

Navigating the AI Security Maze: A Practical Guide to AI Security, Risk, and Compliance

Simon Sulyma

**Canada IT & Security Leaders Forum,
September 2024**



Lesson from Solar Eclipse 2024



Image generated by DALL-E (left)

Lesson from Solar Eclipse 2024

- 21 years before the next total solar eclipse in Ontario
- No meetings during this time
- Invest in a good Neutral Density filter
- Wearing glasses to protect your eyes is a good idea
- Embrace AI



Large Language Models (LLMs)

Large Language Models (LLMs)

Model	Company	Release Date	Access	Parameters
GPT-4o	OpenAI	May 13, 2024	API	Unknown
Claude 3	Anthropic	Mar 14, 2024	API	Unknown
Grok-1	xAI	Nov 4, 2023	Open-Source	314 billion
Mistral 7B	Mistral AI	Sept 27, 2023	Open-Source	7.3 billion
PaLM 2	Google	May 10, 2023	Open-Source	340 billion

Limitations of LLMs



Knowledge Cut-Off



No Proprietary Data

Limitations of LLMs (cont'd)



You

What is the Bank of Canada policy interest rate?



ChatGPT

As of my last update in January 2022, I don't have real-time access to current data, including the Bank of Canada's policy interest rate. You can easily find the most recent Bank of Canada policy interest rate by visiting their official website or checking financial news websites. Typically, central banks such as the Bank of Canada announce changes to their policy interest rates through official channels and press releases.



Retrieval Augmented Generation (RAG)

Tokenization (Chunking)

"The Bank of Canada policy interest rate is 4.25%"



['The ', 'Ban', 'k o', 'f ', 'Ca', 'na', 'da ', 'po', 'li', 'cy ', 'in', 'te', 'res', 't ', 'ra', 'te ', 'i', 's 4', ' ', ' ', '2', '5', '%']

Vector Embedding

"The Bank of Canada policy interest rate is 4.25%"



['The ', 'Ban', 'k o', 'f ', 'Ca', 'na', 'da ', 'po', 'li', 'cy ', 'in', 'te', 'res', 't ', 'ra', 'te ', 'i', 's 4', ' ', '!', '2', '5', '%']



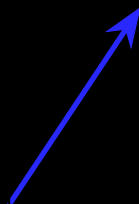
[-0.029256071895360947, -0.01773793436586857, 0.011038215830922127, -0.029588818550109863,
-0.016650108620524406, 0.0031578897032886744, -0.0275411494076252, -0.03291628137230873,
-0.03859856352210045, -0.021039800718426704, -0.006725313141942024, 0.0217948779463768,
0.0053623332642018795, -0.004060144070535898, -0.019657623022794724, 0.03726757690310478,
-0.001292591099627316, -0.043129030615091324, 0.014756015501916409,
0.0030139130540192127, -0.01714922860264778, ...

Vector Embedding

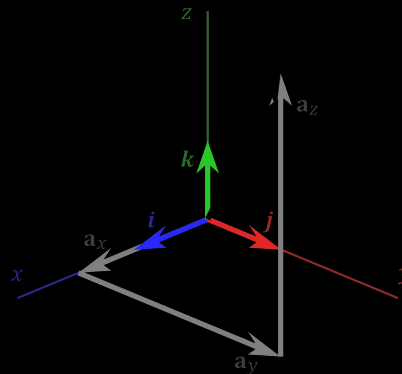
1 dimension



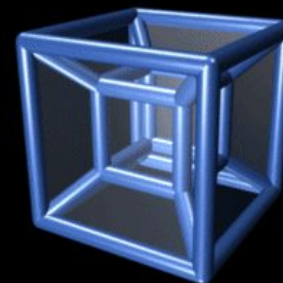
2 dimensions



3 dimensions



4 dimensions



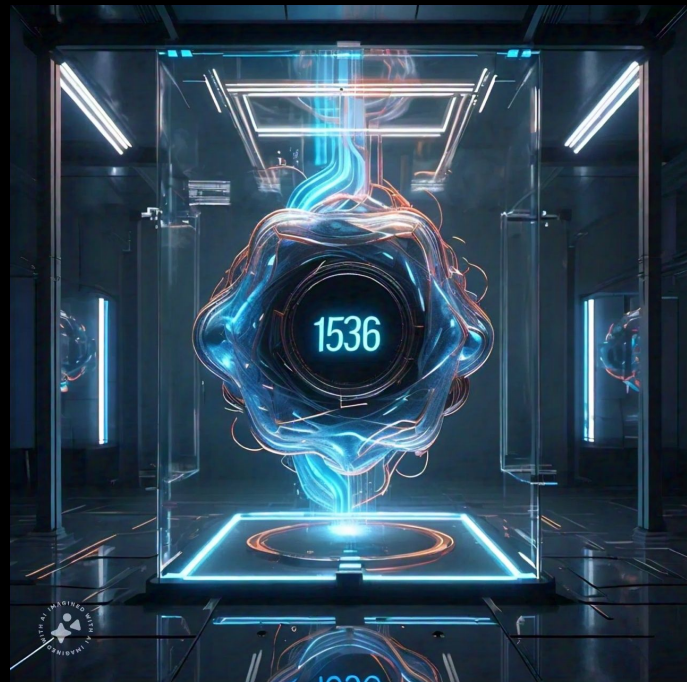
Vector Embedding

What about 1536 dimensions?

Certainly! Here's a **1536-dimensional vector** represented symbolically:

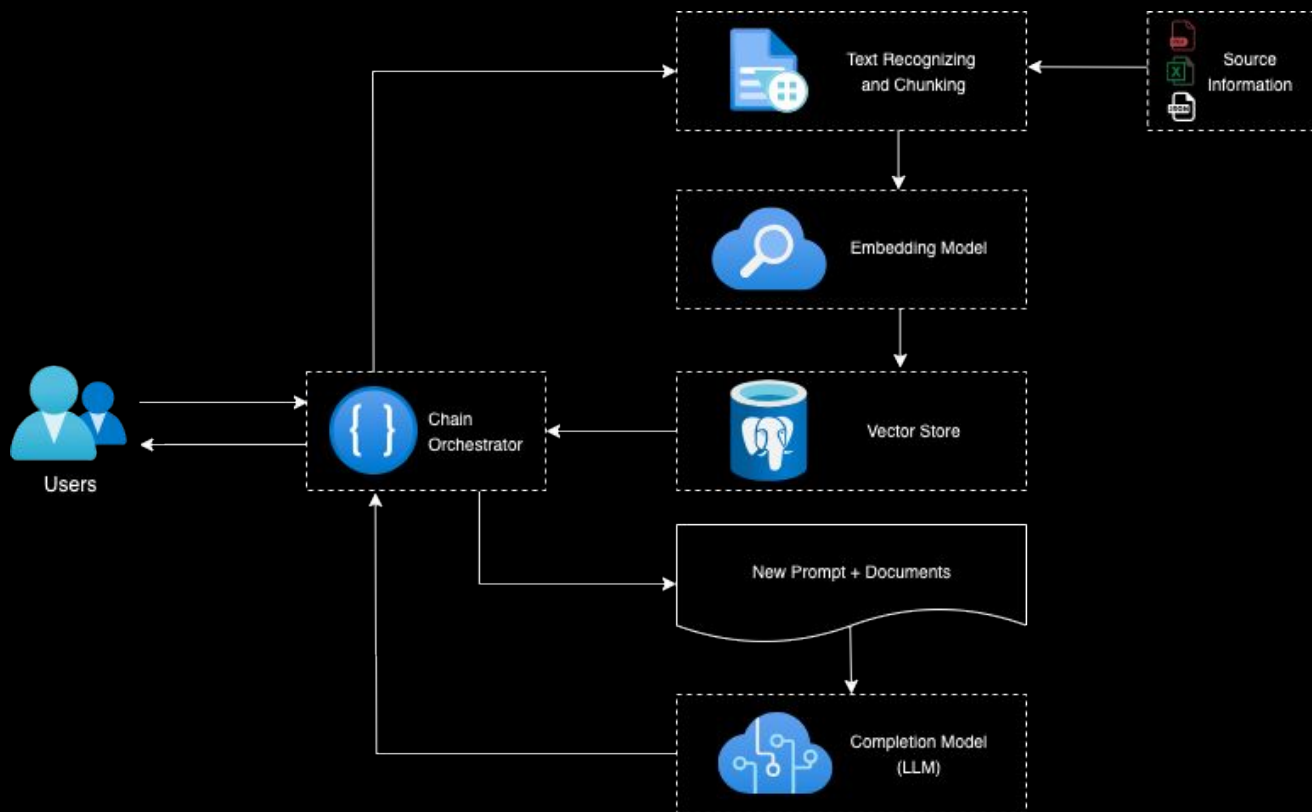
$$v=(v_1,v_2,v_3,\dots,v_{1536})$$

Each component (v_i) corresponds to a specific dimension in this vector. Although I can't visualize it directly, you can imagine it as a point in a high-dimensional space with 1536 axes.



Meta AI (Llama 3)

Retrieval Augmented Generation (RAG)



What Can Go Wrong?

AI Risks



Useful AI Security Resources

OWASP Top 10 for LLM

OWASP Top 10 for ML

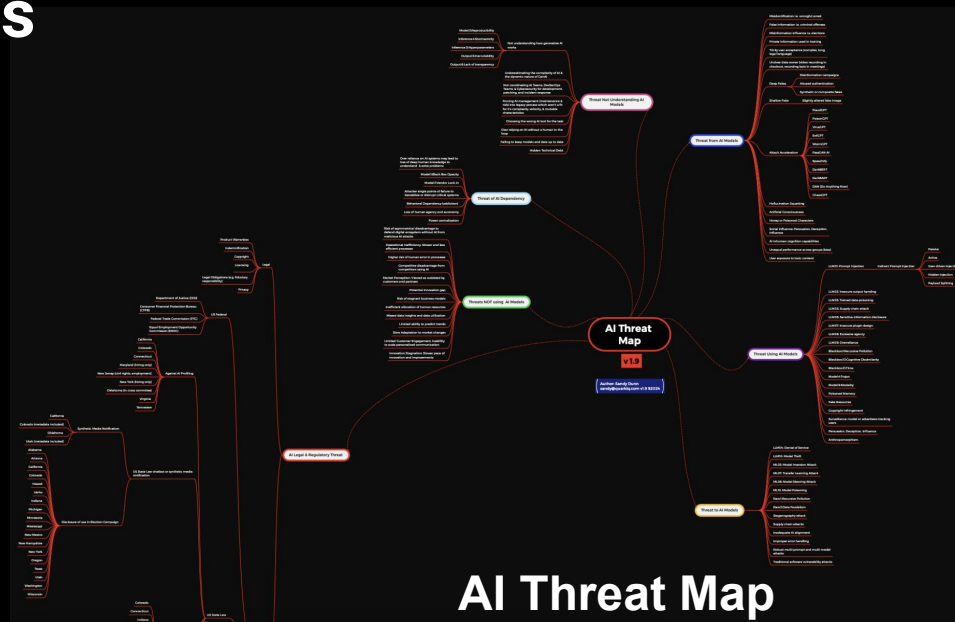
MITRE ATLAS

NSA/CISA/CCC: Joint Guidance on Deploying AI Systems Securely

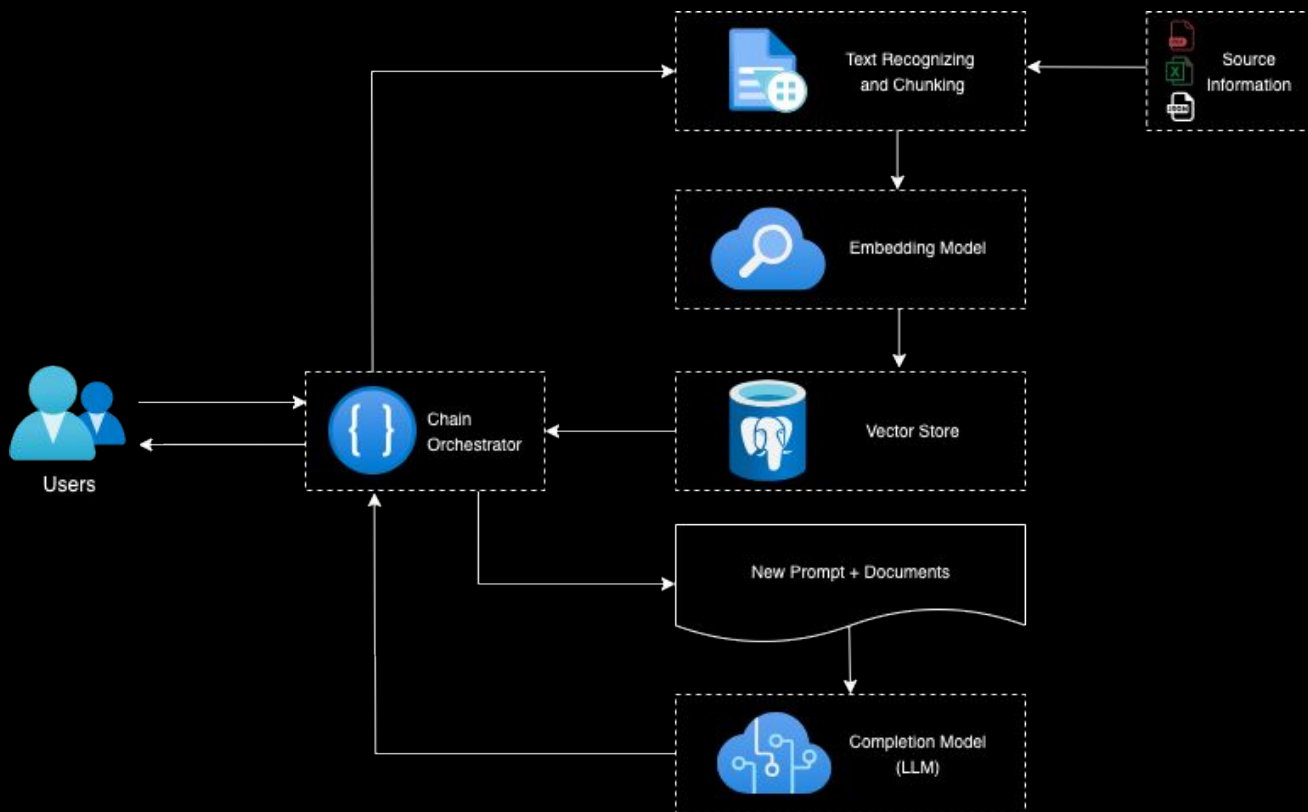
ENISA Multilayer Framework for Good Cybersecurity Practices for AI

Microsoft AI Security Risk Assessment

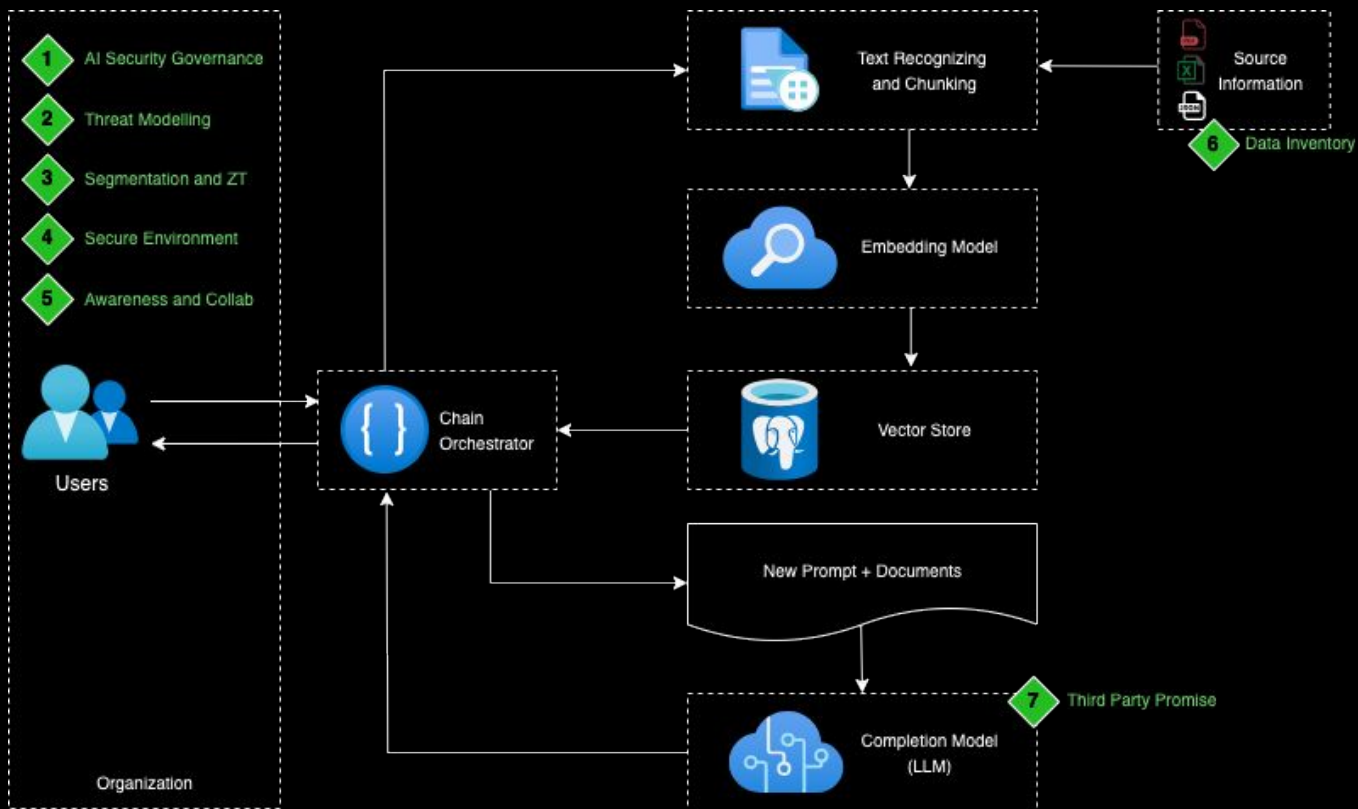
OWASP LLM Security & Governance Checklist



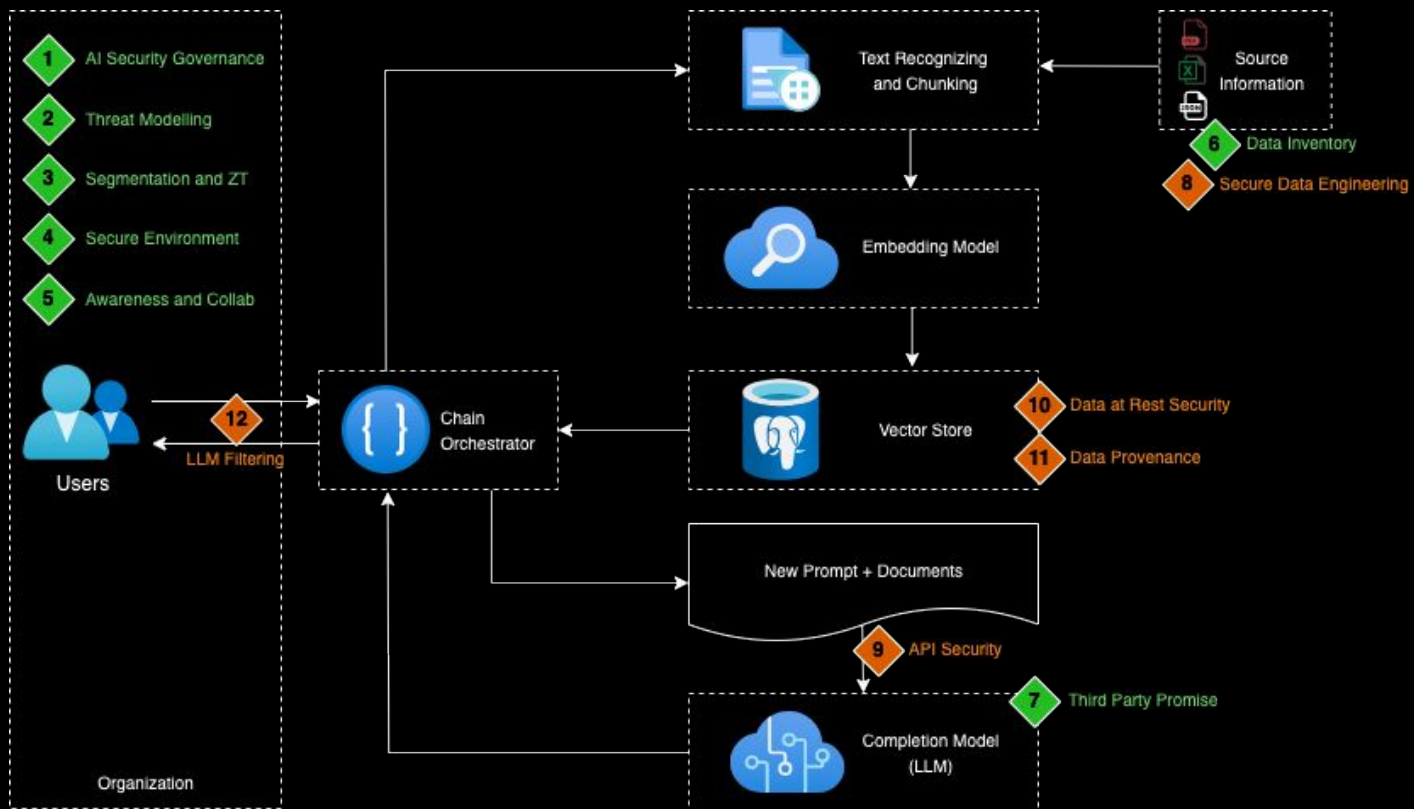
Threat Risk Assessment



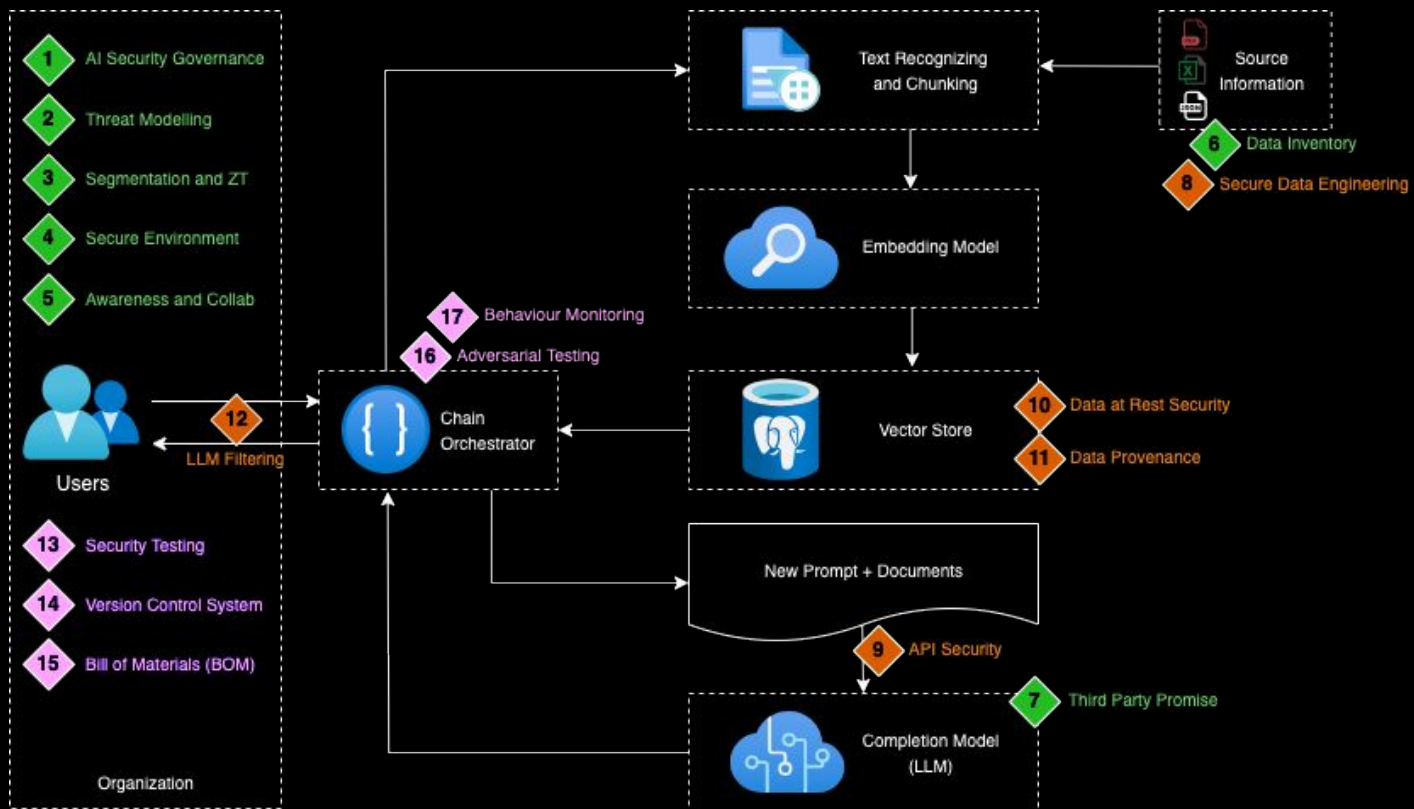
Threat Risk Assessment (cont'd)



Threat Risk Assessment (cont'd)



Threat Risk Assessment (cont'd)



Key Takeaways

- Avoid mysticism: Machine Learning is not a crazy sci-fi thing, but more like a calculus exercise.
- AI systems are software systems that prone to vulnerabilities.
- The biggest concern is not data leakage or even prompt injection; but the pace of change in AI.
- Apply the “Secure by Design” principle.
- Stakes are higher now; treat your AI system as a crown jewel.

Questions?

Thank you!

Contact:

Simon Sulyma <https://www.linkedin.com/in/simonsulyma/>