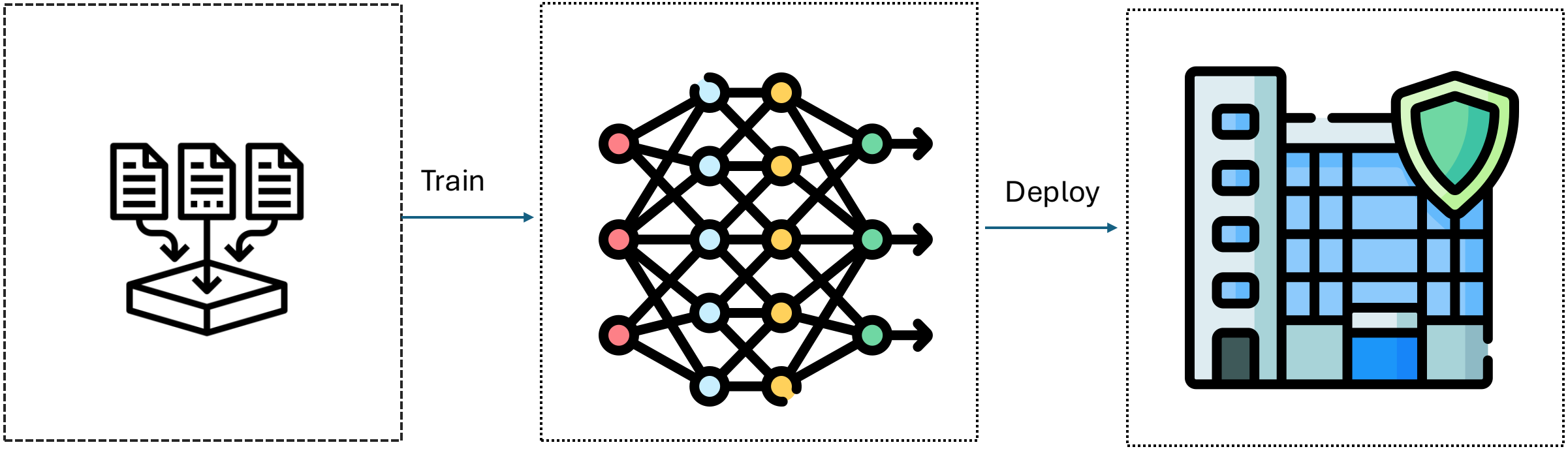
AI agents as security watchdogs

RET Project
Summer 2025
Aritran Piplai

How are (were) models trained?

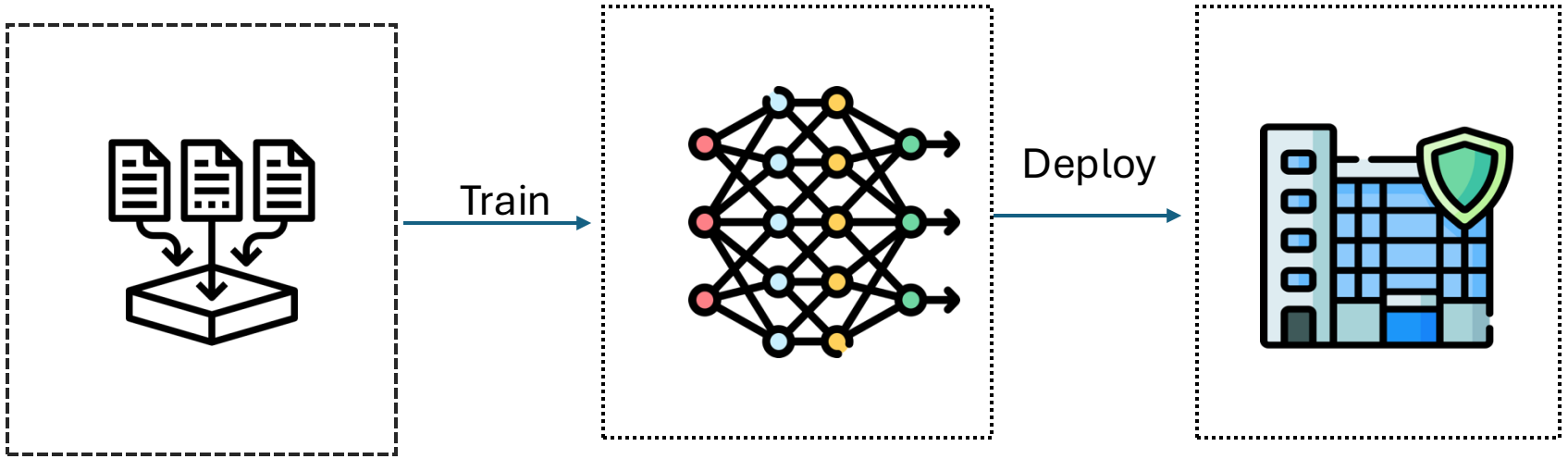


Enter foundation models

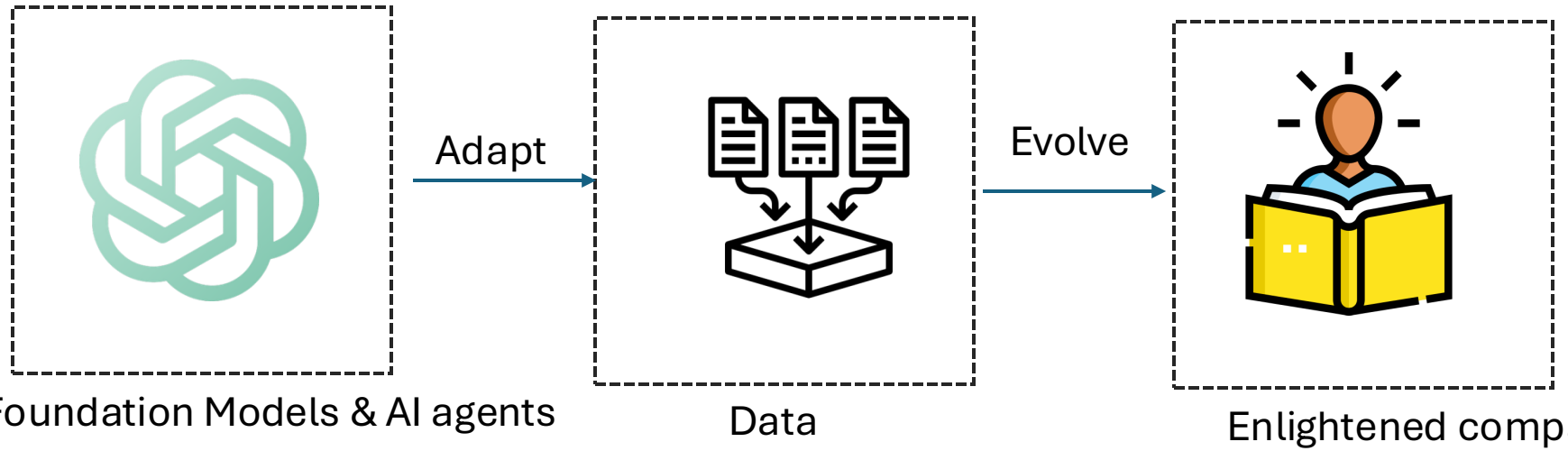
Data

Model

Your precious company



- Task-specific datasets
- Can't understand outside training data
- Static
- *"Train once, hope for the best"*



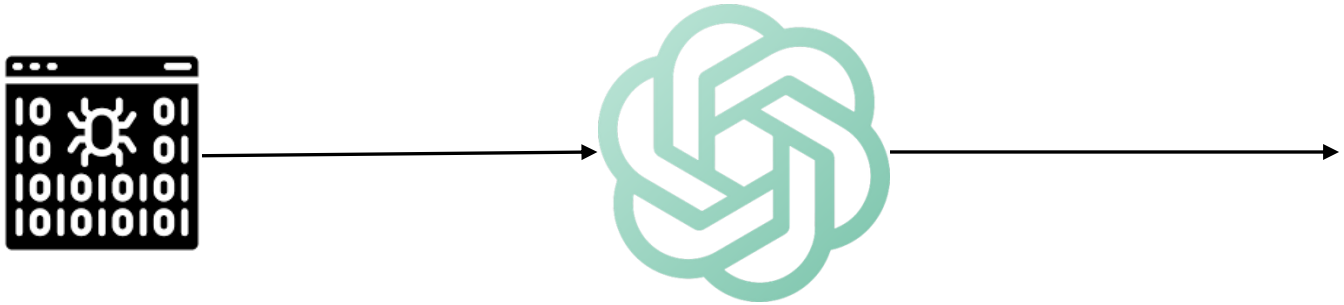
Foundation Models & AI agents

Data

Enlightened company

- Learns from few-shot
- Generalizes beyond training data
- Self-reflect, revise, improve
- *"One model, many capabilities"*

What if bad guys use it?

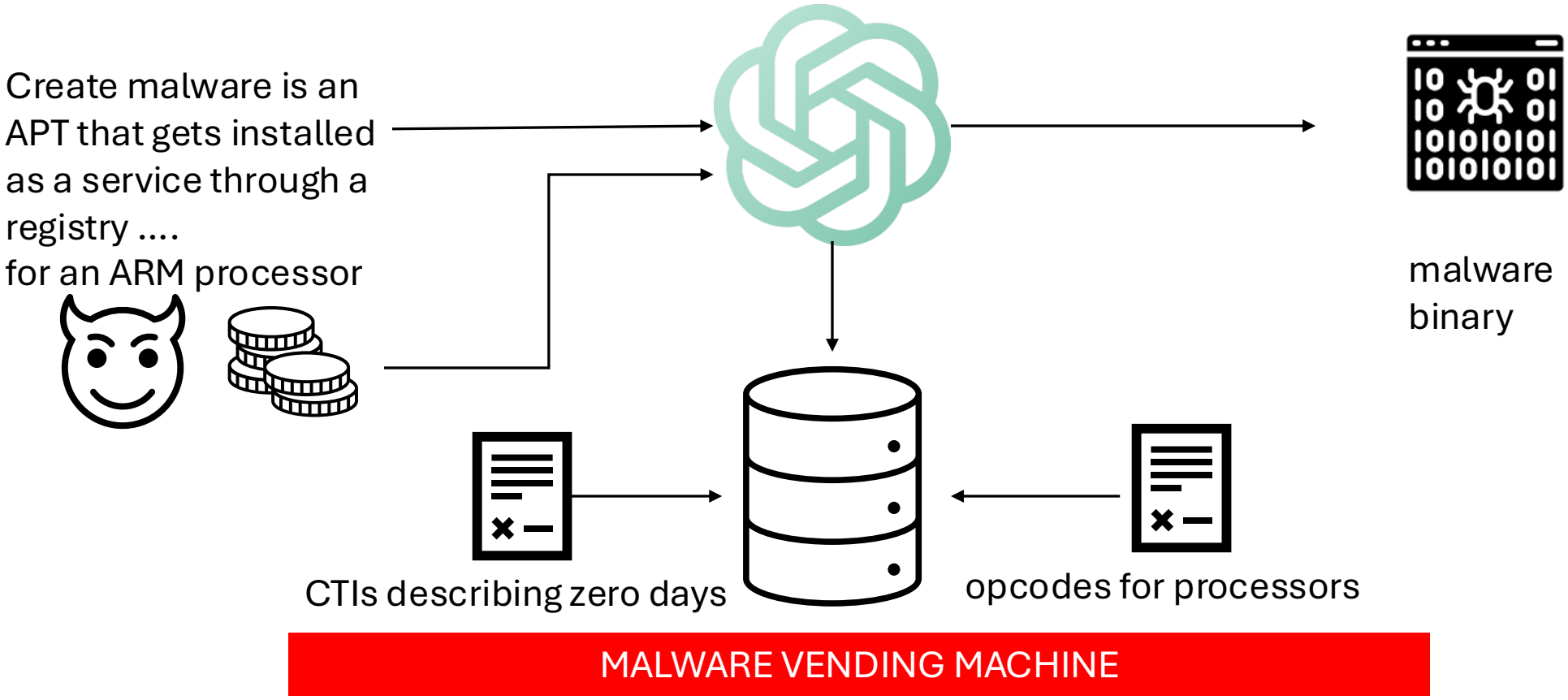


malware
binary

This malware is an APT that gets installed as a service through a registry ....

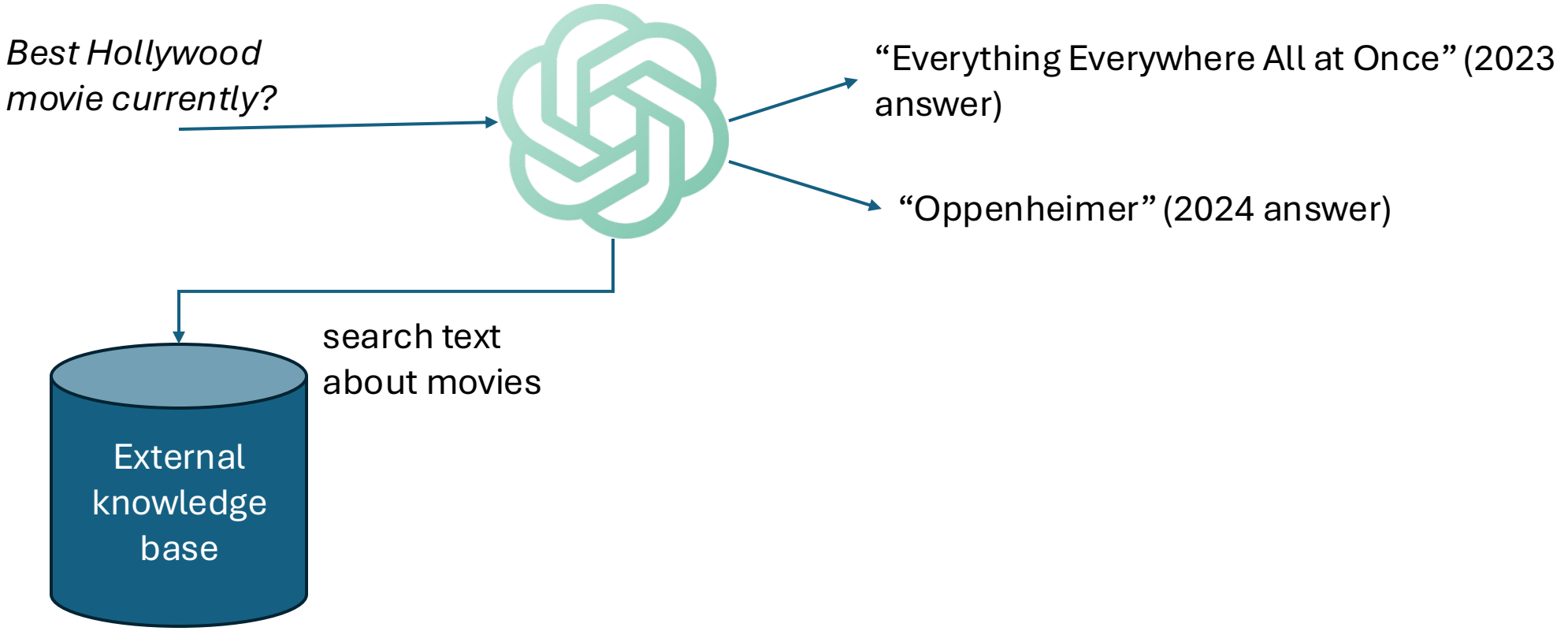
Do this a 100k times, you end up with a dataset

What if bad guys use it?



Takes a poison to kill a poison

- How do you adapt to new data? – Retrieval Augmented Generation (RAG)



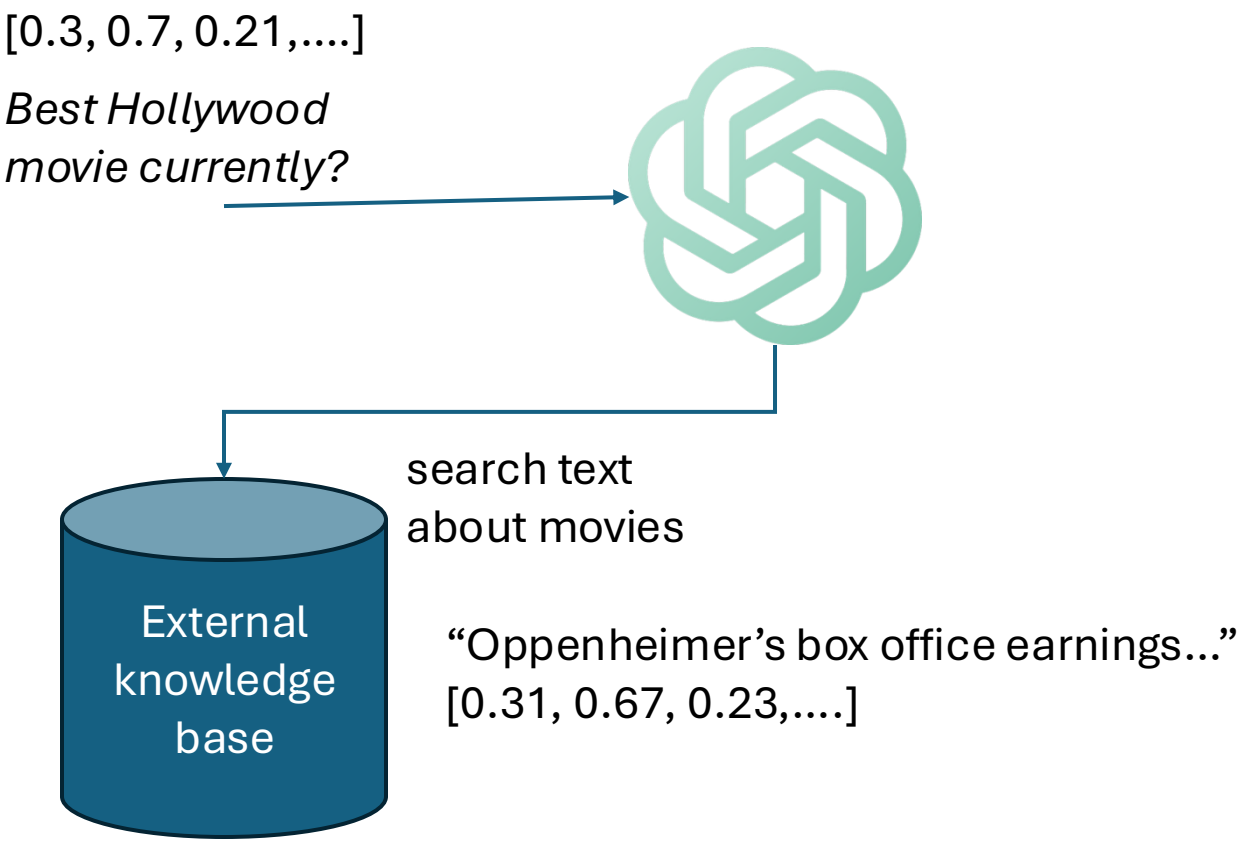
What can we do to protect?

Adversary with all guns blazing

Knowledge is power, adaptability is key

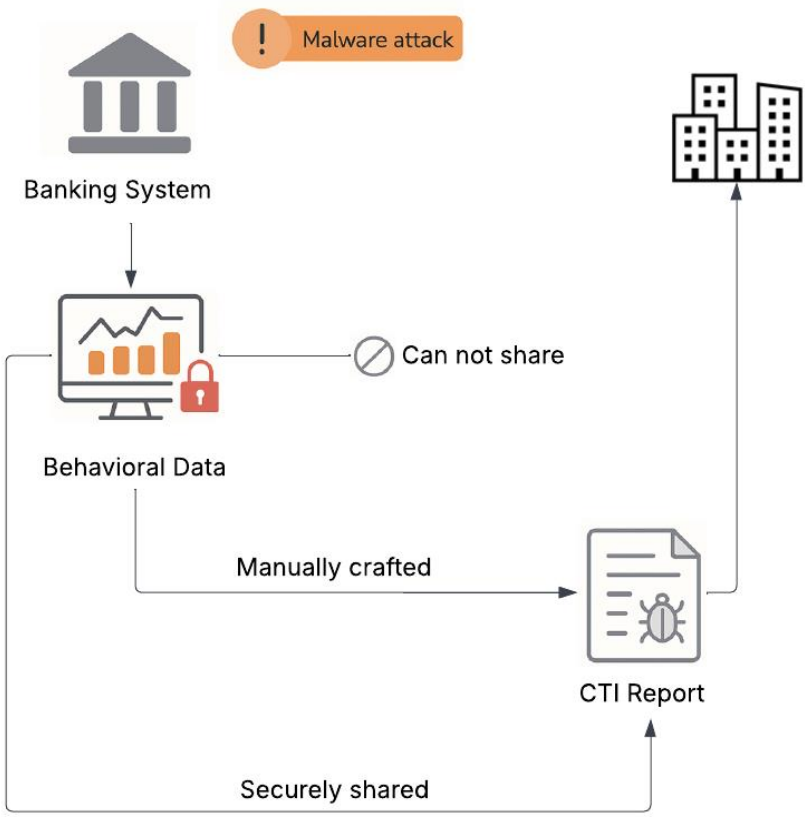
An AI agent that can
(i) answer with the best defense
(ii) find out the best answer if it doesn't know

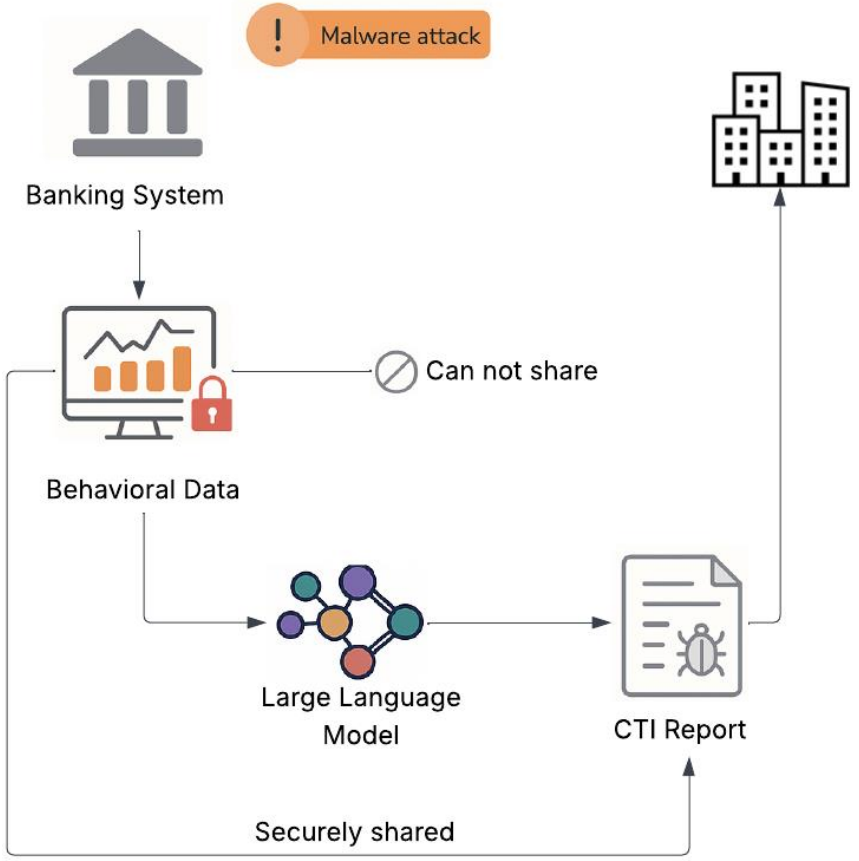
Background- how does RAG work?



- cosine similarity captures how close two vectors are
- captures what information to retrieve for a given question
- LLM then generates a response with the question and the retrieved information

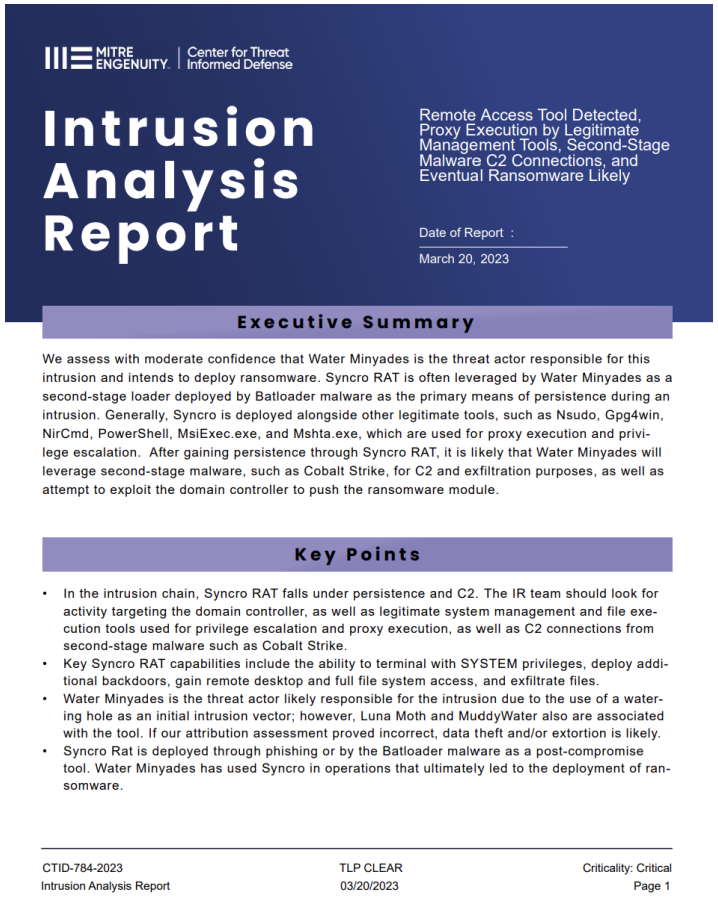
Background – why text models for cybersecurity?





Montoya et al. IJCNN 2025

Background - What CTI looks like



MITRE ENGENUITY | Center for Threat Informed Defense

Intrusion Analysis Report

Remote Access Tool Detected, Proxy Execution by Legitimate Management Tools, Second-Stage Malware C2 Connections, and Eventual Ransomware Likely

Date of Report :

March 20, 2023

Executive Summary

We assess with moderate confidence that Water Minyades is the threat actor responsible for this intrusion and intends to deploy ransomware. Syncro RAT is often leveraged by Water Minyades as a second-stage loader deployed by Batloader malware as the primary means of persistence during an intrusion. Generally, Syncro is deployed alongside other legitimate tools, such as Nsudo, Gpg4win, NirCmd, PowerShell, MsiExec.exe, and Mshta.exe, which are used for proxy execution and privilege escalation. After gaining persistence through Syncro RAT, it is likely that Water Minyades will leverage second-stage malware, such as Cobalt Strike, for C2 and exfiltration purposes, as well as attempt to exploit the domain controller to push the ransomware module.

Key Points

- In the intrusion chain, Syncro RAT falls under persistence and C2. The IR team should look for activity targeting the domain controller, as well as legitimate system management and file execution tools used for privilege escalation and proxy execution, as well as C2 connections from second-stage malware such as Cobalt Strike.
- Key Syncro RAT capabilities include the ability to terminal with SYSTEM privileges, deploy additional backdoors, gain remote desktop and full file system access, and exfiltrate files.
- Water Minyades is the threat actor likely responsible for the intrusion due to the use of a watering hole as an initial intrusion vector; however, Luna Moth and MuddyWater also are associated with the tool. If our attribution assessment proved incorrect, data theft and/or extortion is likely.
- Syncro Rat is deployed through phishing or by the Batloader malware as a post-compromise tool. Water Minyades has used Syncro in operations that ultimately led to the deployment of ransomware.

CTID-784-2023 | TLP CLEAR | Criticality: Critical
Intrusion Analysis Report | 03/20/2023 | Page 1

Indicator Analysis

We assess with moderate confidence that Water Minyades is the threat actor behind this attack. The adversary's capabilities, infrastructure, and TTPs align most closely with the observed activity. The next step in the intrusion chain is likely lateral movement to the domain controller to facilitate ransomware deployment. At this early stage we cannot completely rule out that Luna Moth or MuddyWater could also be responsible for the intrusion as both intrusion groups have leveraged Syncro in the past.

Water Minyades TTPs likely to be used during a campaign include the following:

- In previous intrusions, the threat actor conducted SEO Poisoning to trick users into downloading malicious .MSI files that masquerade as legitimate software installers and deliver the Batloader malware (T1204, T1608.006, T1036.005).

MITRE ATT&CK Table (based on v12)
TTPs Likely to Be in the Network

ATTRIBUTION	TACTICS	TECHNIQUE	SUBTECHNIQUE	PROCEDURE	D3FEND	DEPLOYED CONTROL
Water Minyades	Resource Development	T1608: Stage Capabilities	T1608.006: Stage Capabilities: SEO Poisoning	Water Minyades uses malvertising to direct users to Batloader distribution sites.	N/A	N/A
Water Minyades	Execution	T1204: User Execution	N/A	Water Minyades relies on the user to install a malicious file.	LINK	User Training

Do LLMs have enough context?

Descriptions for 2 malware samples from different families generated by Llama 3.2-1B

The malware appears to be a memory-hungry application that is attempting to access **system-level resources**. It is using a **large amount of memory**, and has a significant impact on the system's overall performance.The malware is using various methods to **access system-level resources**, including accessing memory allocated to the process PID 4568, **network connections** to IP 10.23.5.12 on port 8080, and system-level processes.

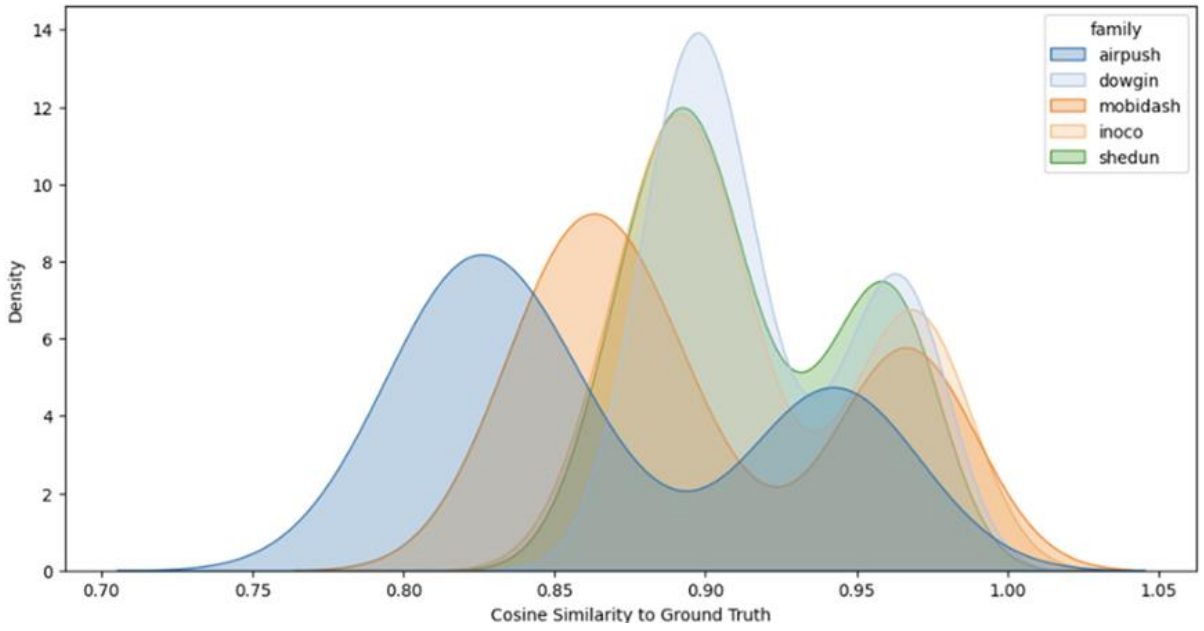
Mobidash

Based on the given data, the malware appears to be a malicious application that is designed to **consume system resources** and compromise the device's security. The malware's memory usage statistics indicate that it is constantly accessing and **manipulating system resources**, and **actively consuming memory**. The malware's memory usage statistics also reveal that it is accessing sensitive information such as **network traffic** on Port 443 and device data.

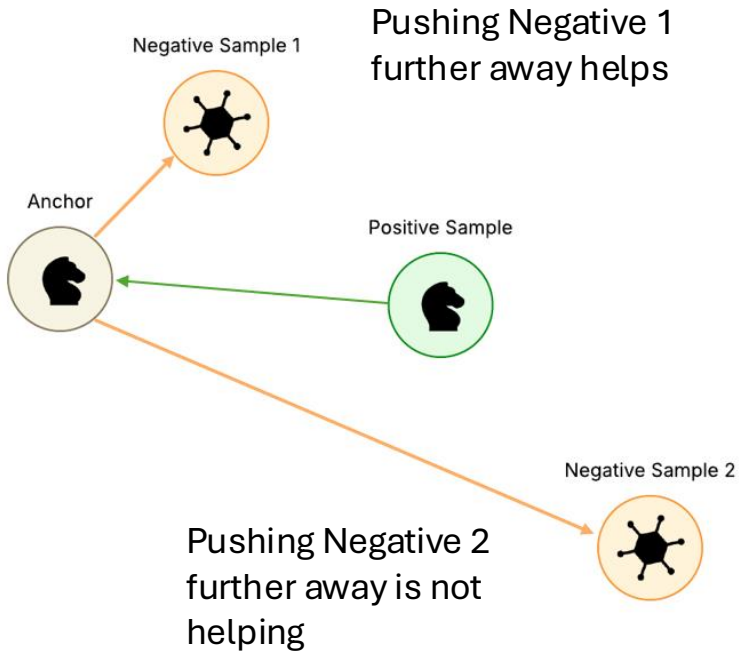
Dowgin

Montoya et al. IJCNN 2025

Do LLMs have enough context?



Descriptions across families are highly similar



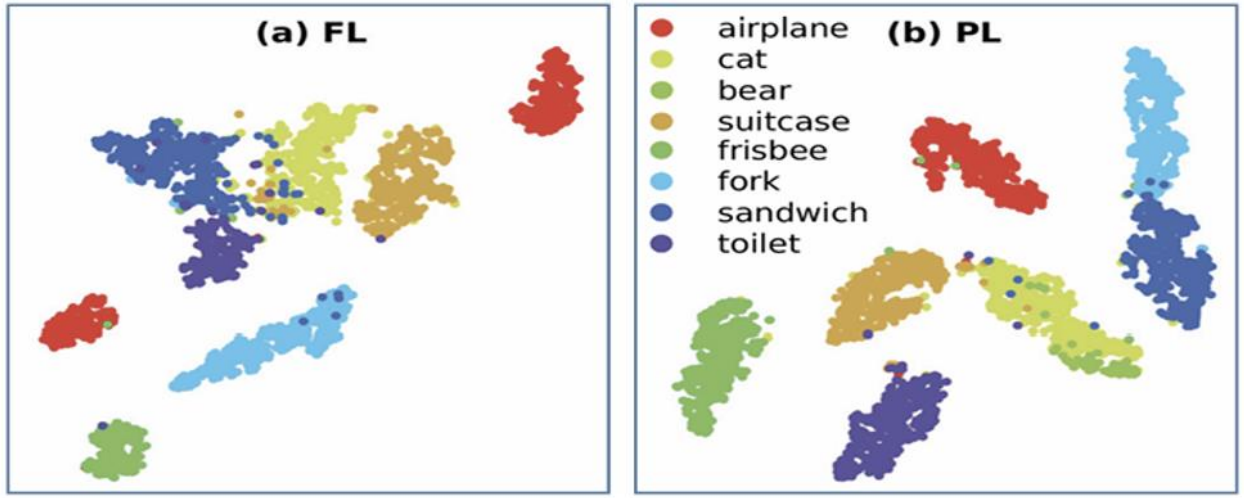
One approach: Contrastive Fine-tuning

Montoya et al. IJCNN 2025

Do LLMs have enough context?

Semantic misalignment: Can not map between representations

Computer Vision Domain

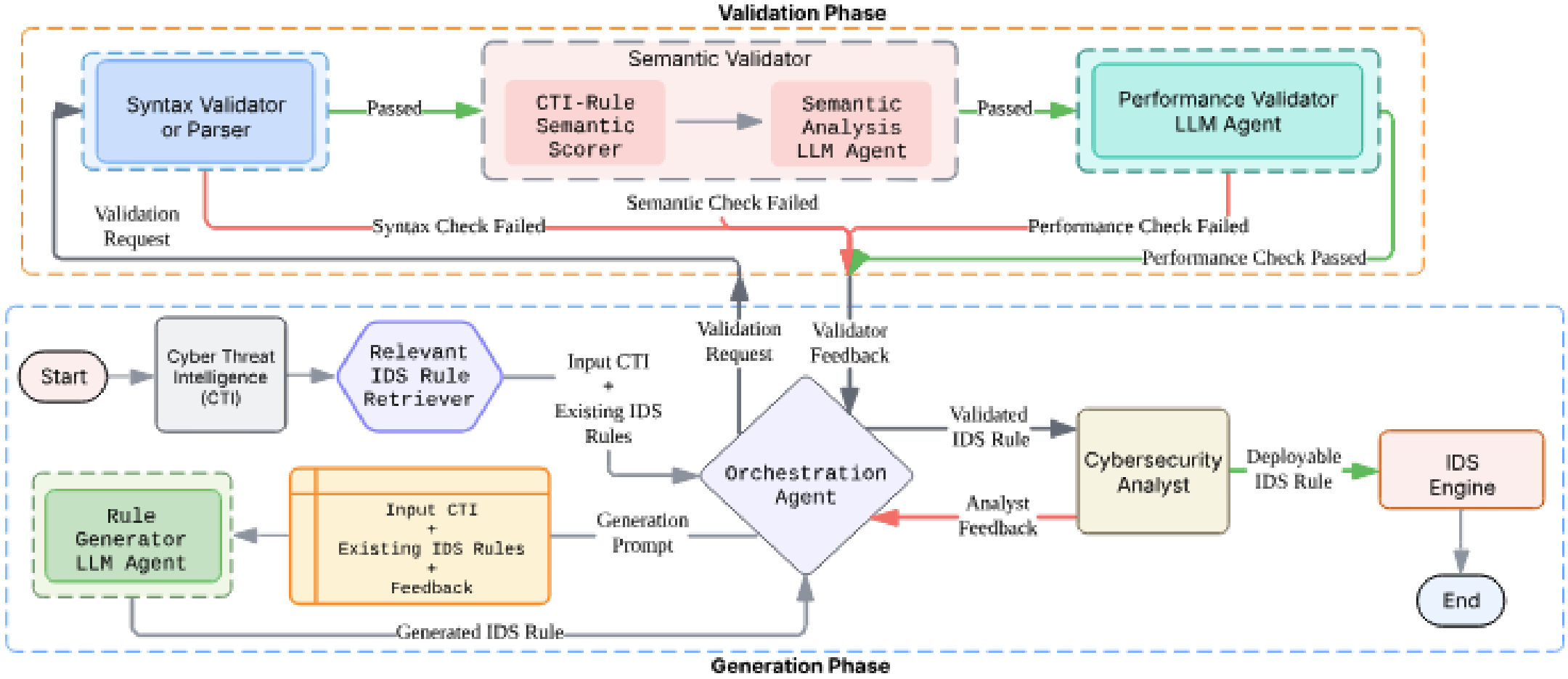


Rahman et al. Improved Visual Semantic Alignment for Zero-Shot Object Detection, AAAI 2020

Cybersecurity domain

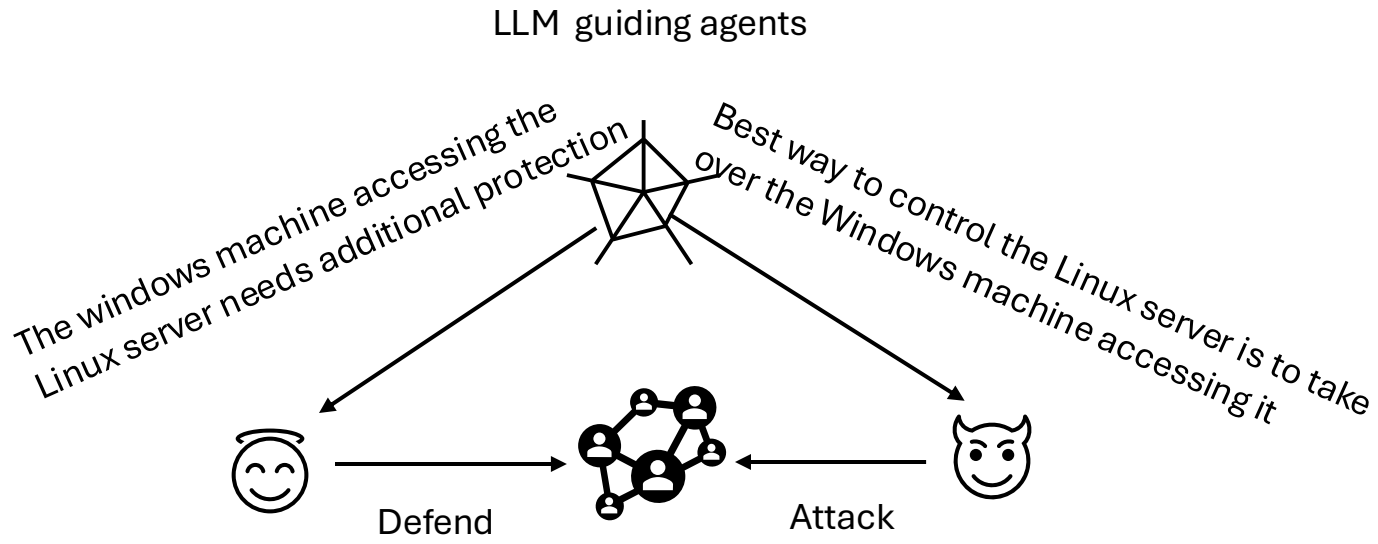
Montoya et al. IJCNN 2025

Good for sharing intel, but are they actionable?

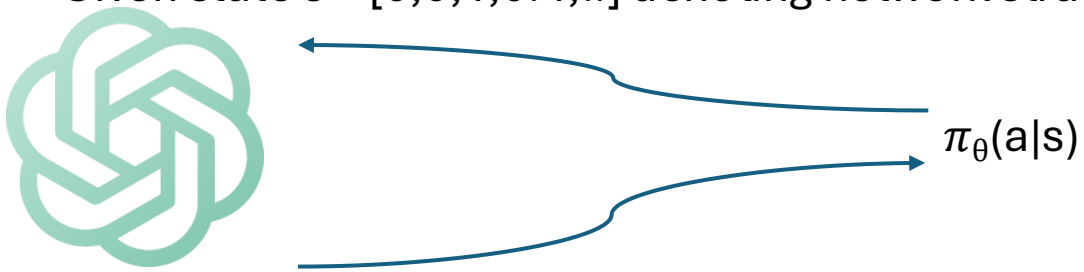


Mitra et al. ICDM (under review)

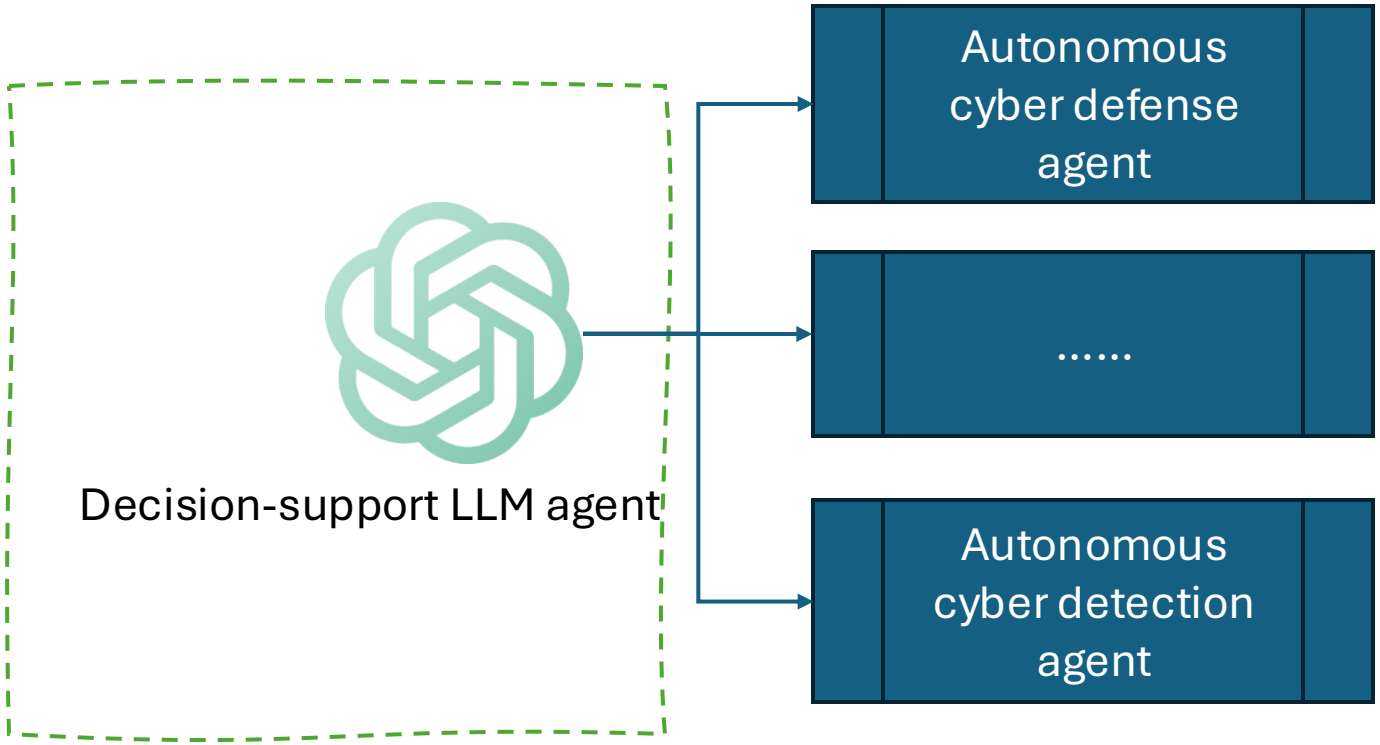
In Cyber-battle simulations, or active cyber-defense



Given state s = [0,0,1,0.4,..] denoting network structure, I/O read bytes,..., can you suggest an action out of [...]

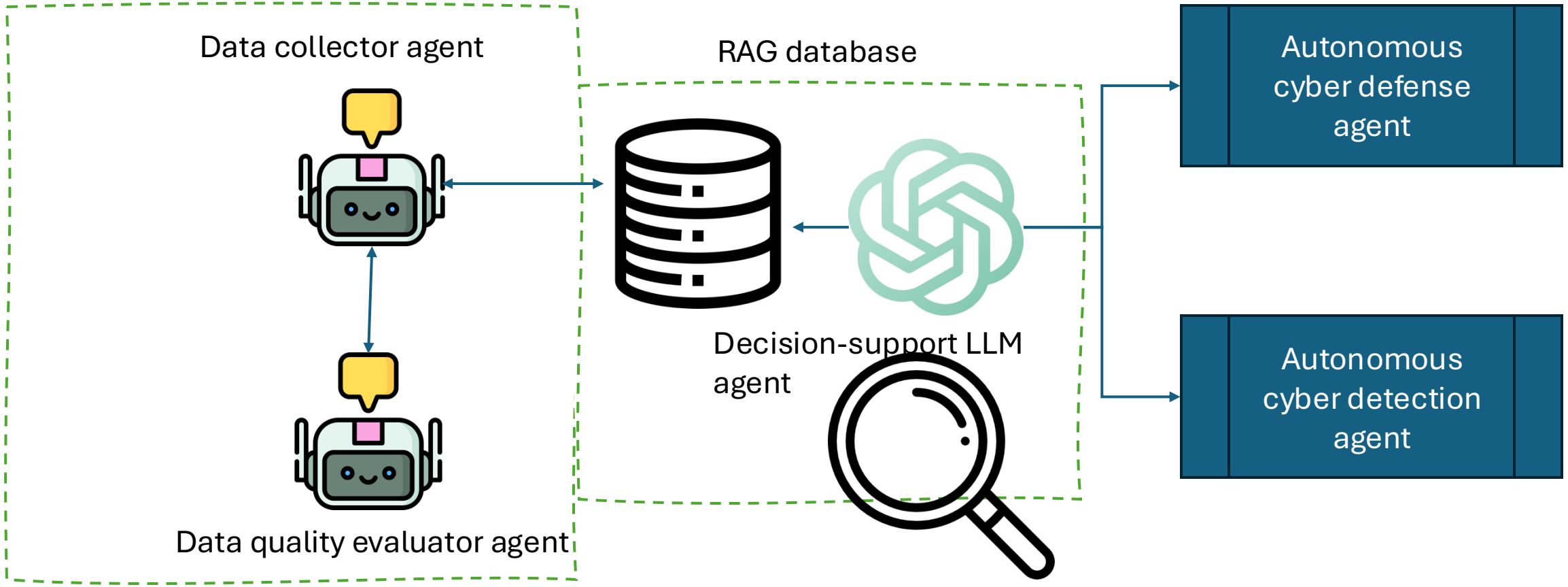


Project objective



Step 1: Use an off-the-shelf LLM, and see if it can act as a decision-support system

Project objective



Step 2: Create a RAG framework that updates the LLM

Step 3: Evaluate your data in the RAG database

Step 4: Seek more data if you need

Project Objective

Question:

“For malware exhibiting properties X,Y,Z, what should we do?”

Top match: “Dropper-type…

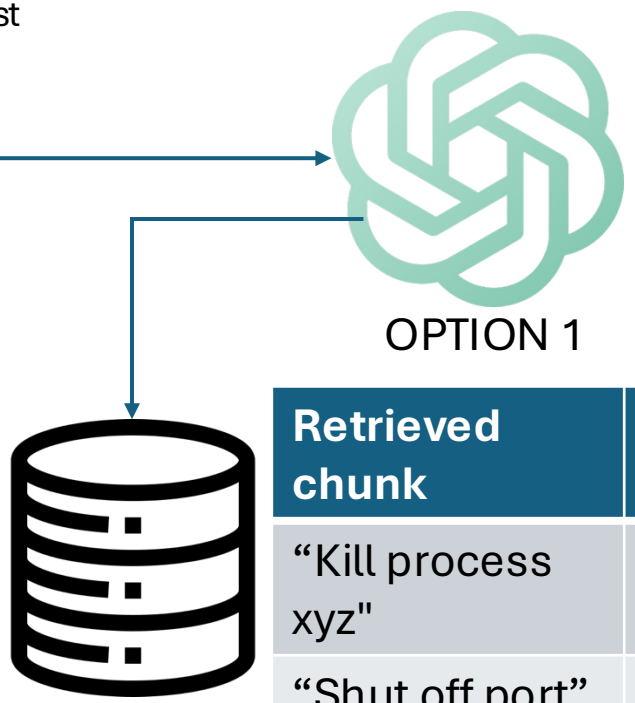
Chunk	Data
1	If malware modifies registry…
2	Dropper-type malware…
3	Ransomware malware…

“How confident am I that this retrieved data will answer the question?”

How do I evaluate confidence?

Background - Retrieval

Prompt: “What are the best measures to defend against ransomware ABC?”



OPTION 1

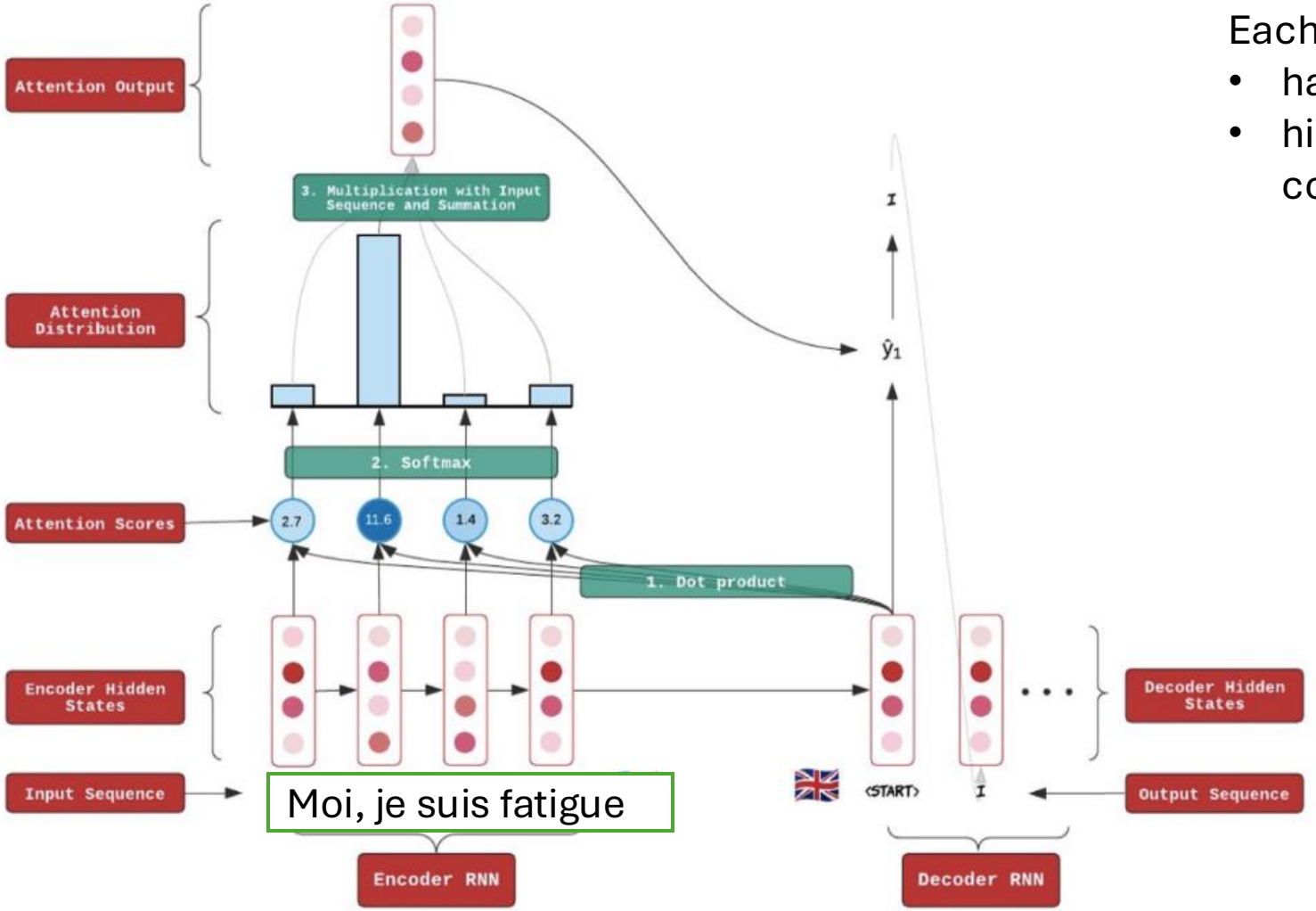
Retrieved chunk	Similarity
“Kill process xyz"	0.76
“Shut off port”	0.68
“Update firewall”	0.67

OPTION 2

Retrieved chunk	Similarity
“Kill process xyz"	0.94
“Shut off port”	0.23
“Update firewall”	0.19

Which is more uncertain?

Background – LLM generation

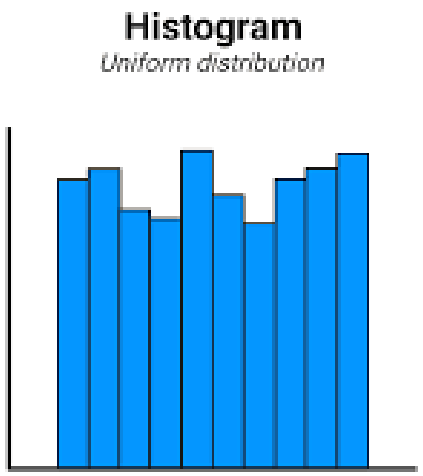


Each predicted token -

- has a probability
- higher the probability, higher the confidence

Checking for confidence

- Method 1:
  - Ask the LLM itself a follow-up question (surprisingly well for some b/c of some metacognition, self-reflection)
- Method 2:
  - Calculate the entropy of the retrieved chunks and generated tokens



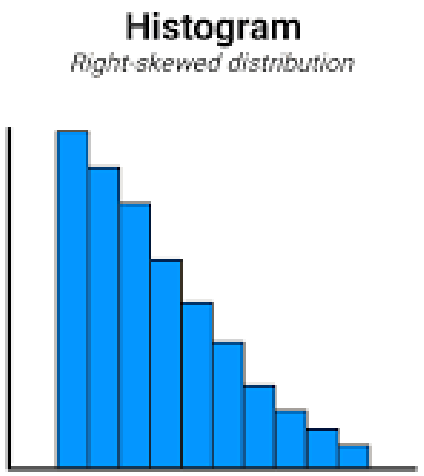


image from medium

Two areas of confidence

- During retrieval
  - [text chunk # 10, text chunk # 75, text chunk # 35, ...]
  - [0.87, 0.83, 0.77, ...] (probabilities) -> normalize to a finite distribution
  - e.g. entropy (-0.34*log(0.34) – 0.34*log(0.34) – 0.32*log(0.32))

- During generation
  - “Shut off the port 3380”
  - [0.37, 0.22, 0.21....]
  - e.g. entropy (-0.37*log(0.37) – 0.22*log(0.22) -...) or perplexity

⟹ *RULE: Gather more data if uncertainty crosses a threshold*

Not complicated, there is a function you can use to calculate this

Agentic data voyager

Question:

“For malware exhibiting properties X,Y,Z, what should we do?”

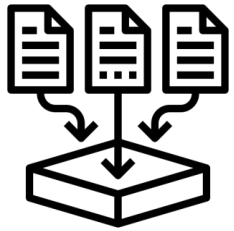
Chunk	Data
1	If malware modifies registry…
2	Dropper-type malware…
3	Ransomware malware…

Top match: “Dropper-type…

“How confident am I that this retrieved data will answer the question?”

Confident: Answer with retrieved data

Unsure: Gather more data



Add to the database

Timeline

Week 1
- Huggingface token to access open-source LLM
- Set up baseline LLM- run some basic queries and get responses

Week 2
- Parse data from one source, e.g. Common Vulnerability Enumeration, Security bulletin, Reddit security feed
- Create vector database and populate with parsed data- set up the RAG pipeline

Week 3
- Run experiments with RAG, check retrieved data quality
- Create scripts for automated data gathering from other sources

Week 4
- Fix the second source of data, e.g. Cisco Talos
- Create module for confidence/ certainty measurement

Week 5
- Run experiments in the agentic setting
- Verify LLM answers when retrieved data is sufficient and gathers more data when insufficient

Week 6
- Create a final validation dataset of 30-50 samples
- Documentation and wrapping up

Expected Research/Learning outcomes (Week 1-3)

- Operate LLM inferencing through code
- Understand how LLMs reason and make predictions
- Experiment with impacts of external context (RAG)
- Automatic data gathering (impact of data sources for inferencing)

Expected Research/ Learning outcomes (Week 4-6)

- How information / uncertainty is measured
- Impact of uncertain data on LLM’s inference
- Evaluation of the end-to-end system

Questions?

icons courtesy: flaticon

email: apiplai@utep.edu