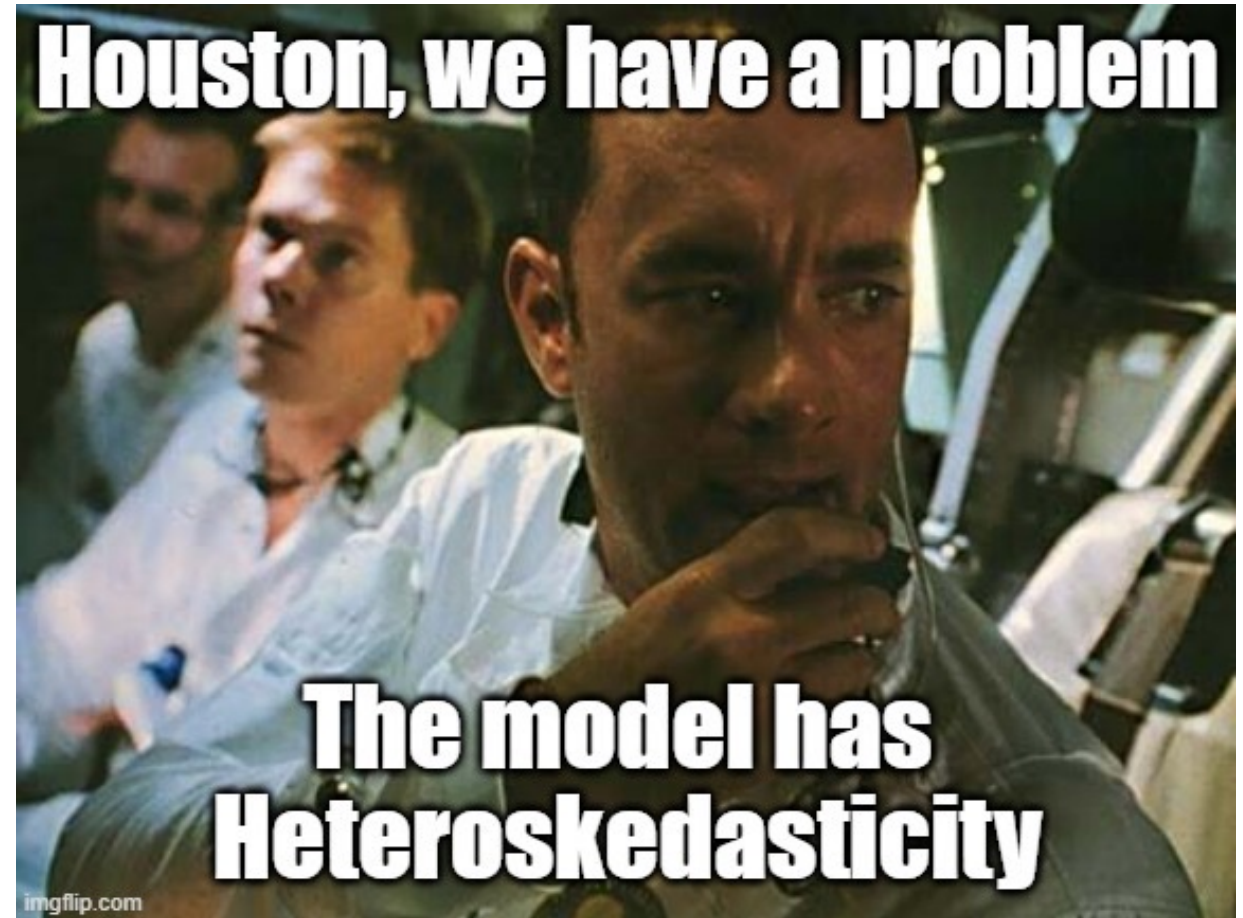
 Dropzone AI

VECTRA®



The Ugly ...What the problem is ?

- GenAI Regulation? Ha. What's that?
- "We have hundreds of internal genAI projects, but any new one must go through X, who is still making a policy on it"
- "All genAI projects are stopped, except 1 pilot for the rest of the year"
- "Our agency does not allow genAI, but 3 of our sister agencies do"
- "Unlike American regulators, we're letting banks take the lead in how AI may be used"
- "We're scrambling to come up with an internal policy"
- "We can't stop our users from using ChatGPT"
- We can implement it, but how do we explain it?

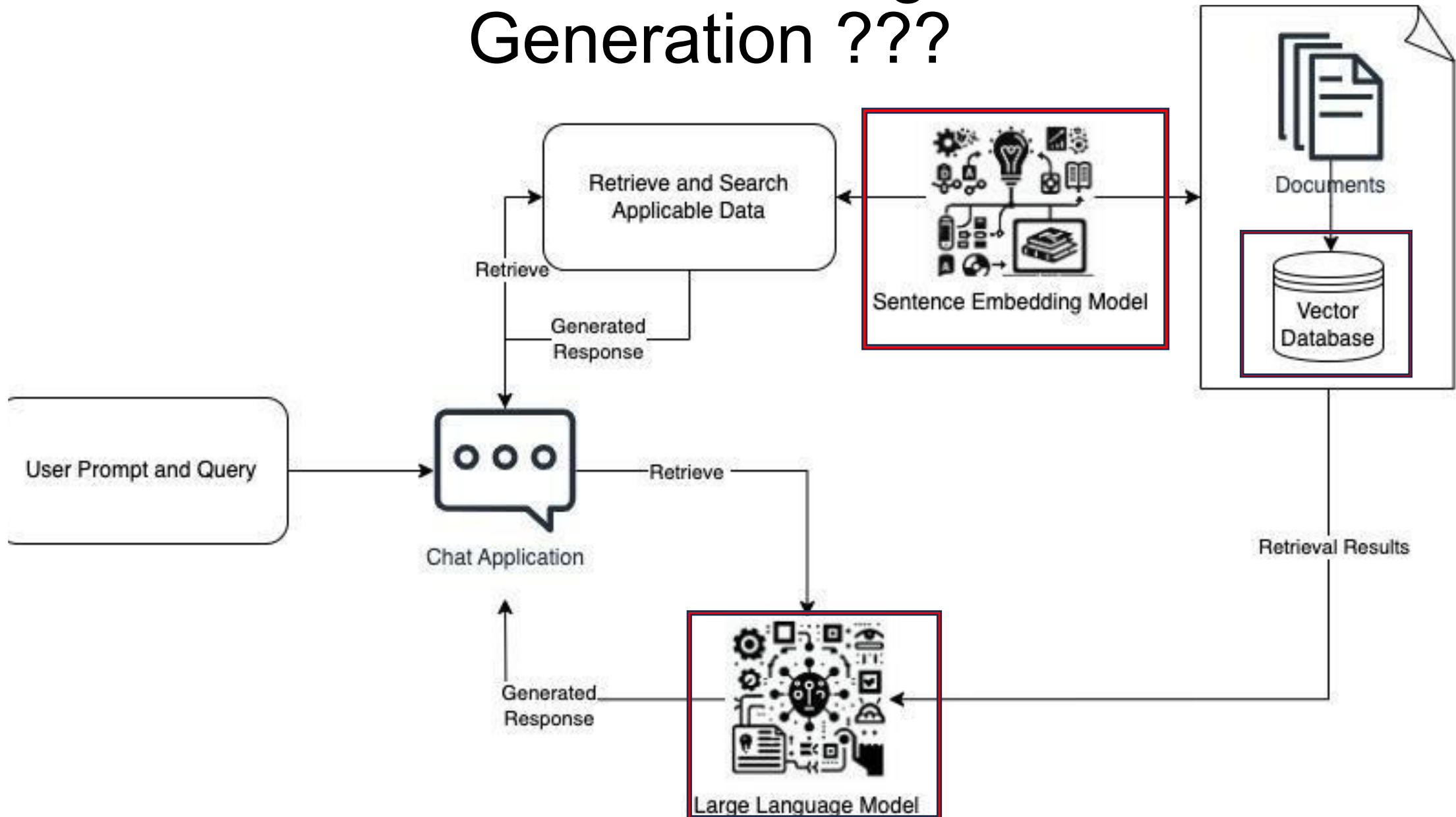


AI as a Guide, Not a Guard

- Understanding Generative AI
- GPT & LLMs:
 - These models analyze and generate text, aiding in threat detection and alerts.
- AI as an Informative Ally:
 - AI guides cybersecurity experts through the maze of threats, risks, and data, providing them with actionable insights.
- Not Owners of Data via API requests (OpenAI):
 - While powerful, they process information without retaining proprietary data, aligning with privacy concerns.
- Purpose: Democratize your cybersecurity tools
 - Real-time AI-driven cybersecurity advisory.
- Rapid Analysis:
 - Instantaneous response.
- Data-driven:
 - Leverages sources like CISA & MITRE ATT&CK, NIST, Information Security Policies, and more.
- Integration:
 - Fits into chat systems for proactive security.



What is Retrieval Augmented Generation ???



Roll your own data

Innovation:

- **RAG + LLM Synergy:** Enhanced cybersecurity analysis through the confluence of Retrieval Augmented Generation and Large Language Models.
- **Real-time Response:** Quick and precise threat intelligence with state-of-the-art AI-driven mechanisms.

Comparison with Existing Solutions:

- **Quicker:** Reduces the time lag in threat detection and response, moving from reactive to proactive defense.
- **Comprehensive:** Taps into vast datasets, including MITRE ATT&CK and CISA advisories for in-depth analysis.
- **Cost-effective:** Democratized access to top-tier AI, making it accessible to all teams.

Feasibility and Potential:

- **Tested Prototype:** Achieves response times within 10-second benchmarks, indicating robustness. Achieved a cost of < \$500 per month (*milage may vary*)
- **Adaptable:** Compatible with both on-premises and cloud deployments, ensuring flexibility for diverse organizations.

"Redefining Cybersecurity Analysis with AI"



How to roll your own data

1. **Prepare Documents:** Start by collecting data. Integrate your Threat Intel feeds, CISA advisories, Policy documents, etc to build a dataset for our RAG application.
2. **Split and Create Embeddings:** Break down the collected documents into smaller snippets and use an embedding model to convert them into compact vector representations. This involves utilizing a splitter,
3. **Build a LangChain + ChromaDB:** Set up the LangChain by combining a prompt generator for context creation, a Retriever for fetching relevant snippets, and an LLM to answer queries from the data you sent to ChromaDB.
4. **Ask a Question:** Learn how to ask questions to the RAG application.
5. **Visualize:** Use Renumics-Spotlight visualize the embeddings in 2D with UMAP and analyze the relationships and proximities between queries and document snippets.

#parsing code

```
!pip install pypdfium2  
|  
https://pypi.org/project/pypdfium2/
```



Hugging Face

 sentence-transformers/[all-MiniLM-L12-v2](#)

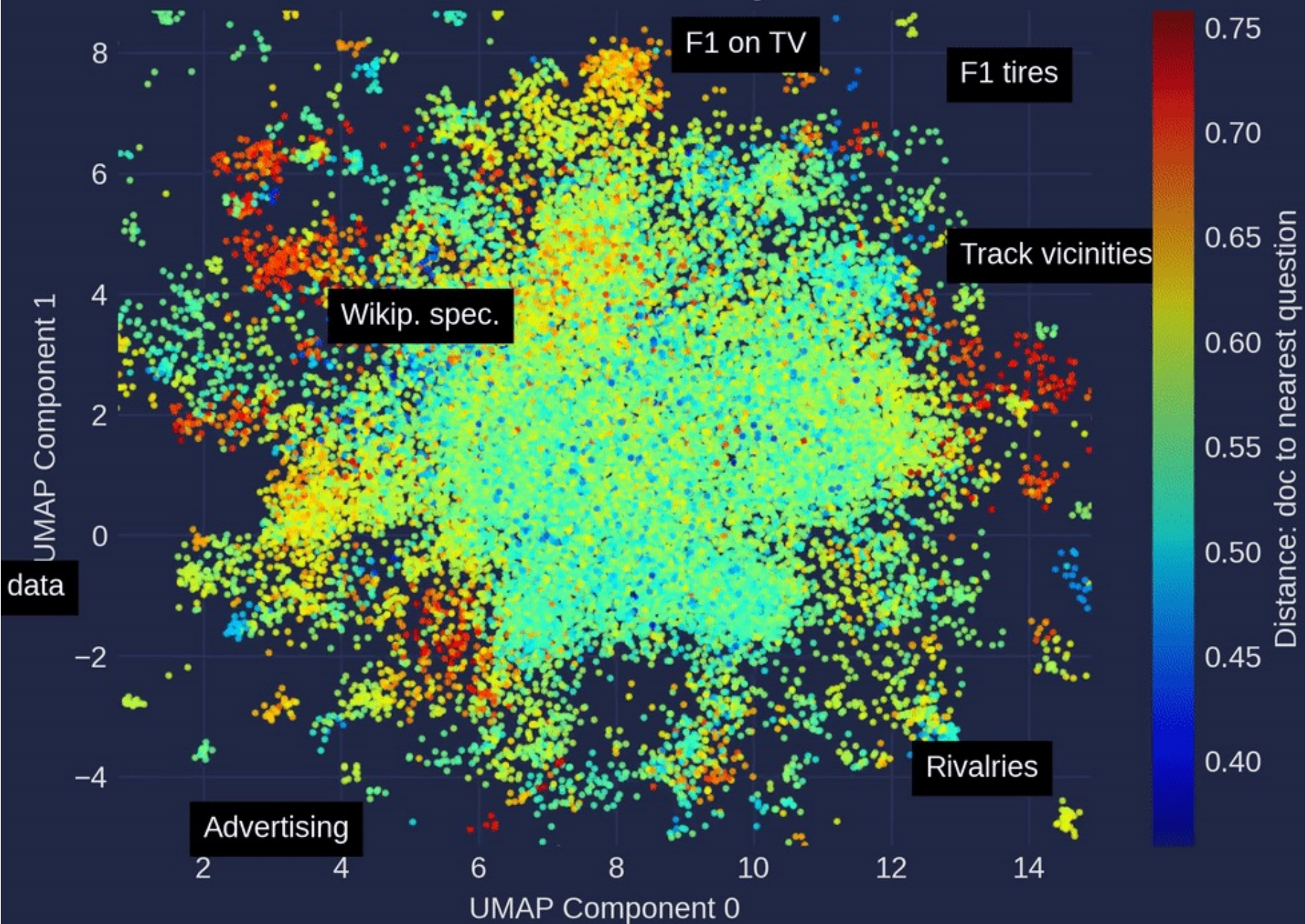
 nomic-ai/[gpt4all-j](#) 

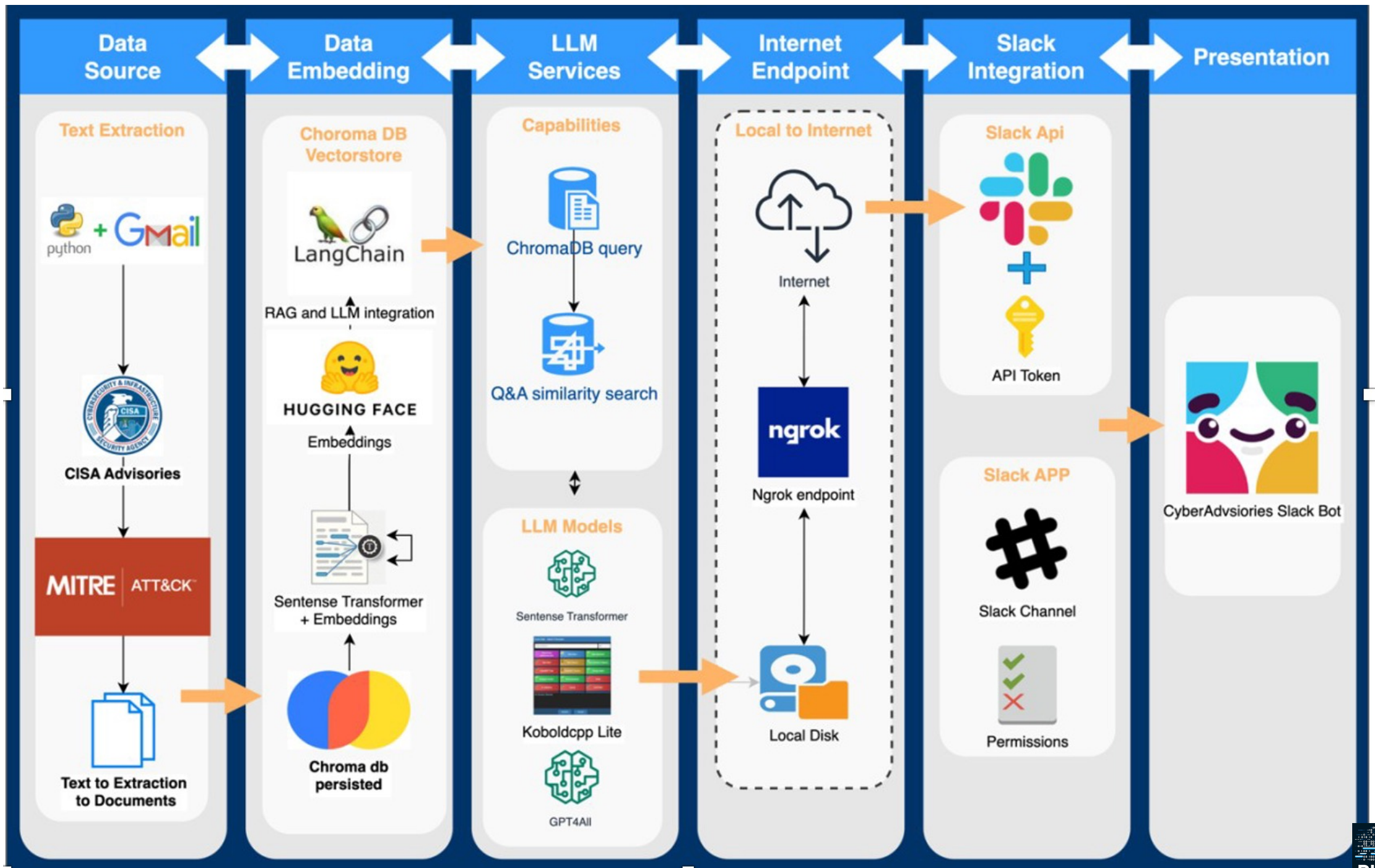
atlas by NOMIC



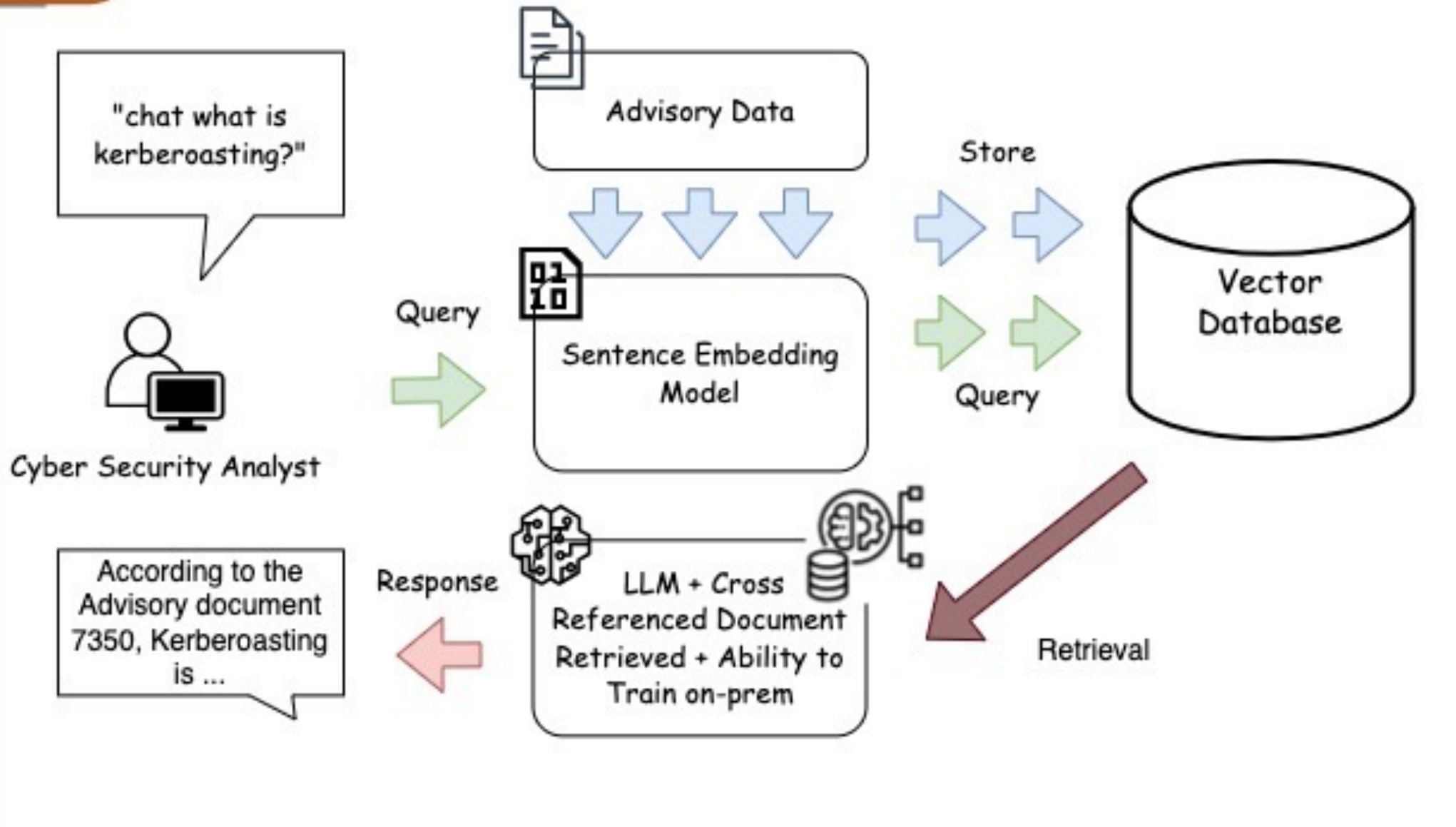
LangChain + Chroma

UMAP on document embedding, Iteration 6

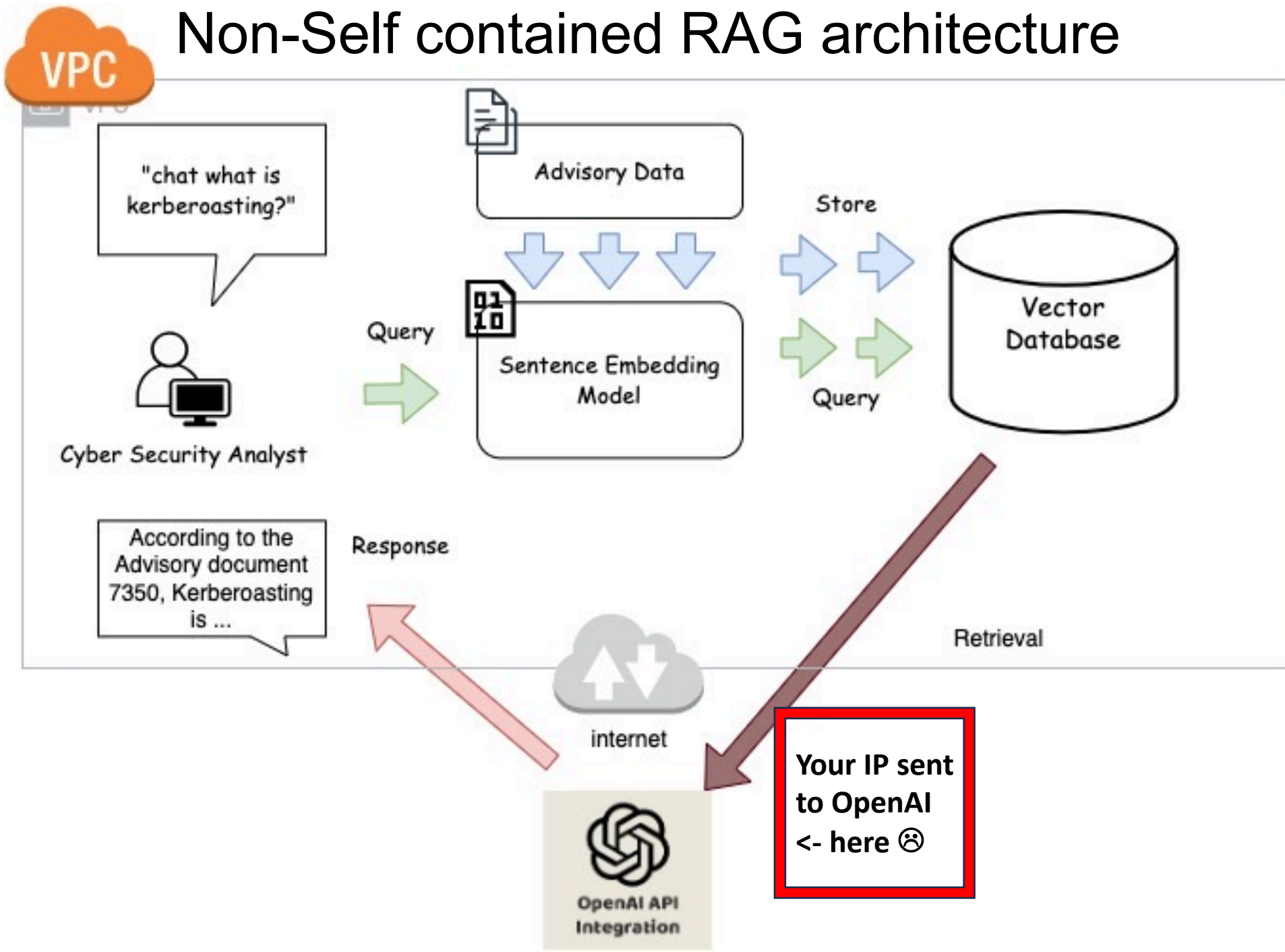




Self contained RAG architecture



Non-Self contained RAG architecture



Typical usage for a RAG solution



GRC or Cyber Security Analysts X5



20 queries each day
requests at 5000 total
tokens utilization X 5
analysts = ~100,000k
tokens each day

Co\$t breakdown for the infra

Full self contained RAG in your VPC

g4dn.4xlarge

The g4dn.4xlarge instance is in the gpu instance family with 16 vCPUs, 64.0 GiB of memory and up to 25 Gbps of bandwidth starting at \$1.204 per hour.

💰 Pricing

\$1.204 On Demand
\$0.5196 Spot
\$0.7590 1 Yr Reserved
\$0.5200 3 Yr Reserved

US East (N. Virginia) ▾

Linux ▾

Per Hour ▾

No Upfront ▾

\$379.60 per month (-57%) with [Autopilot](#)

[Learn more](#)

📏 Family Sizes

Size	vCPUs	Memory (GiB)
g4dn.xlarge	4	16
g4dn.2xlarge	8	32
g4dn.4xlarge	16	64

📘 Instance Details

Compute	Value
vCPUs	16
Memory (GiB)	64.0
Memory per vCPU (GiB)	4.0
Physical Processor	Intel Xeon Family
Clock Speed (GHz)	2.5
CPU Architecture	x86_64
GPU	1
GPU Architecture	nvidia t4 tensor core
Video Memory (GiB)	16
GPU Compute Capability (?)	7.5
FPGA	0

Networking	Value
Network Performance (Gbps)	Up to 25
Enhanced Networking	true
IPv6	true

~\$500 per month

Co\$t breakdown for the infra

EC2 instance + 100k tokens request per day

t2.large

The t2.large instance is in the general purpose family with 2 vCPUs, 8.0 GiB of memory and low to moderate network performance starting at \$0.0928 per hour.

💰 Pricing

\$0.0928 On Demand	\$0.0425 Spot	\$0.0575 1 Yr Reserved	\$0.0401 3 Yr Reserved
------------------------------	-------------------------	----------------------------------	----------------------------------

US East (N. Virginia) ▾

Linux ▾

Per Hour ▾

No Upfront ▾

\$29.27 per month (-57%) with [Autopilot](#)

[Learn more](#)

🚩 Family Sizes

Cost Calculation for 100,000 Tokens per Day

GPT-3.5 Turbo-0125 Model

- **Total Cost per 1,000 Tokens:** \$0.0020 (since this includes both input and output costs as explained before)
- **Daily Cost for 100,000 Tokens:** $100,000 \text{ tokens} / 1,000 * \$0.0020 = \$0.20 * 100 = \20.00
- **Monthly Cost (30 Days):** $\$20.00 * 30 = \600.00

~\$600 per month

RAGs can be secure and privacy conscious...not the AI conscious kind

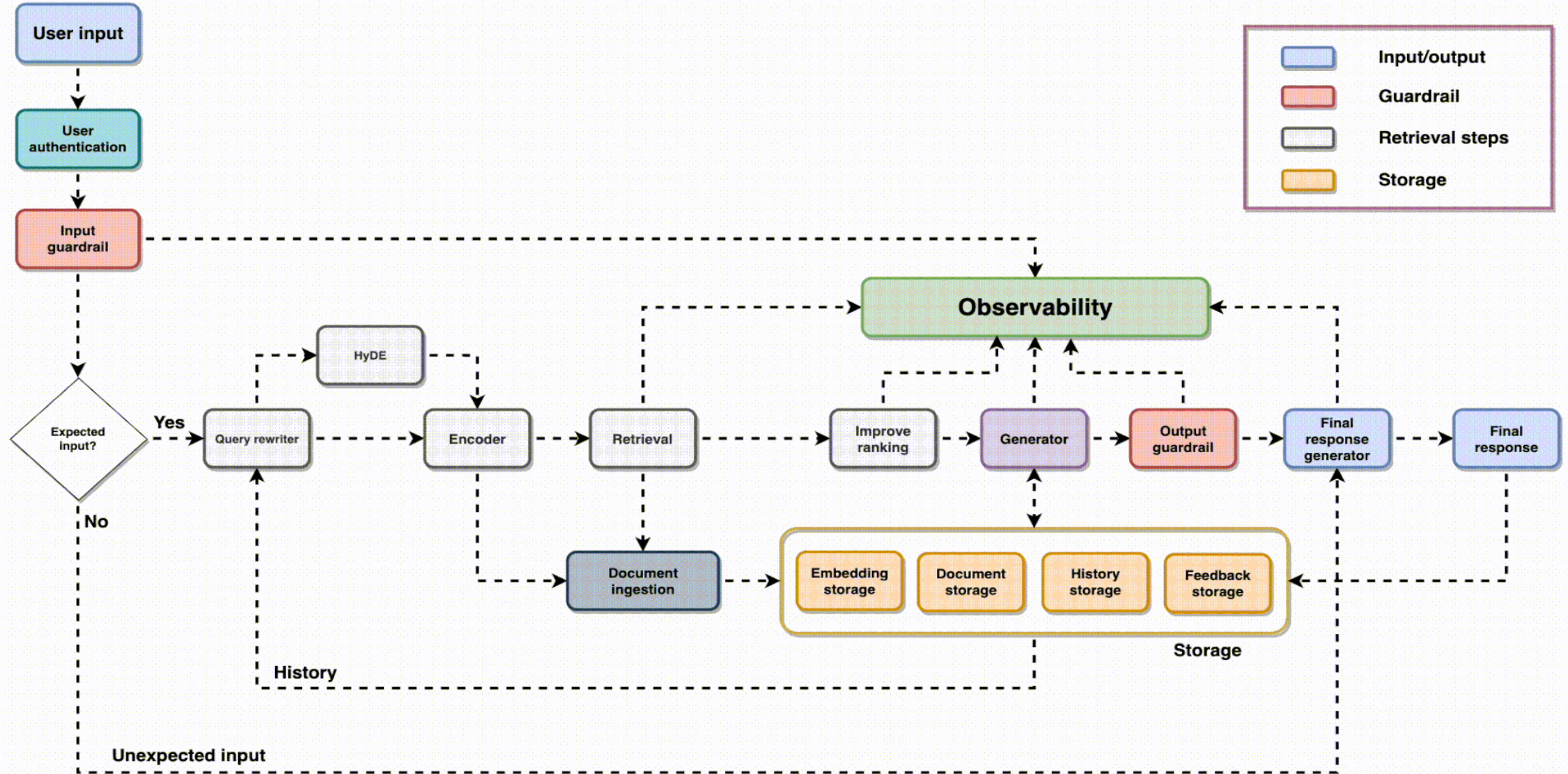
- **User Authentication**
 - Ensures only authorized access, enhancing security and privacy.
- **Input Guardrails**
 - Protect against harmful inputs and maintain data integrity.
- **Query Rewriter**
 - Clarifies user queries for precision and relevance.
- **Document Ingestion and Encoders**
 - Manages data processing and converts text to vectors for retrieval.
- **Vector Databases**
 - Supports efficient document retrieval, balancing recall and latency.
- **Techniques for Improving Retrieval**
 - Utilizes methods like Hypothetical Document Embeddings (HyDE) for more accurate information retrieval.
- **Output Guardrails**
 - Ensures responses are accurate, relevant, and secure.

The Phoenix Project



The Phoenix Project: A Novel about IT, DevOps, and Helping Your Business Win

Architecture For Enterprise RAG



Introducing CADDIE: *The Future of AI- Driven Cybersecurity*

- What is CADDIE?
- **Cybersecurity Ai Designed Defense Intelligence Engine**
- We make a cyber security generative AI tool for organizations who need to inform and upskill their cyber security defenders so they can solve cybersecurity challenges
- An advanced AI-driven tool and “*retriever*” designed to revolutionize cybersecurity analysis.
- **Core Technologies:**
 - **RAG (Retrieval Augmented Generation):**
 - Seamlessly blends retrieval and generation for effective threat detection.
 - **LLM (Large Language Model):**
 - Powering comprehension, prediction, and problem-solving for intricate cybersecurity challenges.



Innovation:

Comparison with Existing Solutions:

- **Quicker:** Reduces the time lag in threat detection and response, moving from reactive to proactive defense.
- **Comprehensive:** Taps into vast datasets, including MITRE ATT&CK and CISA advisories for in-depth analysis.
- **Cost-effective:** Democratized access to top-tier AI, making it accessible to all teams.

- **Tested Prototype:** Achieves response times within 10-second benchmarks, indicating robustness. Achieved a cost of < \$500 per month
- **Adaptable:** Compatible with both on-premises and cloud deployments, ensuring flexibility for diverse organizations.

"Redefining Cybersecurity Analysis with AI"



Advantages of CADDIE's Architecture

•Transition in Action

- By 2024, 75% of enterprises will shift from Gen AI concepts to actual production (Gartner).
- Companies are moving to build small, in-house AI models tailored to their specific needs.

•Market Direction

- Emphasis on specialized solutions (cybersecurity regulation and threats) over general-use models

•Ownership & Control:

- Complete ownership ensures adaptability and customizability tailored to specific organizational needs.

•Speed & Efficiency:

- Localized access results in faster query times and real-time responses.

• Enhanced Analysis

- Dynamic cross-referencing of real-time data offers deeper and more accurate insights.

• Your Data Stays Local

- Preserve your Intellectual Property
- Eliminates the risk of IP loss that arises from querying external databases or the internet.
- Enhanced Privacy & Security
 - Local data handling means minimized external exposure, reducing the risk of data breaches and staying compliant.
- While many LLM solutions risk IP loss by querying the internet, CADDIEs architecture ensures security, speed, and efficiency by leveraging localized resources and advanced RAG infrastructure.



CADDIE's Moat



Custom Sentence Embedding: Harnesses the power of fine-tuned models to understand company-specific terminology.



Unique to Your Business: Targets cybersecurity, risk, governance, and regulatory lexicons that general LLMs overlook.



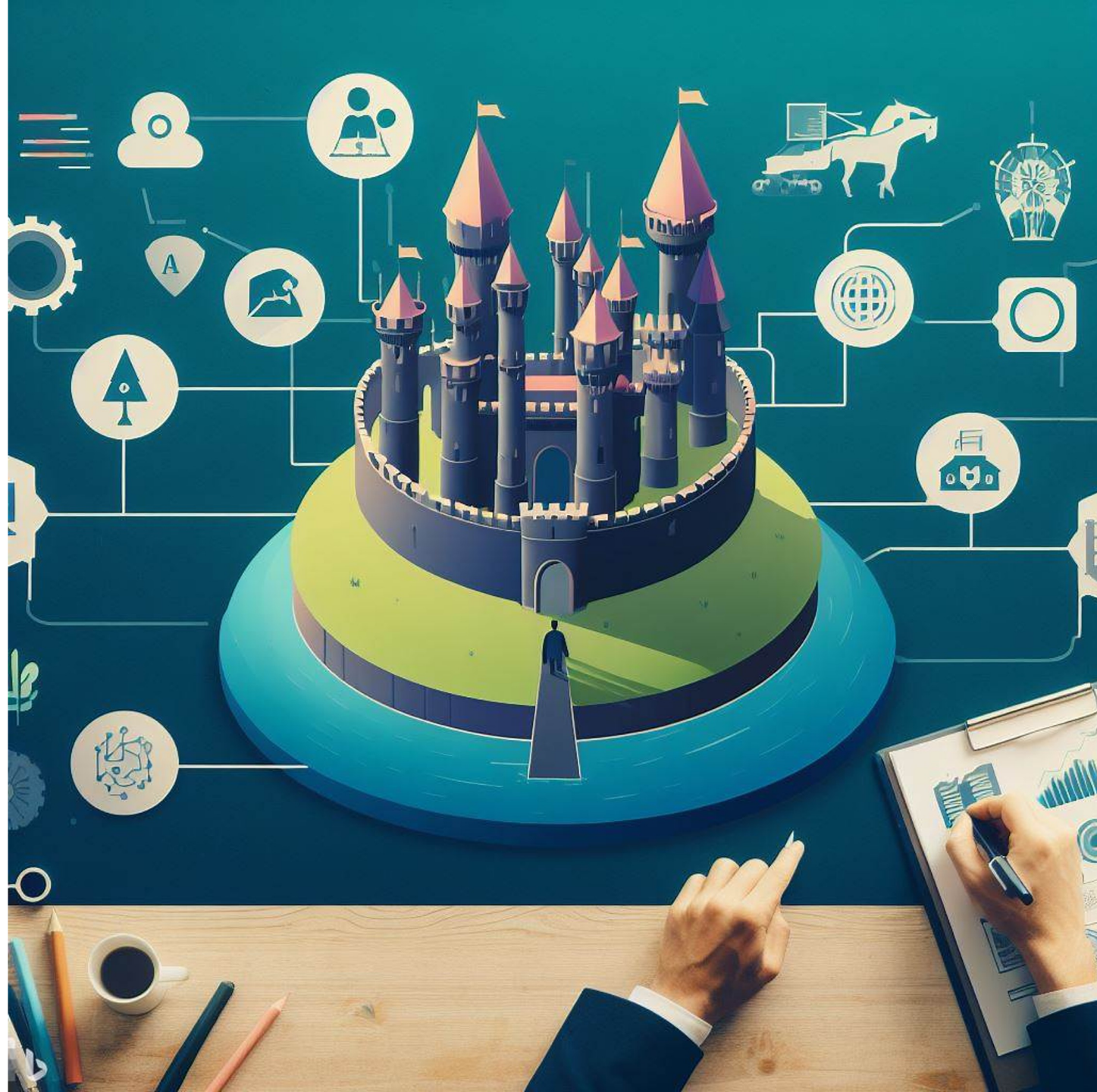
Deep Moat Strategy: Specialized integration ensures a competitive edge and high barriers for competitors.



Precision and Protection: Cybersecurity models can offer precise, context-aware responses to threats, adapted to your unique business environment.



Beyond Broad LLMs: While general models provide a one-size-fits-all solution, CADDIE custom embeddings ensure a glove-fit for your specific needs.

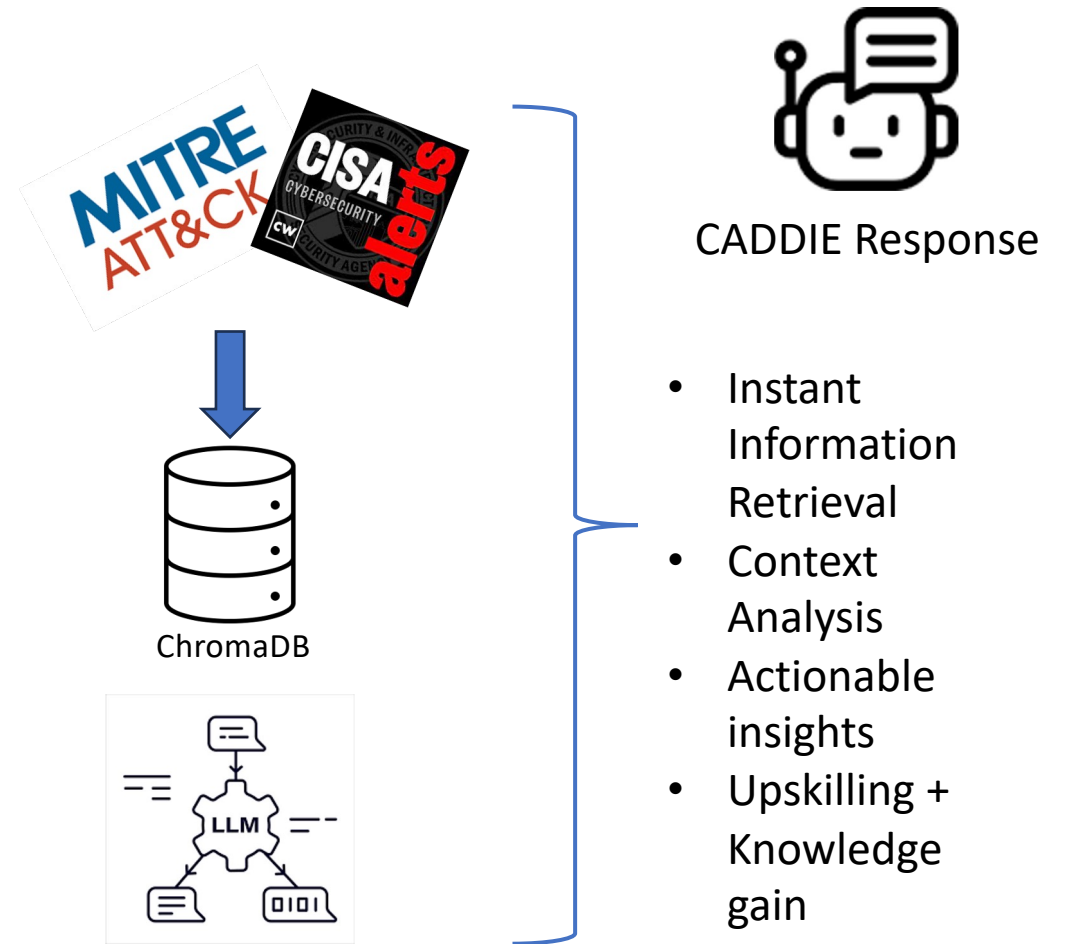


Leveraging Slack for a RAG solution

- **Innovative Integration:**
 - Utilize Slack's platform to integrate a specialized RAG system for dynamic cybersecurity collaboration.
- **Real-Time Interaction:**
 - Facilitate direct access to RAG-driven insights within Slack, streamlining the process of query and response.
- **Community Engagement:**
 - Join Slack's dedicated communities to discuss GRC, tech compliance, and cybersecurity, sharing the art of the possible with RAG technology.
- **Knowledge Sharing:**
 - Contribute use-cases and expert insights to foster an understanding of RAG's impact on cybersecurity operations.
- **Collaborative Alert System:**
 - Leverage Slack channels for distributing timely alerts and compliance notifications, enhancing organizational response capabilities.
- **User-Driven Growth:**
 - Develop a culture of sharing by encouraging user-driven referrals and highlighting the community's shared successes.
- **Success Narratives:**
 - Present narratives from users who have seen measurable improvements in their cybersecurity workflows, emphasizing community-driven innovation.
 - By incorporating RAG into Slack, we open a gateway to proactive, informed, and collaborative cybersecurity defense, tailored to the dynamic needs of modern organizations.



Use Case 2: CYBER SECURITY ADVISORY



+ Add a topic mark

**CyberAdvisories**

APR 3:25 PM

Tuesday, January 23rd

📄 Based on the provided context from the CISA Advisory, it appears that the latest threat actor being discussed is those who have exploited the Citrix CVE-2023-3519 vulnerability in order to implant webshells on a critical infrastructure organization's NetScaler ADC appliance. These threat actors were able to use the webshell to perform discovery on the victim's Active Directory and collect and exfiltrate AD data 📄 Thu, 20 Jul 2023 21:05:11 +0000 📄 CISA Advisory 📄 CISA Releases Cybersecurity Advisory on Threat Actors Exploiting Citrix CVE-2023-3519

Today

**Brennan Lodge** 7:34 PM

chat tell me about the MoveIT vulnerability?

**CyberAdvisories**

APR 7:34 PM

📄 The MoveIT vulnerability refers to three specific vulnerabilities (CVE-2023-36934, CVE-2023-36932, and CVE-2023-36933) that were discovered in the MOVEit Transfer software developed by Progress Software. These vulnerabilities could potentially be exploited by cyber threat actors to gain unauthorized access to sensitive information or even take control of affected systems. 📄 Fri, 07 Jul 2023 17:51:39 +0000 📄 CISA Advisory 📄 Progress Software Releases Service Pack for MOVEit Transfer Vulnerabilities

B I E ↻ ≡ ≡ ≡ 🔍 🗑

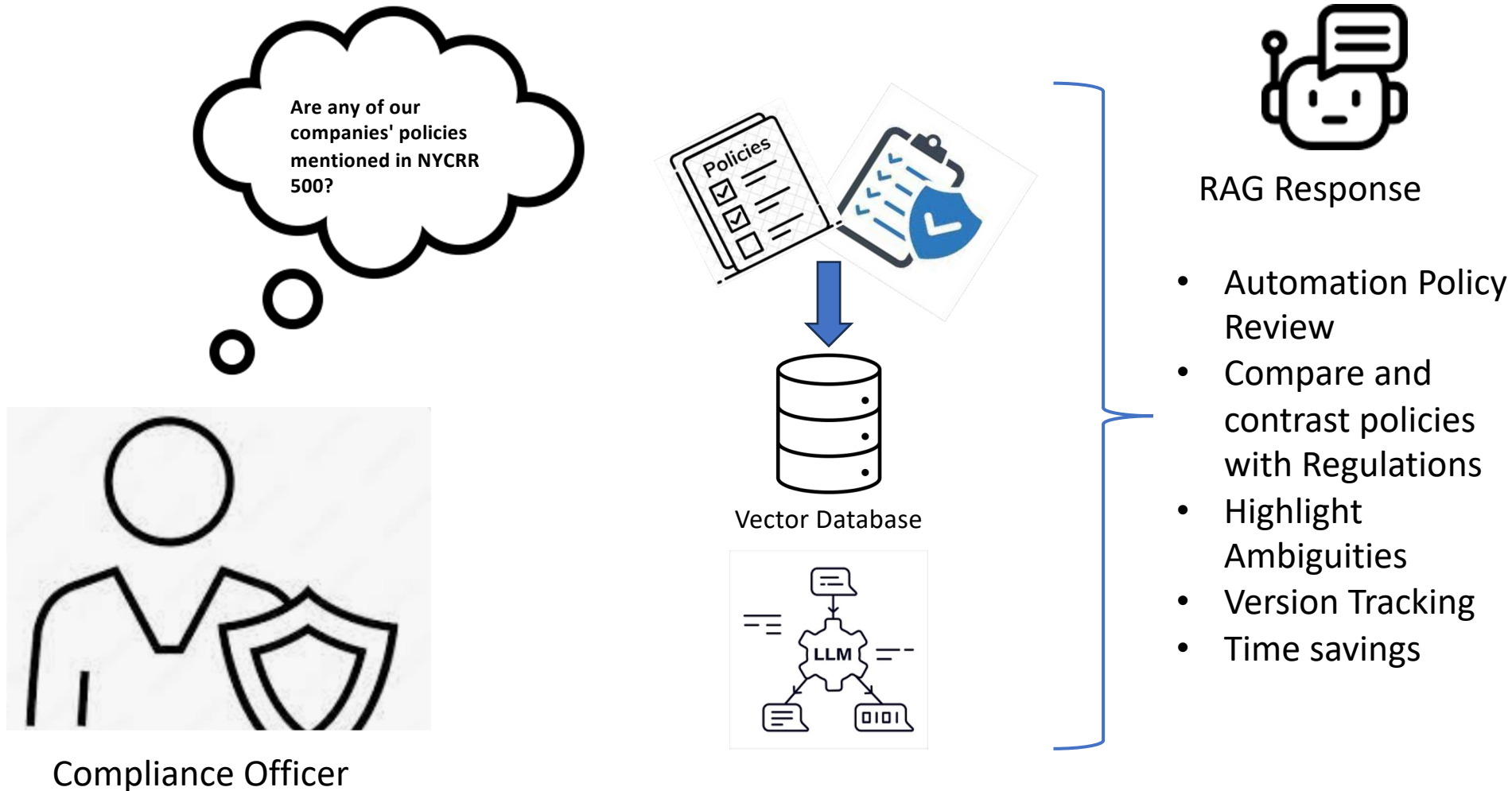
Message #cyberadvisor_bot

I

+ Aa @ 📎 📎 📎

➤

Use Case 1: Leveraging CADDIE for NYCRR 500 Compliance



After analyzing both sets of policies, here are some of the key requirements mentioned in the New York State regulations that may not be covered by our SANS Policies:

1. Cybersecurity program (Section 500.2): Each covered entity must develop and implement a cybersecurity program that is designed to protect its information systems and the information stored therein. This program should address identified risks, including those associated with external threats, internal vulnerabilities, and mobile devices.

2. Risk assessment (Section 500.3): Covered entities must conduct a risk assessment periodically to understand the potential risks and vulnerabilities to their information systems and nonpublic information. The risk assessment should be based on the size and complexity of the entity, as well as the nature and scope of its activities.

3. Cybersecurity policies and procedures (Section 500.4): Covered entities must establish cybersecurity policies and procedures to ensure the security of their systems and protect their information. These policies should address various aspects such as access control, data security, incident response, and security awareness training for employees.

4. Third-party service providers (Section 500.6): Covered entities must take appropriate steps to ensure the cybersecurity of any third-party service providers that have access to their nonpublic information. This includes assessing the risks posed by such service providers and requiring them to implement adequate cy

[]:

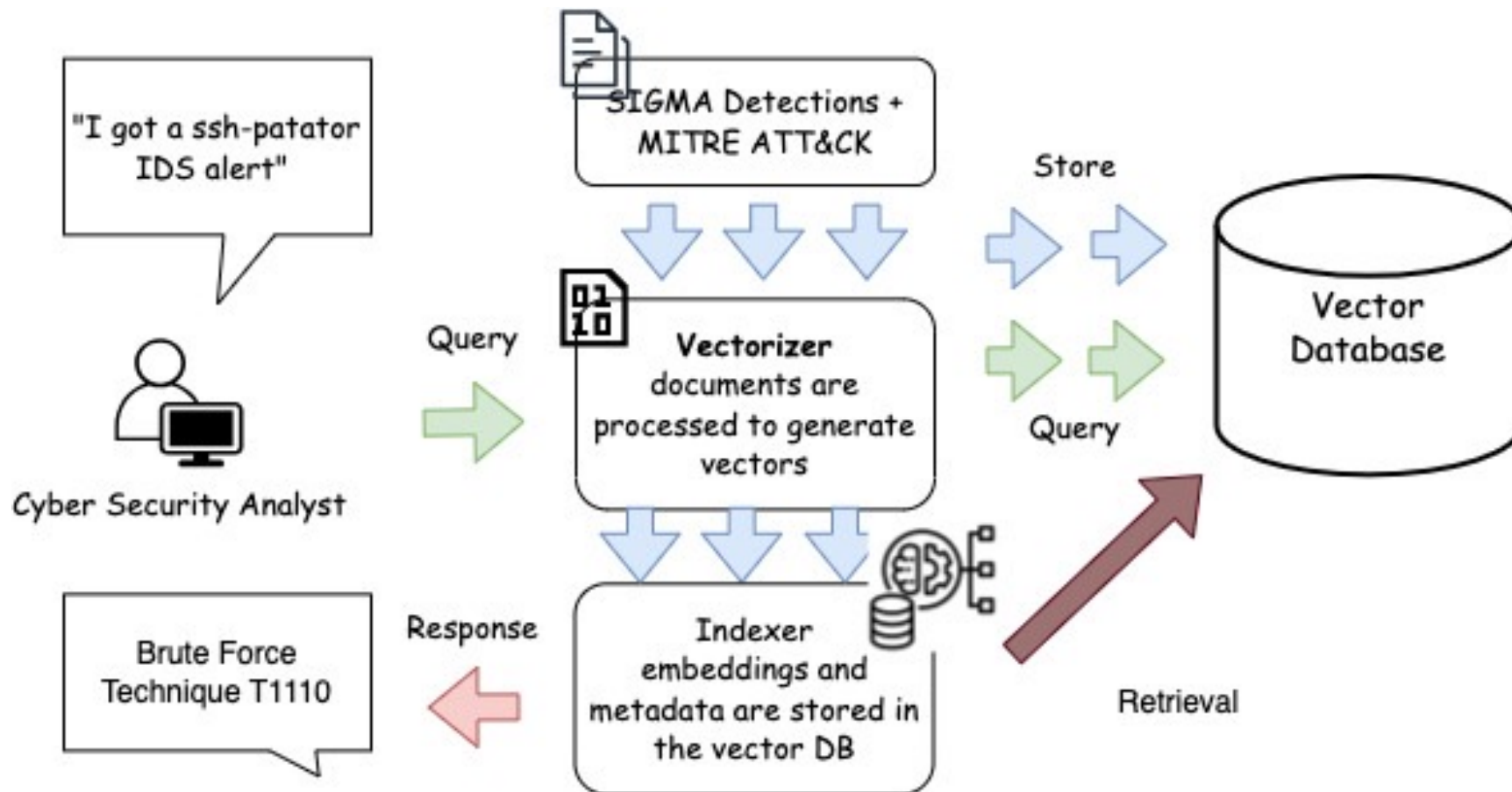
[]:

[]:

[]:

[]:

OPEN SOURCE PROJECT – MITREembed



- OpenSecurity + Data Community Discord
 - <https://discord.gg/dd9FzF29>
- MITRE EMBED Project
 - <https://github.com/deeps ecoss/MITREembed>

Next Steps for RAG



Feedback Loop Data Generation

Add feedback dialog into the chat after each bot response and record the results for future training and fine-tuning of the LLM



Cybersecurity Domain-Specific Embedding Model

Construct a specialized embedding model tailor-made for cybersecurity nuances and terminologies.



Real-time Data Feeds to the Vector Database

Refresh the data by integrating real-time data feeds into our vector database.



Pass Cybersecurity and Risk Certification exams with CADDIE

Prove value and knowledge by passing certification examples like CISSP, SANS, etc.

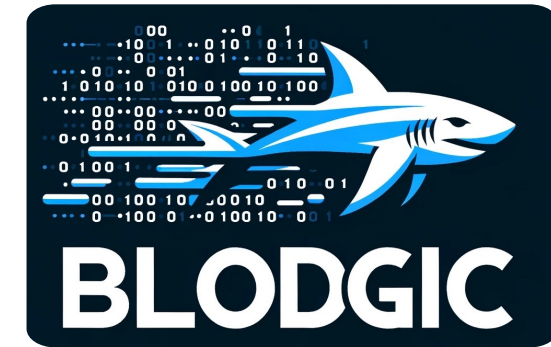


API and System Integration

App and system integration with API for other system integration and other embeddings in security products

LINKS TO LEARN MORE

- BLODGINC
 - <https://blodgin.com/>
- OpenSecurity + Data Community
 - <https://discord.gg/dd9FzF29>
- MITRE EMBED Project
 - <https://github.com/deepsecoss/MITREembed>
- IronBark AI cybersecurity defense for AI
 - <https://ironbarkai.com>



GRC Use Case 1






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

GRC Use Case 2







Brennan Lodge 12:32 PM

chat If our company uses the SANS Policies are we compliant to Basel III ?



CyberAdvisories APP 12:32 PM

 No, using the SANS policies alone does not ensure compliance with Basel III regulations. While the SANS policies provide guidance on information security controls, they do not cover all aspects required under Basel III. Basel III is a comprehensive global regulatory framework aimed at enhancing the resilience of banks and banking systems. It covers various areas such as capital adequacy, risk management, supervision, and disclosure requirements. To be compli  17 2023-10-18  Regulatory  BASELIII

Cyber Defender Use Case 1





Brennan Lodge 12:36 PM

chat what is Kerberoasting?



CyberAdvisories APP 12:37 PM

 Kerberoasting is a technique used by attackers to exploit vulnerabilities in Kerberos authentication systems. It involves stealing or forging Kerberos tickets to enable unauthorized access to sensitive resources. This method allows attackers to bypass authentication mechanisms and gain access to systems or data they should not have access to. In some cases, this technique can lead to complete compromise of an organization's security infrastructure. To protect against Kerberoasting attacks  17 16 August 2023



MITRE  Steal or Forge Kerberos Tickets

Cyber Defender Use Case 2



Brennan Lodge 4:41 PM

chat tell me about Fortinet CISA advisories



CyberAdvisories APP 4:42 PM

BOT: resposne: Based on the provided context, it seems like you want information related to Fortinet CISA advisories. The Cybersecurity and Infrastructure Security Agency (CISA) issues advisories to inform the public about cybersecurity threats and vulnerabilities.

In the given context, there is a specific Fortinet security update released by Fortinet to address a critical vulnerability (CVE-2023-33308) affecting FortiOS and FortiProxy. You are subscribed to Cybersecurity Advisories for Cybersecurity and Infrastructure Security Agency. This information has recently been updated, and is now available. Fortinet Releases Security Update for FortiOS and FortiProxy 07/11/2023 06:40 PM EDT Fortinet has released a security update to address a critical vulnerability (CVE-2023-33308) affecting FortiOS and FortiProxy. A remote attacker can exploit this vulnerability to take control of an affected system. CISA encourages users and administrators to review the FortiNet security release FG-IR-23-183 and apply the necessary updates. metadata={ 'Date': 'Tue, 11 Jul 2023 22:55:10 +0000', 'Source': 'CISA Advisory', 'Subject': 'Fortinet Releases Security Update for FortiOS and FortiProxy', 'filepath':