

Enhancing DevOps with MLOps and MLSecOps- Guardrails around AI powered Applications

Jatin Sachdeva
Principal Security Architect

cisco Live !

Cisco Webex App

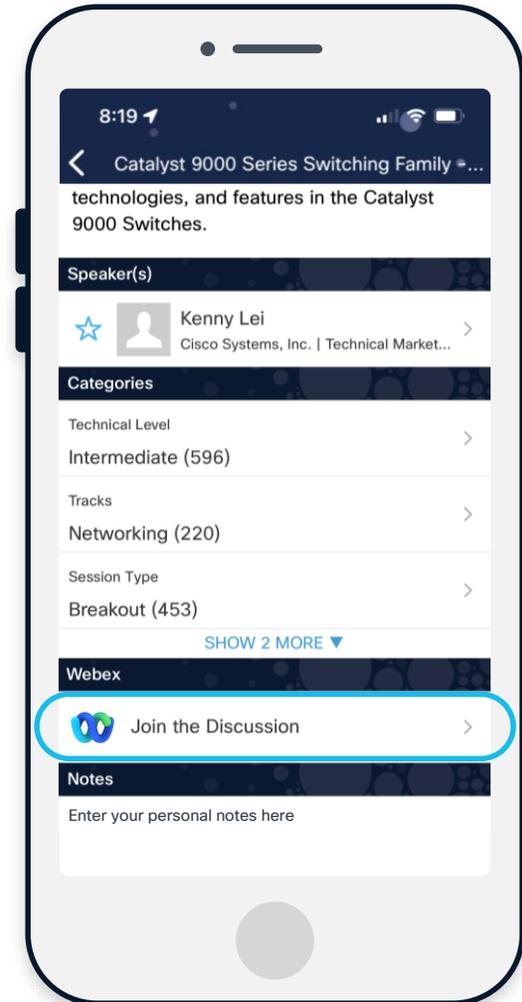
Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 13, 2025.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKCLD-1006>

Agenda

- 01 **Intro - What is MLOps?**
- 02 **A sample LLM application**
- 03 **Where are the Threats!**
- 04 **Securing AI with Cisco**
- 05 **Outro**

About your speaker

Fun fact – I am Indian, but I am not into cricket or spicy food!

Play

Live in Melbourne, Australia with my lovely family and this crazy fellow

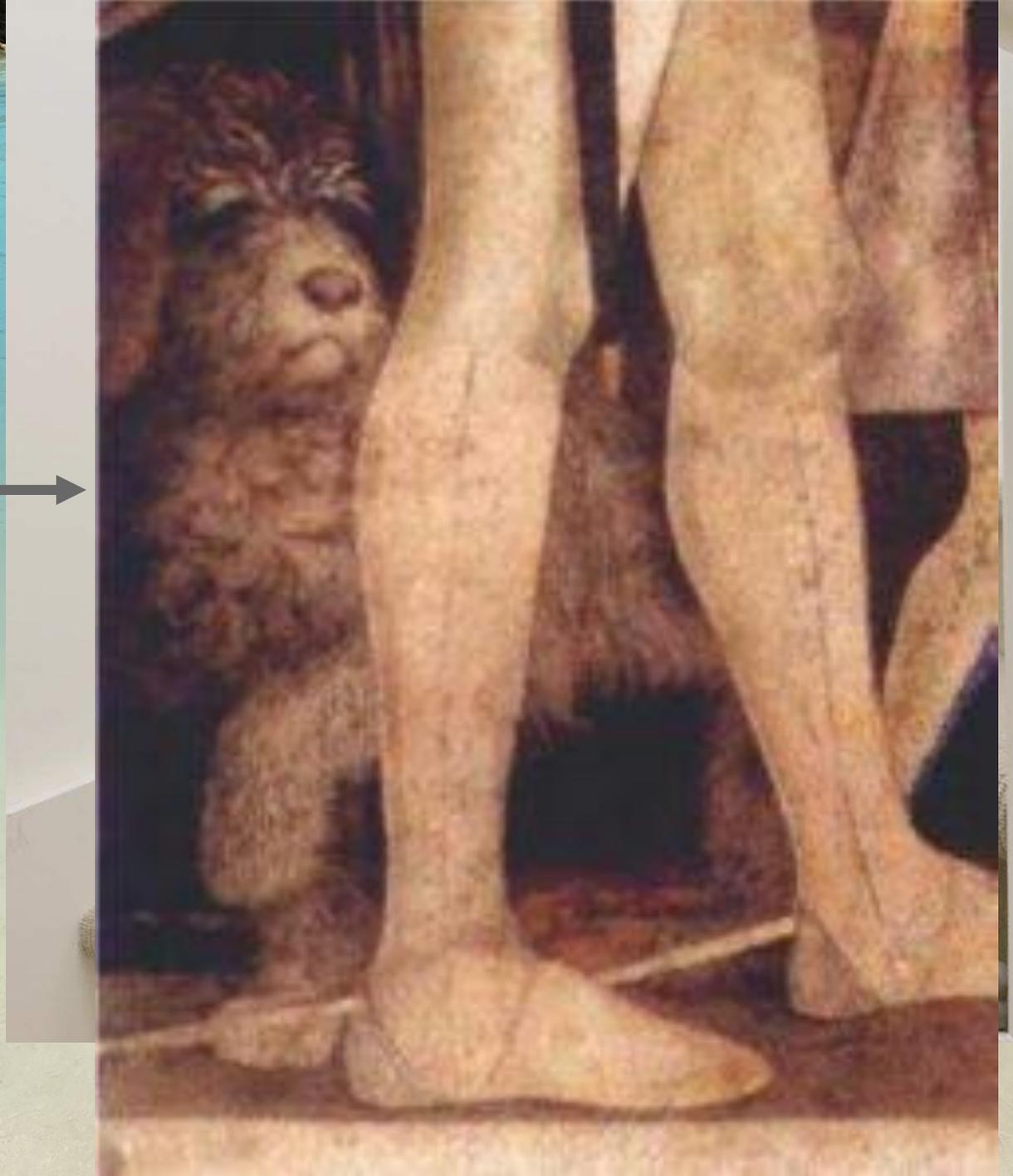
Can nerd out on anything from tech and cars to fitness and nutrition!

Work

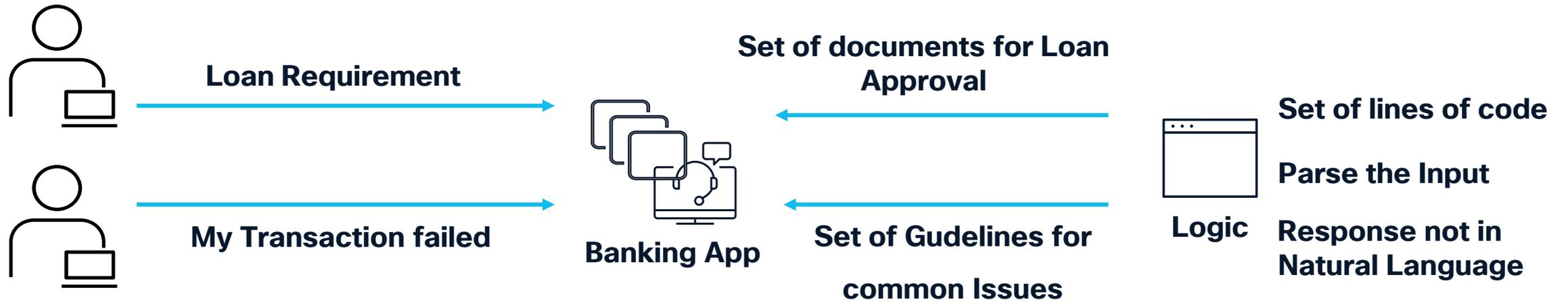
23 years in security industry, 20 in Cisco.

Knowledge seeker hence certs a plenty – CISSP, CISA, CEH, GWAPT, GSEC, GCIA, GCIH, GCSA, GPCS, SFCE, Associate C|CISO

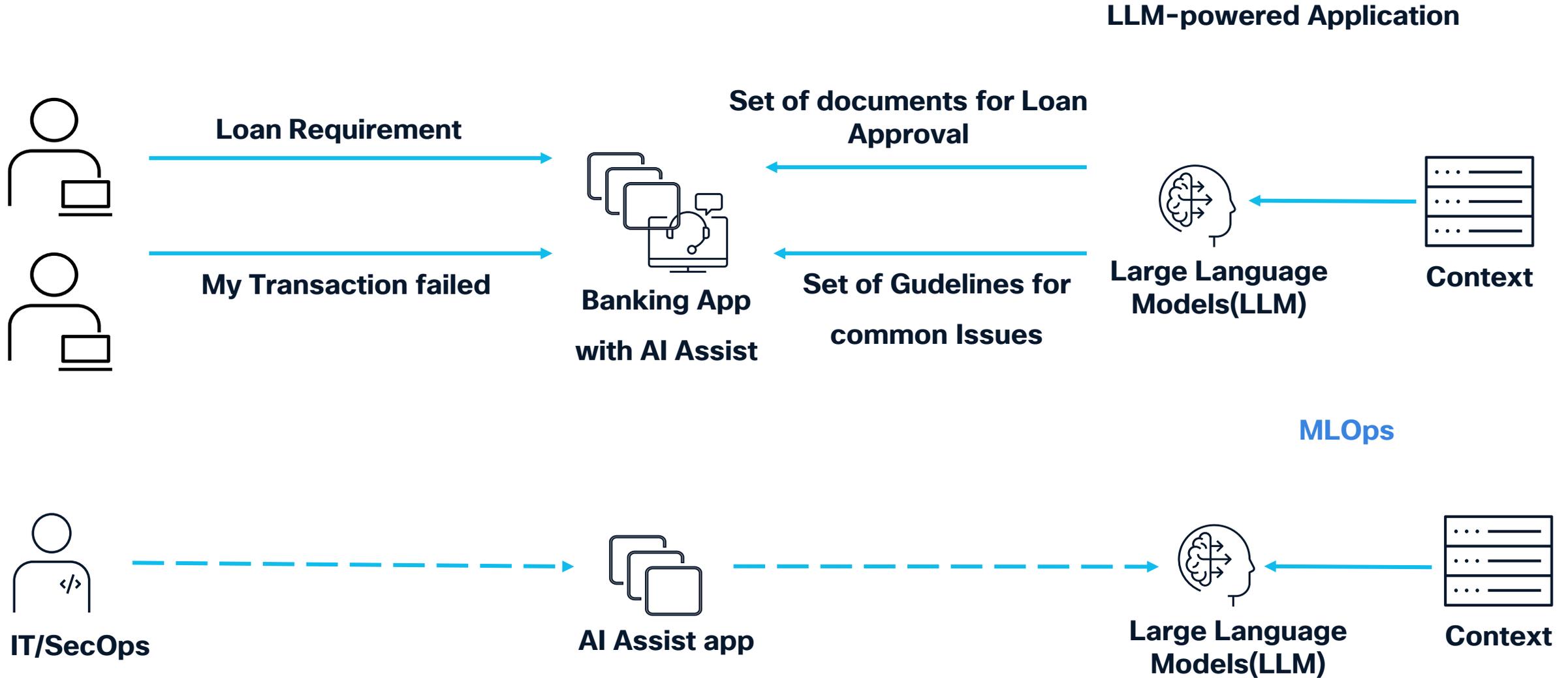
Prior to Cisco – security consulting, implementation and audit



Let's start with a customer facing application

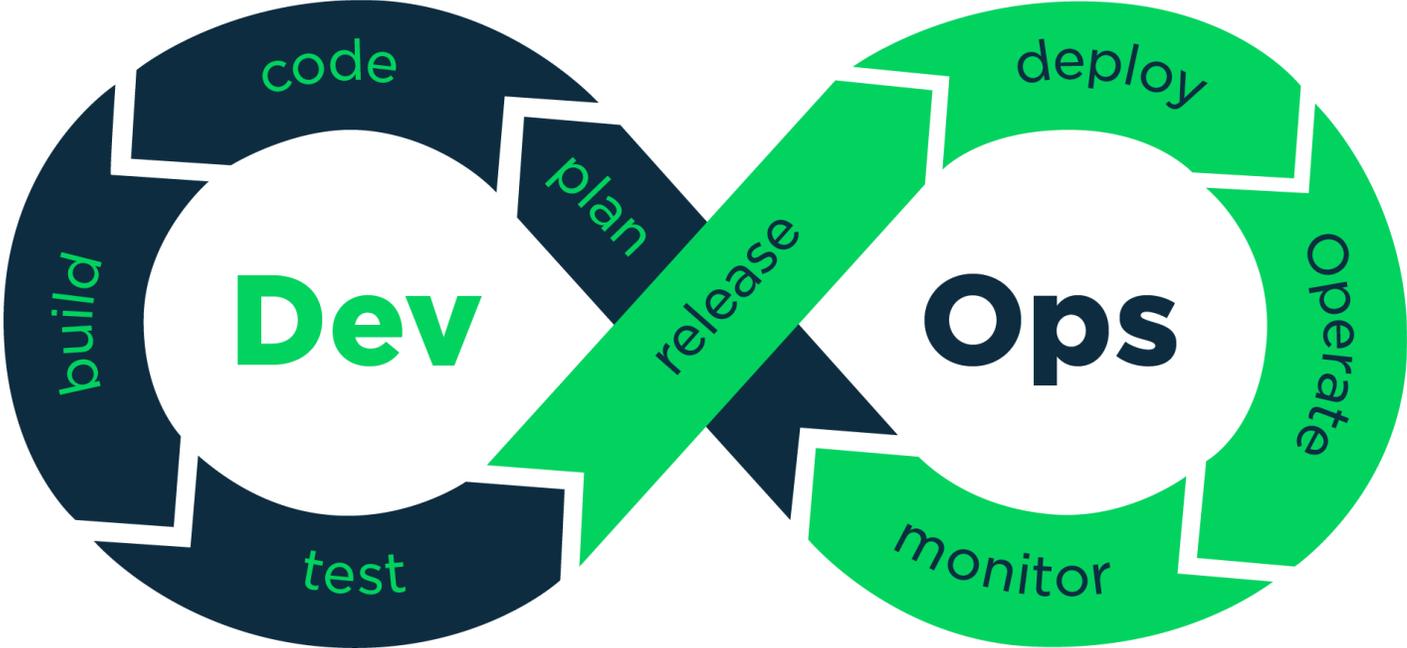


Now let's add AI Assistance to it



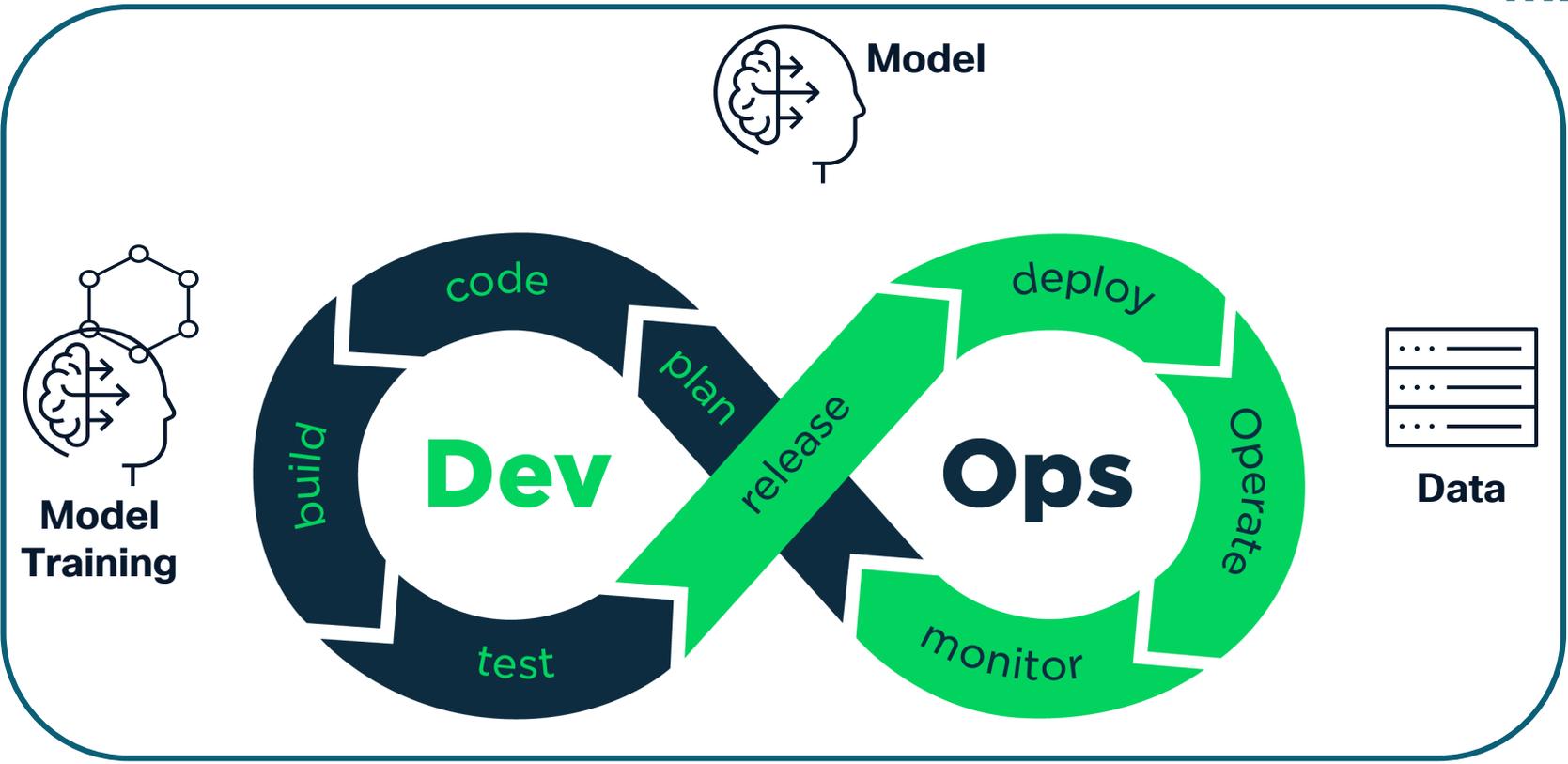
DevOps -> MLOps

Application Lifecycle



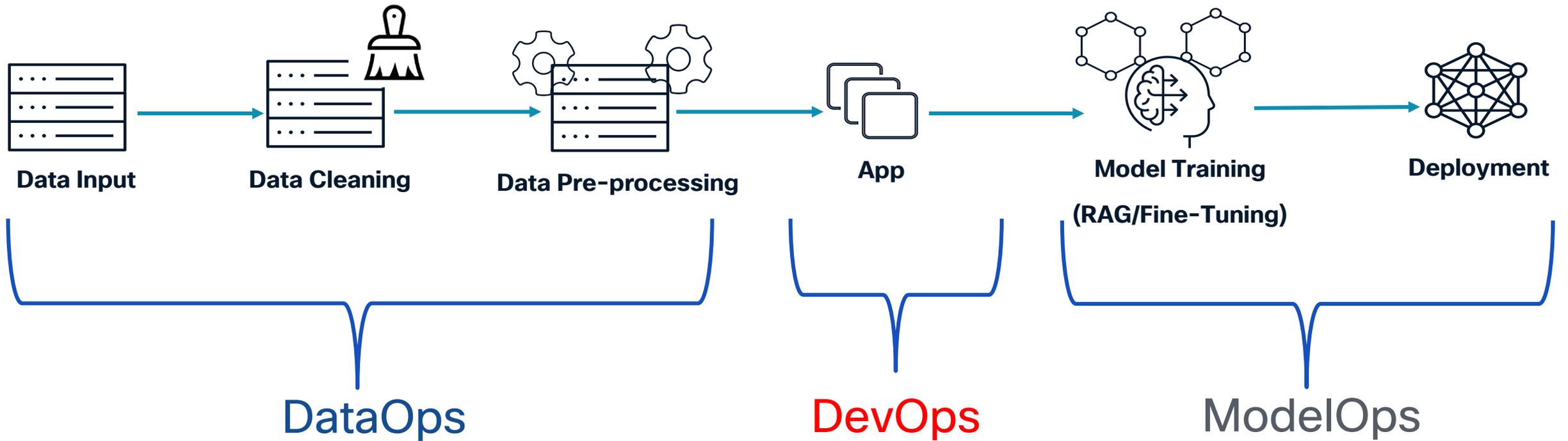
Adding LLM requirements into Application Lifecycle

MLOps



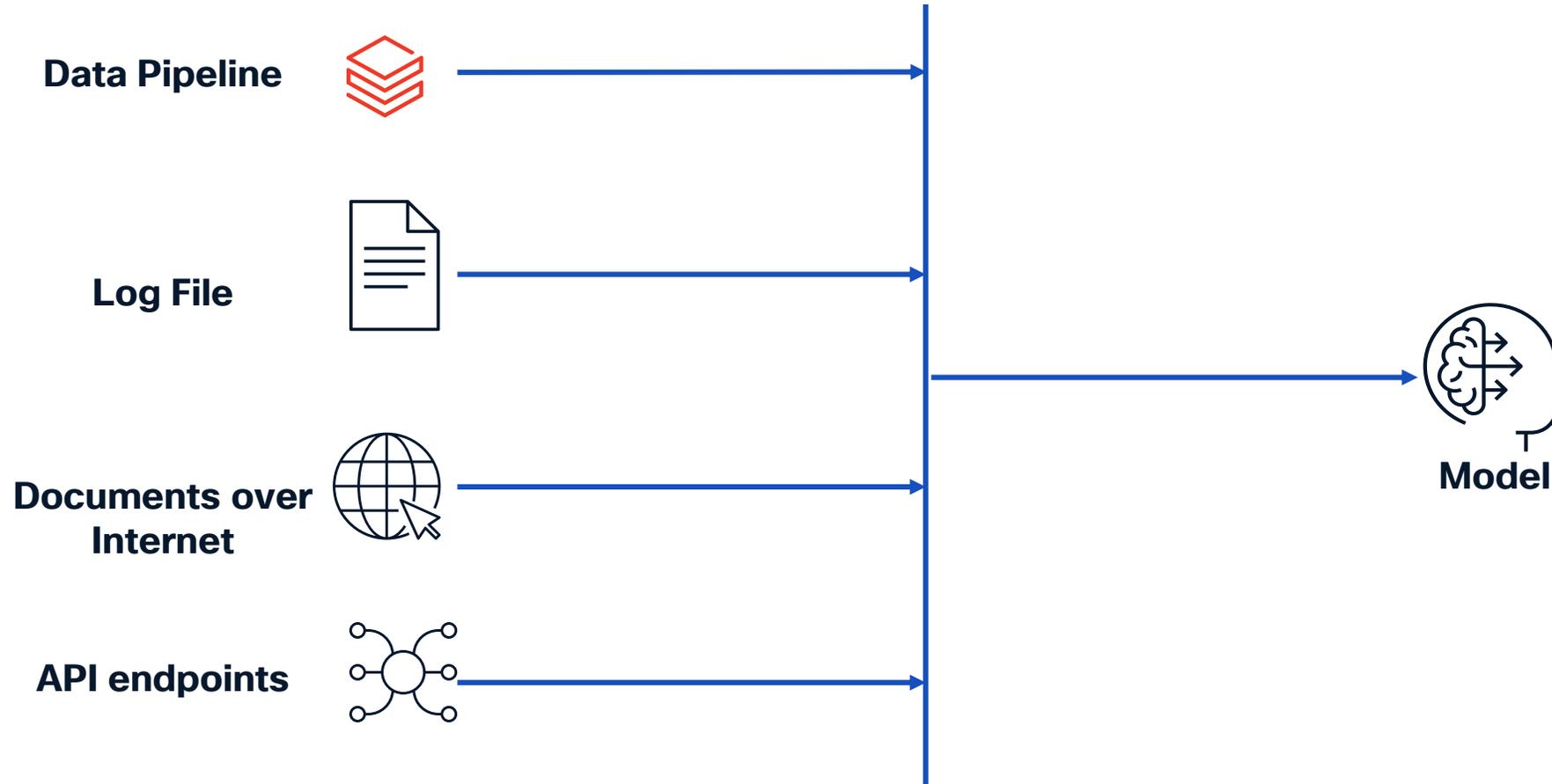
MLOps = DataOps + DevOps + ModelOps

MLOps Core Components



DataOps

Data Input



Data Cleaning Example – Handling log files

2025-04-01 09:59:12 [5568] [WARN] < 14> Dns Protection **IPv4** State Machine: Error: Can only have one disconnected session - we have 2 - not using user specific policies



2025-04-01 09:59:12 [5568] [WARN] < 15> Dns Protection **IPv6** State Machine: Error: Can only have one disconnected session - we have 2 - not using user specific policies



2025-03-28 16:16:12 [8976] [**WARN**] < 10> Device Registration: Exception occurred while deserializing device registration response message; there was likely a problem with the communications channel. Registration will be reattempted later.
System.Runtime.Serialization.SerializationException: There was an error deserializing the object of type Core.RoamingDeviceCreateResponseData. Encountered unexpected character 'T'. --->
System.Xml.XmlException: Encountered unexpected character 'T'.



Data Pre-processing example – convert tables to JSON

Management Center Version	Oldest Device Version You Can Manage
7.7	7.2
7.6	7.1
7.4 Last support for NGIPS device management.	7.0
7.3	6.7
7.2	6.6
7.1	6.5
7.0	6.4
6.7	6.3
6.6	6.2.3
6.5	6.2.3
6.4	6.1
6.3	6.1
6.2.3	6.1
6.2.2	6.1
6.2.1	6.1
6.2	6.1
6.1	5.4.0.2/5.4.1.1
6.0.1	5.4.0.2/5.4.1.1
6.0	5.4.0.2/5.4.1.1
5.4.1	5.4.1 for ASA FirePOWER on the ASA-5506-X series, ASA5508-X, and ASA5516-X. 5.3.1 for ASA FirePOWER on the ASA5512-X, ASA5515-X, ASA5525-X, ASA5545-X, ASA5555-X, and ASA-5585-X series. 5.3.0 for Firepower 7000/8000 series and legacy devices.



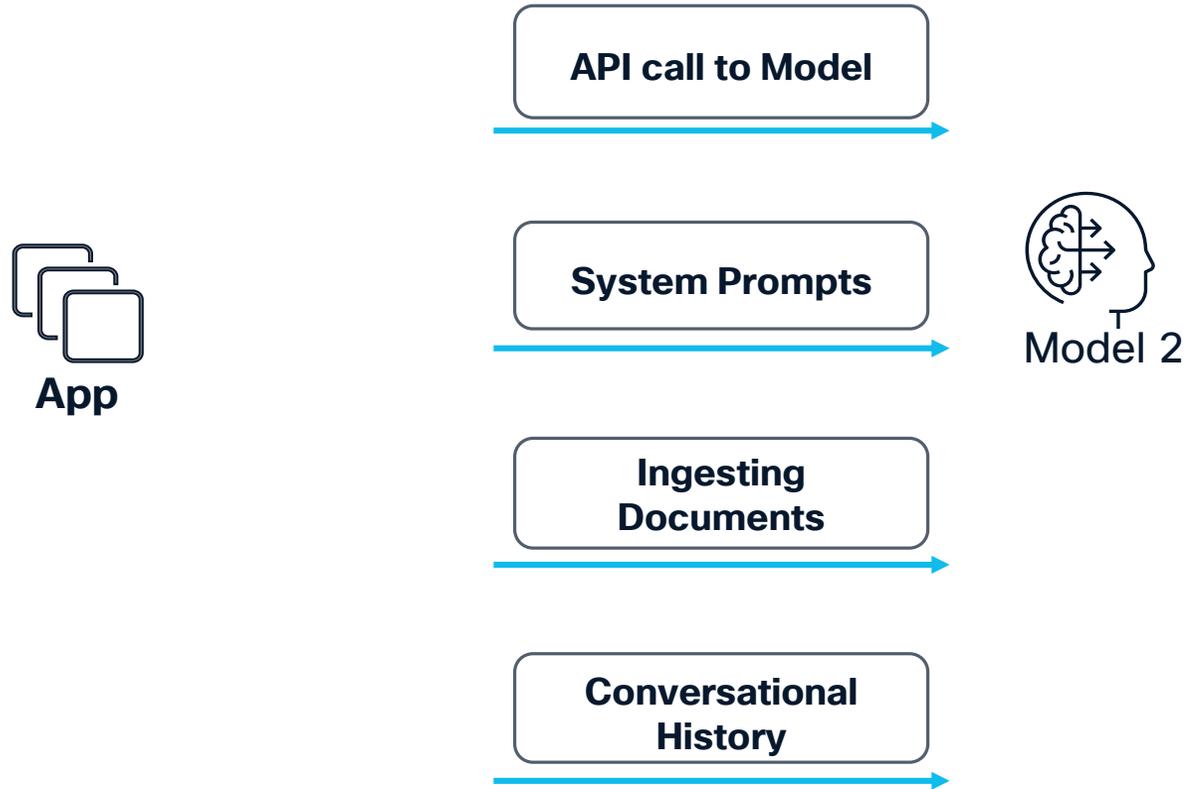
```

Management Center Version Oldest Device
Version You Can Manage\n0 7.7 7.2\n1 7.6
7.1\n2 7.4 Last support for NGIPS device
management. 7.0\n3 7.3 6.7\n4 7.2 6.6\n5
7.1 6.5\n6 7.0 6.4\n7 6.7 6.3\n8 6.6
6.2.3\n9 6.5 6.2.3\n10 6.4 6.1\n11 6.3
6.1\n12 6.2.3 6.1\n13 6.2.2 6.1\n14 6.2.1
6.1\n15 6.2 6.1\n16 6.1
5.4.0.2/5.4.1.1\n17 6.0.1
5.4.0.2/5.4.1.1\n18 6.0
5.4.0.2/5.4.1.1\n19 5.4.1 5.4.1 for ASA
FirePOWER on the ASA-5506-X seri...
    
```

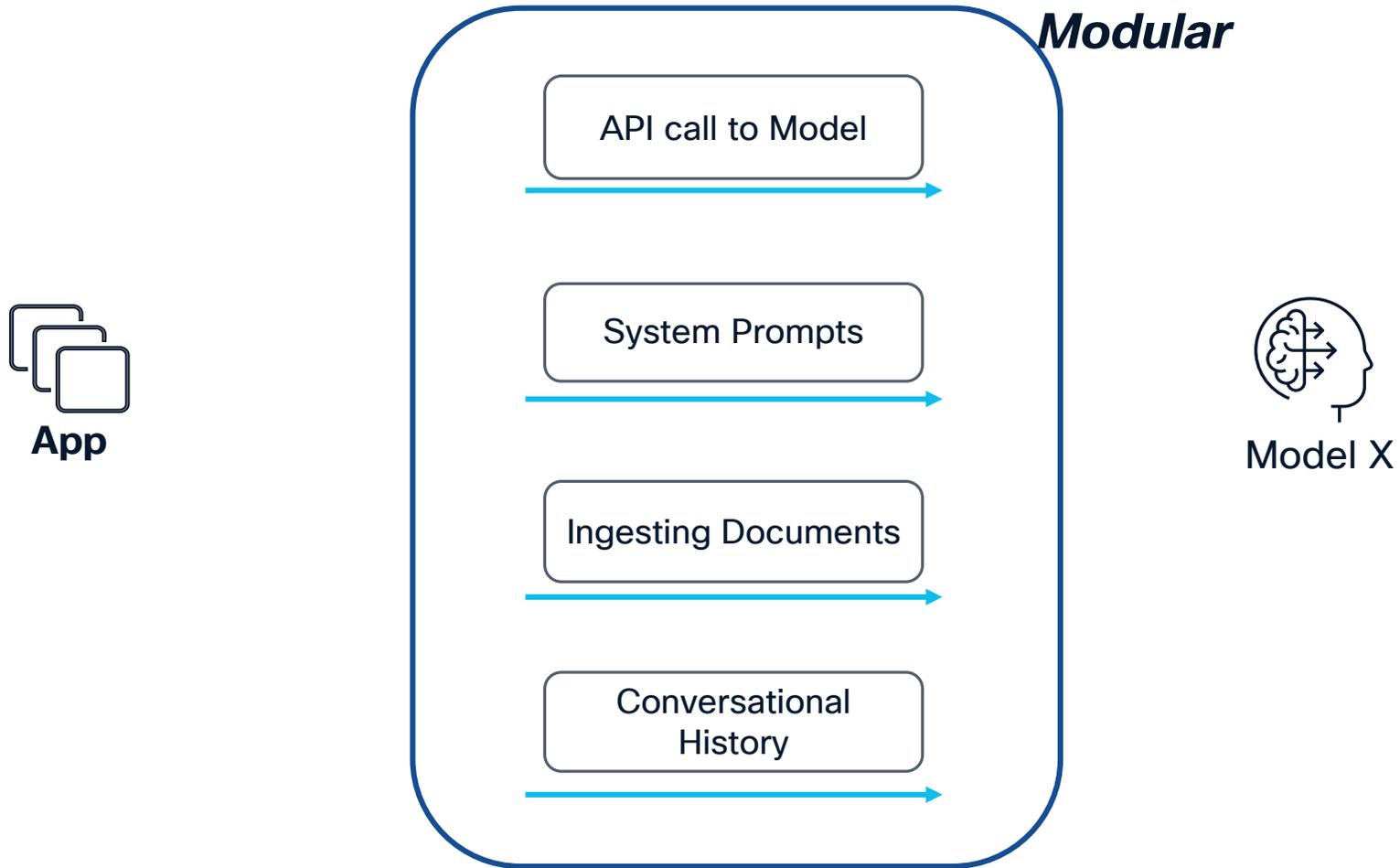


DevOps for LLM inclusion

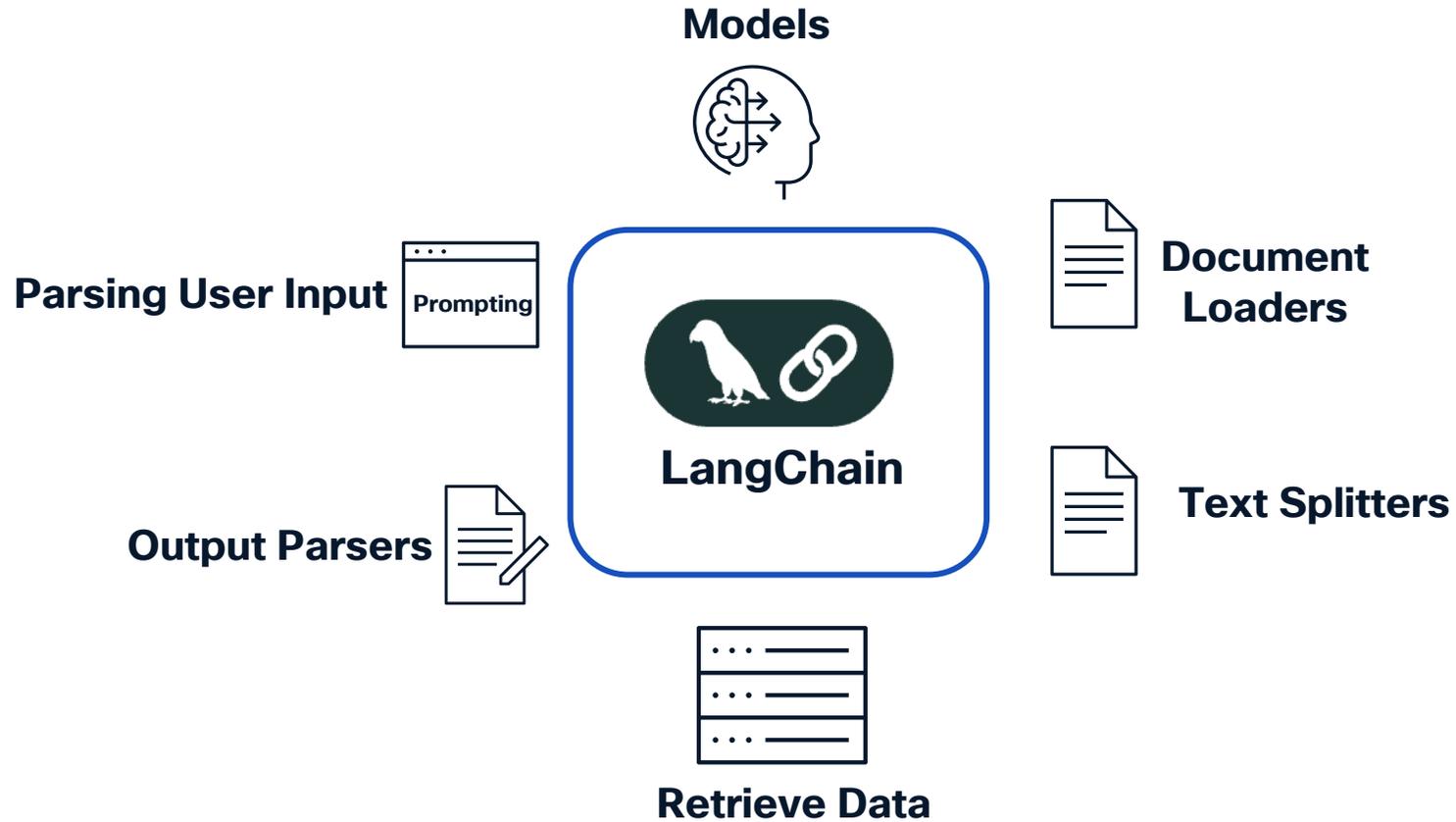
Interactions with LLMs



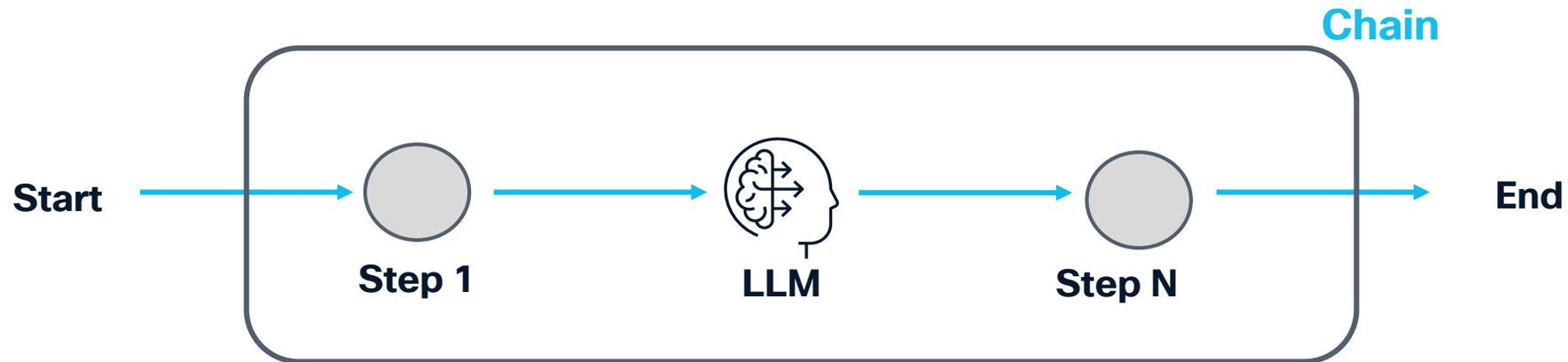
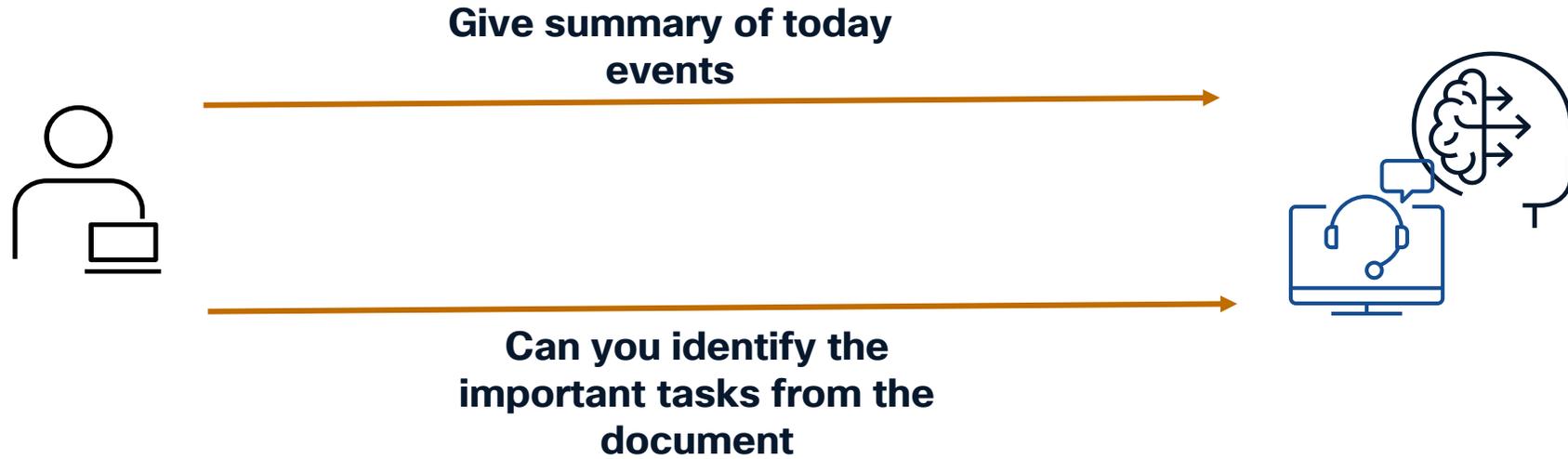
Interactions with LLMs – a modular approach



Integration Layer to the rescue -> Model Agnostic

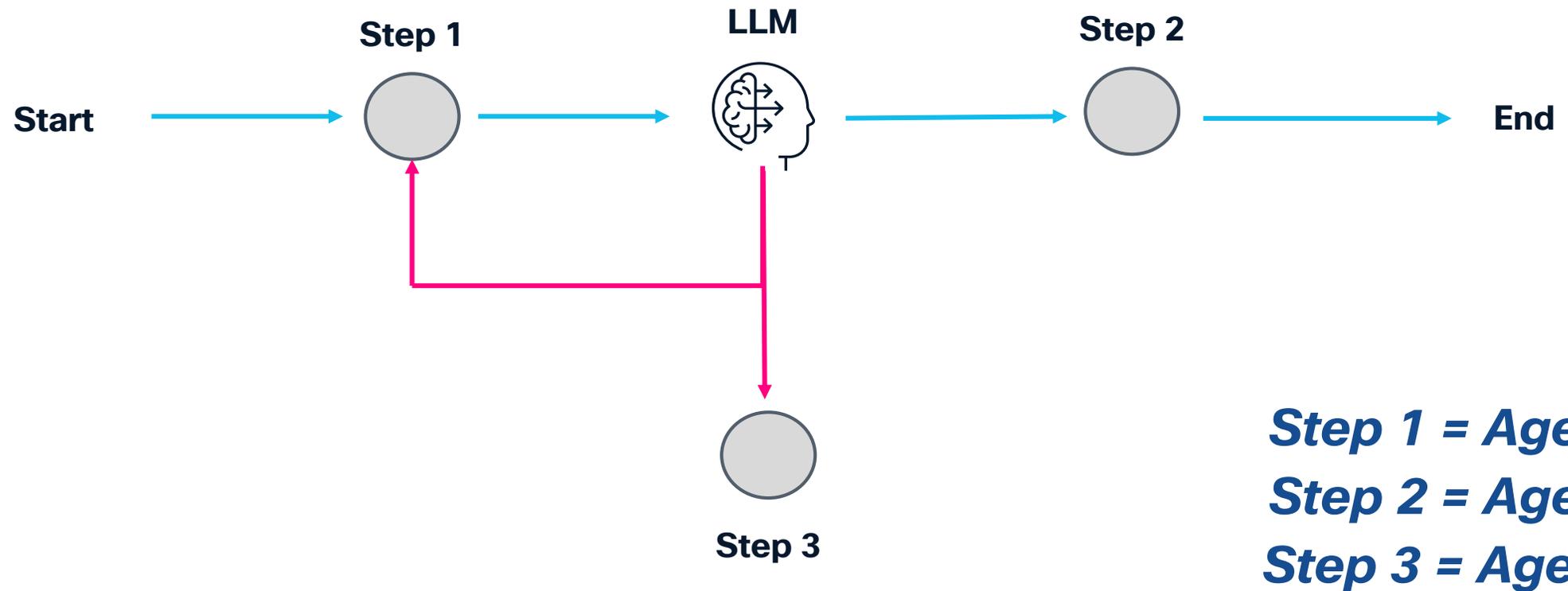


Linear Control Flow – a limiting factor

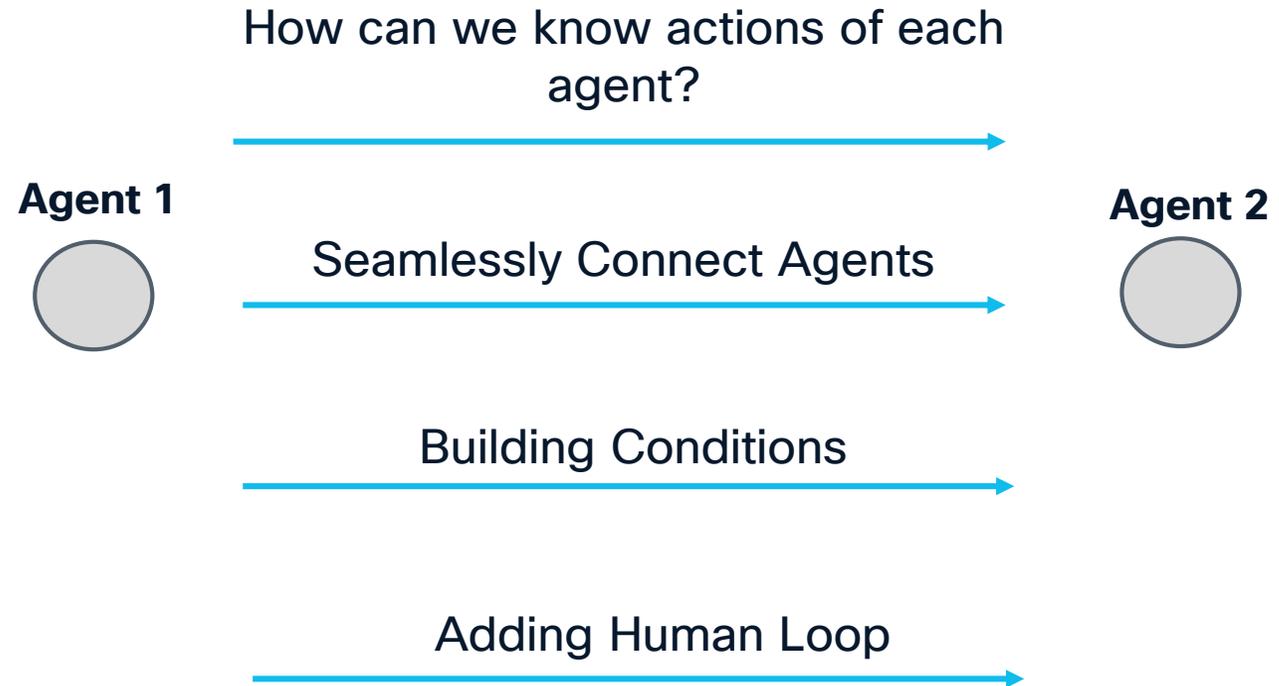


Non-Linear Control Flow with Agentic AI

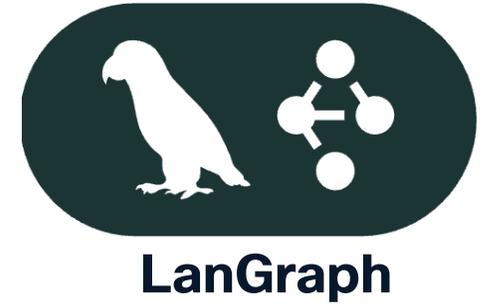
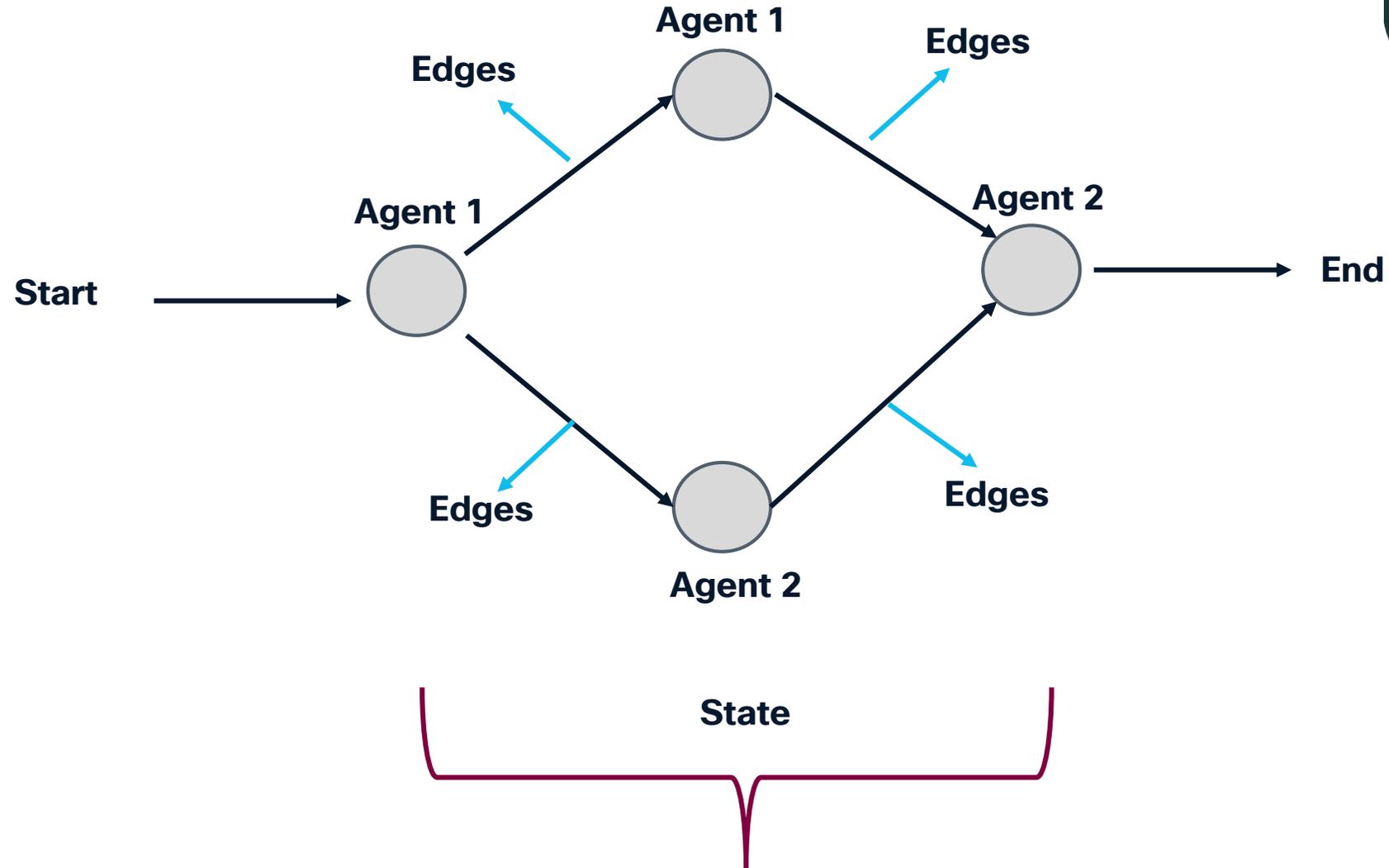
Agentic



Building Blocks for Agentic AI



Orchestration Framework for Agentic AI

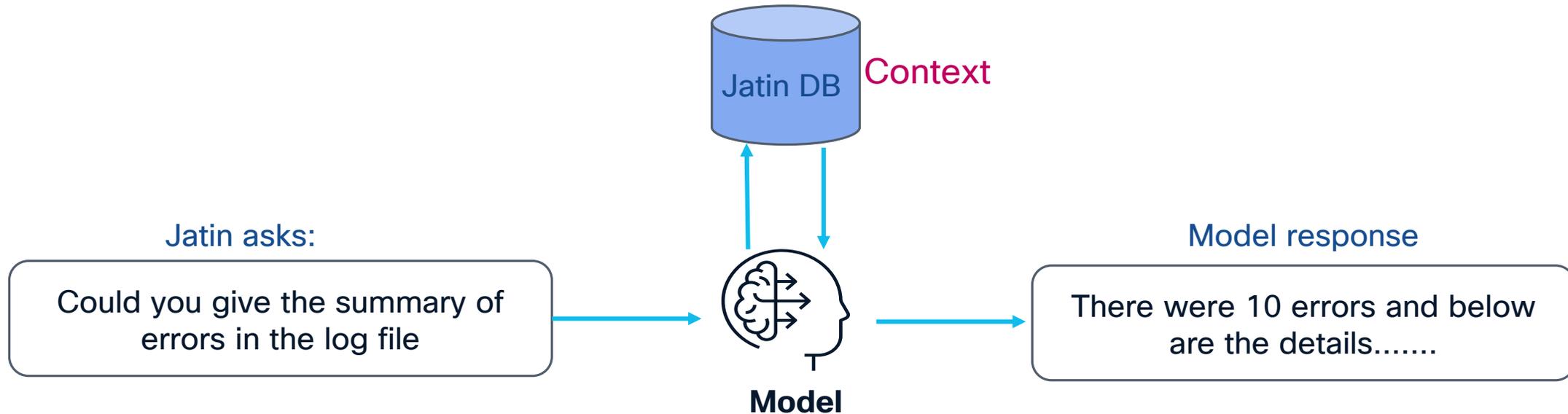


ModelOps

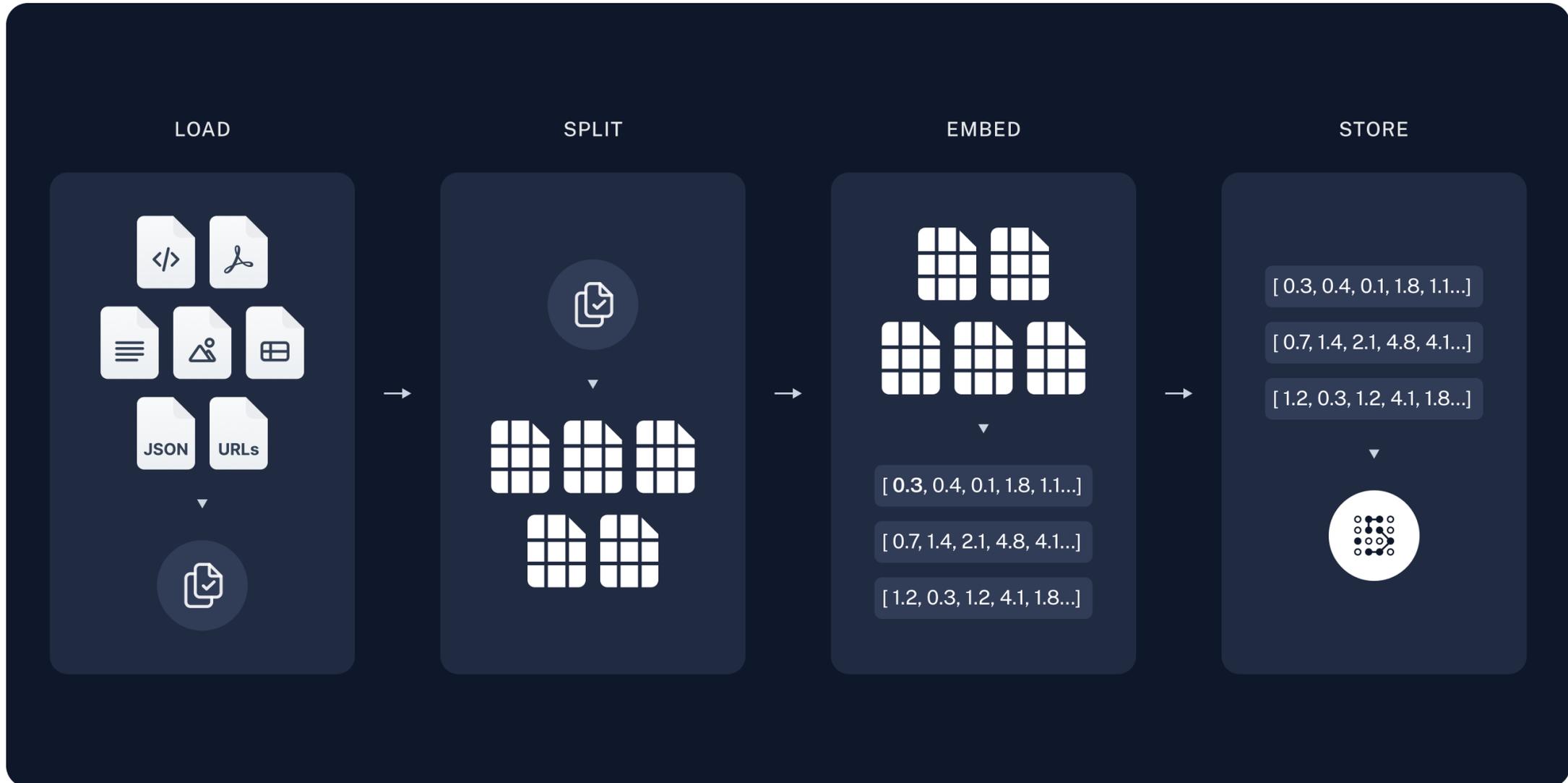
Retrieval Augmented Generation(RAG)

Definition:

- Allows LLM to query external data source to derive context



RAG Pipeline – Steps to make data LLM friendly



Jargon you may hear – for reference only

- **Embeddings** : Objects like text, audio, video and images are represented as “**Vectors**”.
- **Embedding Models**: Specially trained ML models to represent embeddings in multi-dimensional space.
- **Vector Stores** : A database which is is specialized in storing the embeddings.
- **Tokens** : Units of text being processed

However RAG is becoming a Legacy option

Early 2024

32k tokens to 128k tokens



Currently

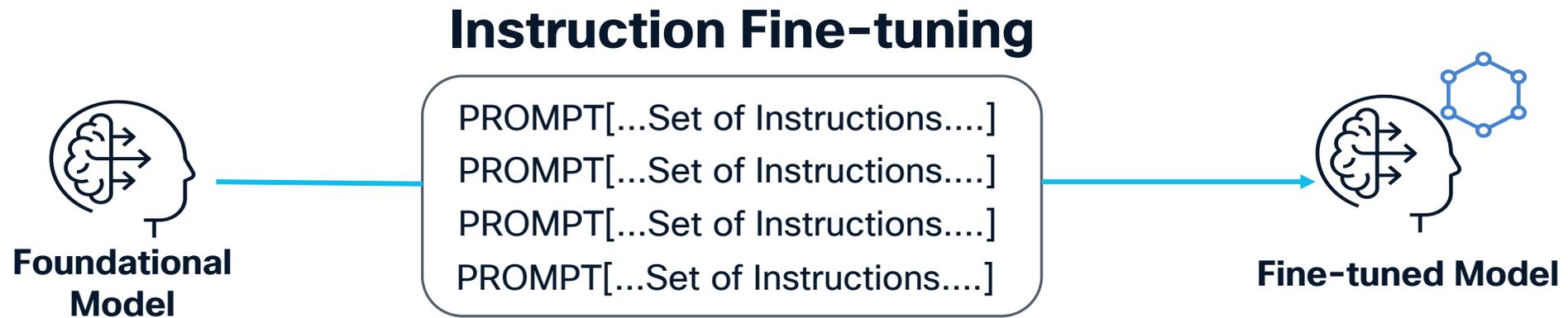
Rate limits

Rate limits ensure fair and reliable access to the API by placing specific caps on requests or tokens used within a given time period. Your usage tier determines how high these limits are set and automatically increases as you send more requests and spend more on the API.

TIER	RPM	TPM	BATCH QUEUE LIMIT
Free		Not supported	
Tier 1	500	30,000	90,000
Tier 2	5,000	450,000	1,350,000
Tier 3	5,000	800,000	50,000,000
Tier 4	10,000	2,000,000	200,000,000
Tier 5	10,000	30,000,000	5,000,000,000



Modern LLMs token limits make Fine Tuning easier

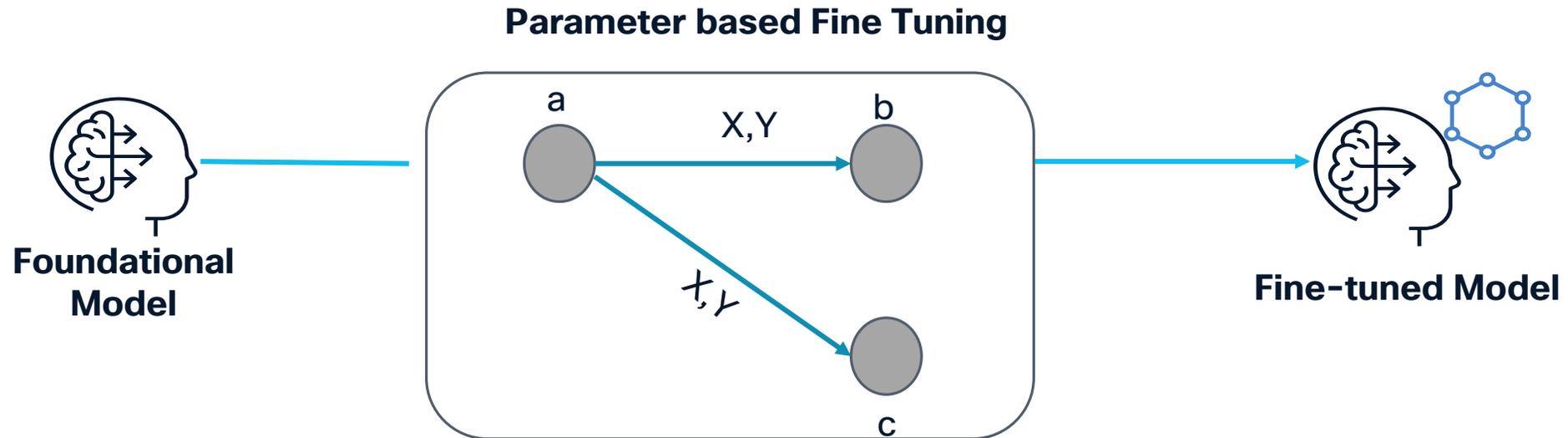


Using Prompts to Fine-Tune LLM

Fine Tuning for advanced use cases

Parameters : a , b , c

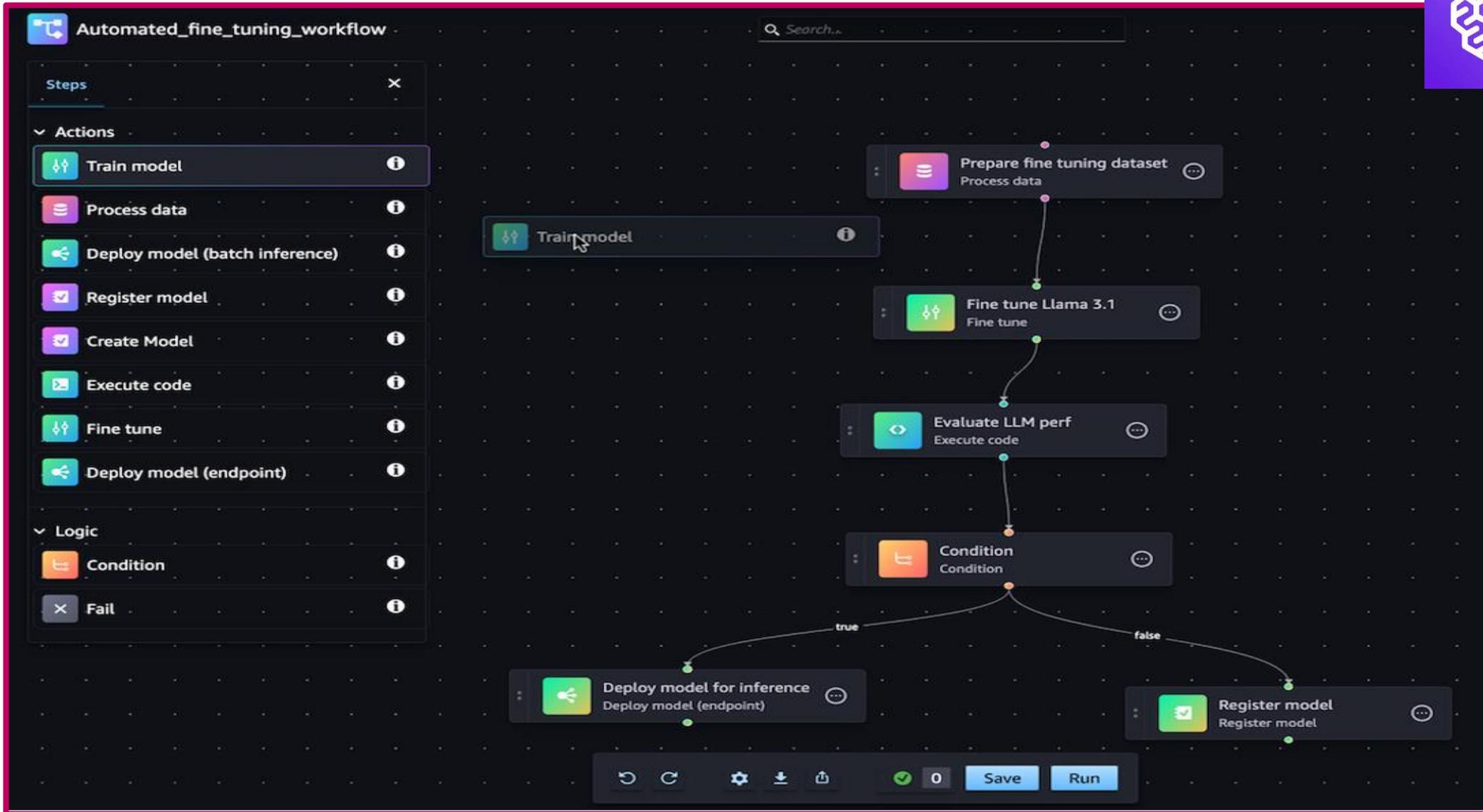
Weights : Y,X



$$Y = a * X + b$$

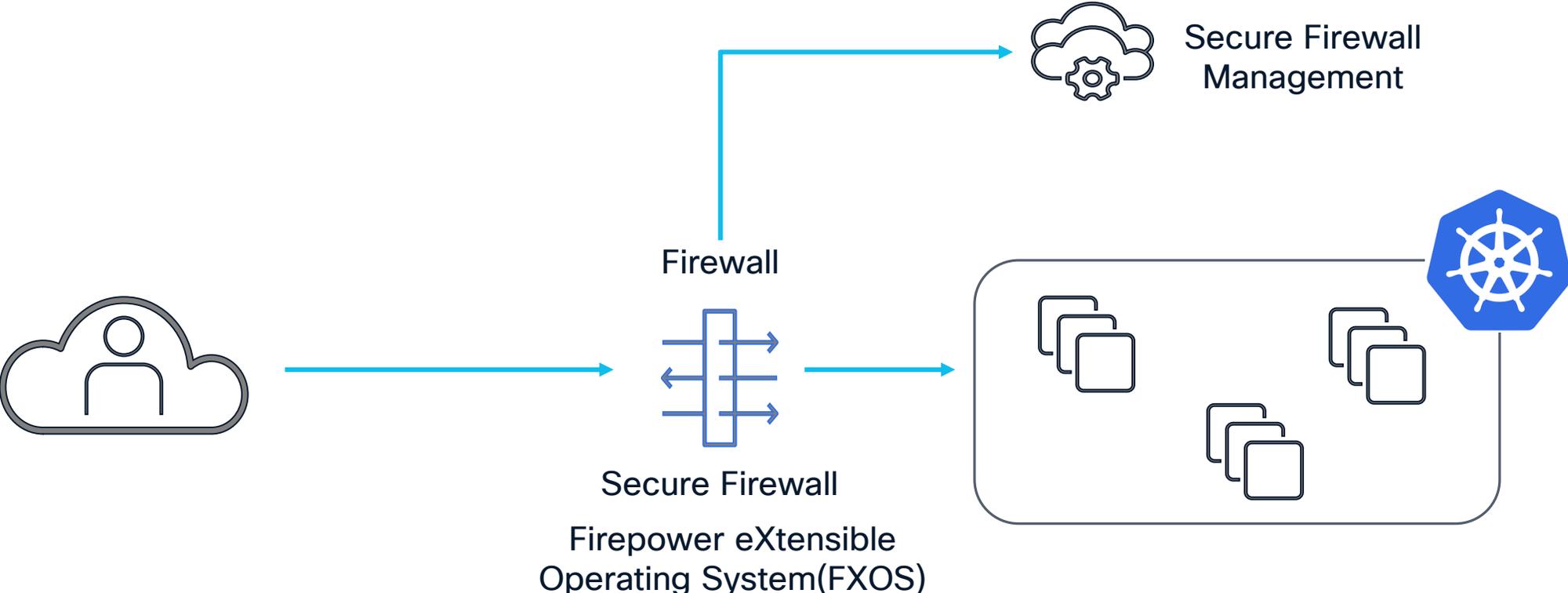
$$Y = a * X + c$$

Model Deployment Example for Reference

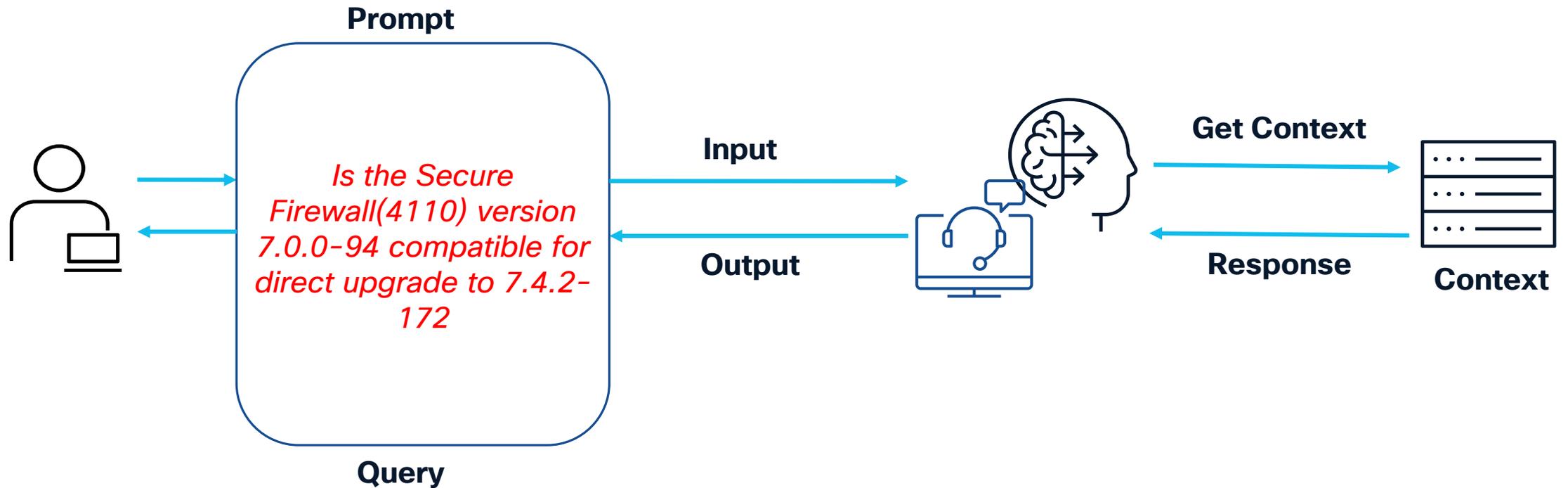


Our Sample LLM application

An example SecOps Use Case – Firewall Upgrades!



Sample LLM Application – Firewall Upgrade Assistant



Upgrade assistant which answers user queries for different Cisco products

Simple Python based LLM app

LLM >  main.py > ...

```
7 from state import UpgradeState
8 from typing import Literal
9 from langchain_core.output_parsers import JsonOutputParser
10 from langchain.agents import create_react_agent, AgentExecutor, create_openai_functions_agent, OpenAIFunctionsA
11 from dotenv import load_dotenv
12 load_dotenv()
13
14
15
16 def compatibility_route(state)->Literal["tools", "end" ]:
17     messages = state["messages"]
18     last_message = messages[-1]
19     if last_message.tool_calls:
20         return "tools"
21     else:
22         return "end"
23 tools = [fmc_hardware_compatibility, fmc_hardware_bios_compatibility, fmc_virtual_300_public_cloud, fmc_virtual_pu
24 compatibility_check_tool_node = ToolNode(tools)
25
```

Simple Python based LLM app

LLM >  main.py > ...

```
26 def compatibility_agent(state):
78     llm = ChatOpenAI(
79         model= "gpt-4o-mini",
80         temperature=0
81     )
82
83     compatibility_with_tools = llm.bind_tools(tools)
84
85     chat_prompt_template = chat_prompt_template.partial(compatibility_prompt=compatibility_prompt,tools=tools)
86
87     chat_prompt_template = chat_prompt_template | compatibility_with_tools
88     #print("chat prompt template\n",chat_prompt_template)
89     chat_response = chat_prompt_template.invoke(state)
90     resp_dict = {"messages": [chat_response]}
```

Knowledge Sources

*Management Center
Compatibility Guide*

<https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/management-center-compatibility.html>

*Secure Firewall Threat
Defense Compatibility Guide*

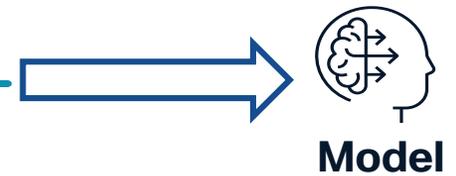
<https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/threat-defense-compatibility.html>

Firepower Compatibility Guide

<https://www.cisco.com/c/en/us/td/docs/security/firepower/fxos/compatibility/fxos-compatibility.html>

ASA Compatibility Guide

<https://www.cisco.com/c/en/us/td/docs/security/asa/compatibility/asamatrix.html>



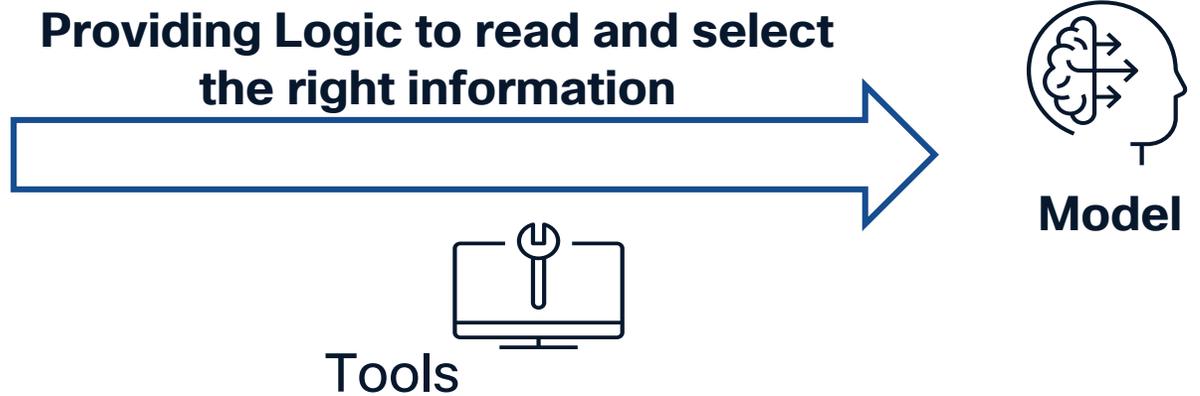
Knowledge sources in code

```
LLM > main.py > compatibility_agent
16 def compatibility_route(state)->Literal["tools", "end" ]:
17     last_message = messages[-1]
18     if last_message.tool_calls:
19         return "tools"
20     else:
21         return "end"
22
23     (variable) compatibility_check_tool_node: ToolNode ;_compatibility,fmc_virtual_300_public_cloud,fmc_virtual_public_cloud,fmc_virtual_300
24     compatibility_check_tool_node = ToolNode(tools)
25
26 def compatibility_agent(state):
27
28
29
30
31     tools = [fmc_hardware_compatibility,fmc_hardware_bios_compatibility,fmc_virtual_300_public_cloud,fmc_virtual_public_cloud,fmc_virtual
32
33
34     COMP_URL = [
35         "https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/threat-defense-compatibility.html",
36         "https://www.cisco.com/c/en/us/td/docs/security/firepower/fxos/compatibility/fxos-compatibility.html",
37         "https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/management-center-compatibility.html",
38         "https://www.cisco.com/c/en/us/td/docs/security/asa/compatibility/asamatrix.html"
39
40     ]
41
42
43     #correct template1
44
45     compatibility_prompt = SystemMessage(f""""You are a TAC expert and your primary role is to solve compatibility queries for different C
46     reference different tables which are present in a variable 'table_name' in tools. Do not generat
47     You need to retrieve tables from the tools and the tables would give the information in terms of
48     Read the table information and answer if the product/versions are compatible or not. If not prov
49     Select the URL defined in list COMP_URL depending on the tool prefix.
```

Functions

Table 9. On-Prem Management Center- Device Compatibility

Management Center Version	Oldest Device Version You Can Manage
7.7	7.2
7.6	7.1
7.4 Last support for NGIPS device management.	7.0
7.3	6.7
7.2	6.6
7.1	6.5
7.0	6.4
6.7	6.3
6.6	6.2.3
6.5	6.2.3
6.4	6.1
6.3	6.1
6.2.3	6.1
6.2.2	6.1
6.2.1	6.1
6.2	6.1
6.1	5.4.0.2/5.4.1.1
6.0.1	5.4.0.2/5.4.1.1
6.0	5.4.0.2/5.4.1.1
5.4.1	5.4.1 for ASA FirePOWER on the ASA-5506-X series, ASA5508-X, and ASA5516-X. 5.3.1 for ASA FirePOWER on the ASA5512-X, ASA5515-X, ASA5525-X, ASA5545-X, ASA5555-X, and ASA-5585-X series. 5.3.0 for Firepower 7000/8000 series and legacy devices.



Knowledge sources in code

```
LLM > main.py > compatibility_agent
16 def compatibility_route(state)->Literal["tools", "end" ]:
17     last_message = messages[-1]
18     if last_message.tool_calls:
19         return "tools"
20     else:
21         return "end"
22
23     (variable) compatibility_check_tool_node: ToolNode ;_compatibility,fmc_virtual_300_public_cloud,fmc_virtual_public_cloud,fmc_virtual_300
24     compatibility_check_tool_node = ToolNode(tools)
25
26 def compatibility_agent(state):
27
28
29
30
31     tools = [fmc_hardware_compatibility,fmc_hardware_bios_compatibility,fmc_virtual_300_public_cloud,fmc_virtual_public_cloud,fmc_virtual
32
33
34     COMP_URL = [
35         "https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/threat-defense-compatibility.html",
36         "https://www.cisco.com/c/en/us/td/docs/security/firepower/fxos/compatibility/fxos-compatibility.html",
37         "https://www.cisco.com/c/en/us/td/docs/security/secure-firewall/compatibility/management-center-compatibility.html",
38         "https://www.cisco.com/c/en/us/td/docs/security/asa/compatibility/asamatrix.html"
39
40     ]
41
42
43     #correct template1
44
45     compatibility_prompt = SystemMessage(f""""You are a TAC expert and your primary role is to solve compatibility queries for different C
46     reference different tables which are present in a variable 'table_name' in tools. Do not generat
47     You need to retrieve tables from the tools and the tables would give the information in terms of
48     Read the table information and answer if the product/versions are compatible or not. If not prov
49     Select the URL defined in list COMP_URL depending on the tool prefix.
```

Tools – function that converts tables to JSON

```
LLM > tools.py > ...
310 @tool
311 def ftd_3100_4200_compatibility(URL):
312     """
313     | This function gives the information on compatibility for 3100 and 4200 ftd devices."""
314     table_name = "Secure Firewall 3100/4200 Series Application Mode Compatibility"
315     heading = None
316     table_context = table_ref_id(URL)
317     return table_context
318
319 @tool
320 def ftd_multi_instance_compatibility(URL):
321     """
322     | This function gives the information on multi-instance mode for 3100/4200 ftd devices."""
323     table_name = "Secure Firewall 3100/4200 Series Multi-Instance Mode Compatibility"
324     heading = None
325     table_context = table_ref_id(URL, table_name, heading)
326     return table_context
327
328 @tool
329 def ftd_4100_9300_compatibility(URL):
330     """
331     | This function gives the information on ftd and fxos compatibility for 4100/9300 ftd devices"""
332     table_name = "Firepower 4100/9300 Compatibility"
333     heading = None
334     table_context = table_ref_id(URL, table_name, heading)
335     return table_context
```

Fine Tuning with Prompt Engineering



Model

```
LLM > main.py > compatibility_agent
26 def compatibility_agent(state):
45     compatibility_prompt = SystemMessage(f"""You are a TAC expert and your primary role is to solve compatibility queries for different C
46     reference different tables which are present in a variable 'table_name' in tools. Do not generat
47     You need to retrieve tables from the tools and the tables would give the information in terms of
48     Read the table information and answer if the product/versions are compatible or not. If not prov
49     Select the URL defined in list COMP_URL depending on the tool prefix.
50
51     1) If the tool name starts with "fmc" kindly use the URL ending in "management-center-compatibility.html"
52     2) If the tool name starts with "ftd" kindly use the URL ending in "threat-defense-compatibility.html"
53     3) If the tool name starts with "asa" kindly use the URL ending in "asamatrix.html"
54     4) If the tool name starts with "fxos" kindly use the URL ending in "fxos-compatibility.html"
55     Below are few acronyms which you should be aware of
56     FMC : Secure Firewall Management Center
57     FTD : Secure Firewall Thread Defense
58     FXOS: Firepower eXtensible Operating System
59     ASA : Adaptive Security Appliance
60     The query input would be given by the user
61     The tools would give the context. Each tool has a description and try to map the query to the description provided. These would h
62     choosing the correct table for answering the query.
63     The list of URL is defined in COMP_URL={COMP_URL}
64     Return the response if the products are compatible to direct or step upgrade required.
65
66     """)
67     chat_prompt_template= ChatPromptTemplate.from_messages(
68     [
69         (
70             "system",
71             "You are intelligent assistant which resolves compatible queries."
72             "Your response should be in 100 words."
73             "You have access to the following tools {tools}\n {compatibility_prompt}"
74         ),
75         MessagesPlaceholder("messages")
```

Observability

The screenshot displays the LangSmith observability interface. On the left is a navigation sidebar with categories like Home, Observability, Tracing Projects, Monitoring, Evaluation, Datasets & Experiments, Annotation Queue, Prompt Engineering, Prompts, Playground, LangGraph Platform, and Deployments. The main area is titled 'TRACE' and shows a waterfall chart for a 'LangGraph - 10.16s' trace. The chart has a time axis from 1s to 10s. Below the chart, a list of trace events is shown, including:

- LangGraph - 10.16s
- __start__ - 0.00s
- ChannelWrite<...> - 0.00s
- ChannelWrite<branch:to:initial... - 0.00s
- initial_agent - 4.30s
- RunnableSequence - 4.19s
- ChatPromptTemplate - 0.00s
- ChatOpenAI - 4.18s
- ChannelWrite<...> - 0.00s
- compatibility_route - 0.00s
- calling_tools - 0.49s
- fmc_on_prem_ftd_compatibility - 0.48s
- ftd_4100_9300_compatibility - 0.48s
- fmc_fxos_compatibility - 0.48s
- ChannelWrite<...> - 0.00s
- ChannelWrite<branch:to:initial... - 0.00s
- initial_agent - 5.31s
- RunnableSequence - 5.25s
- ChatPromptTemplate - 0.00s

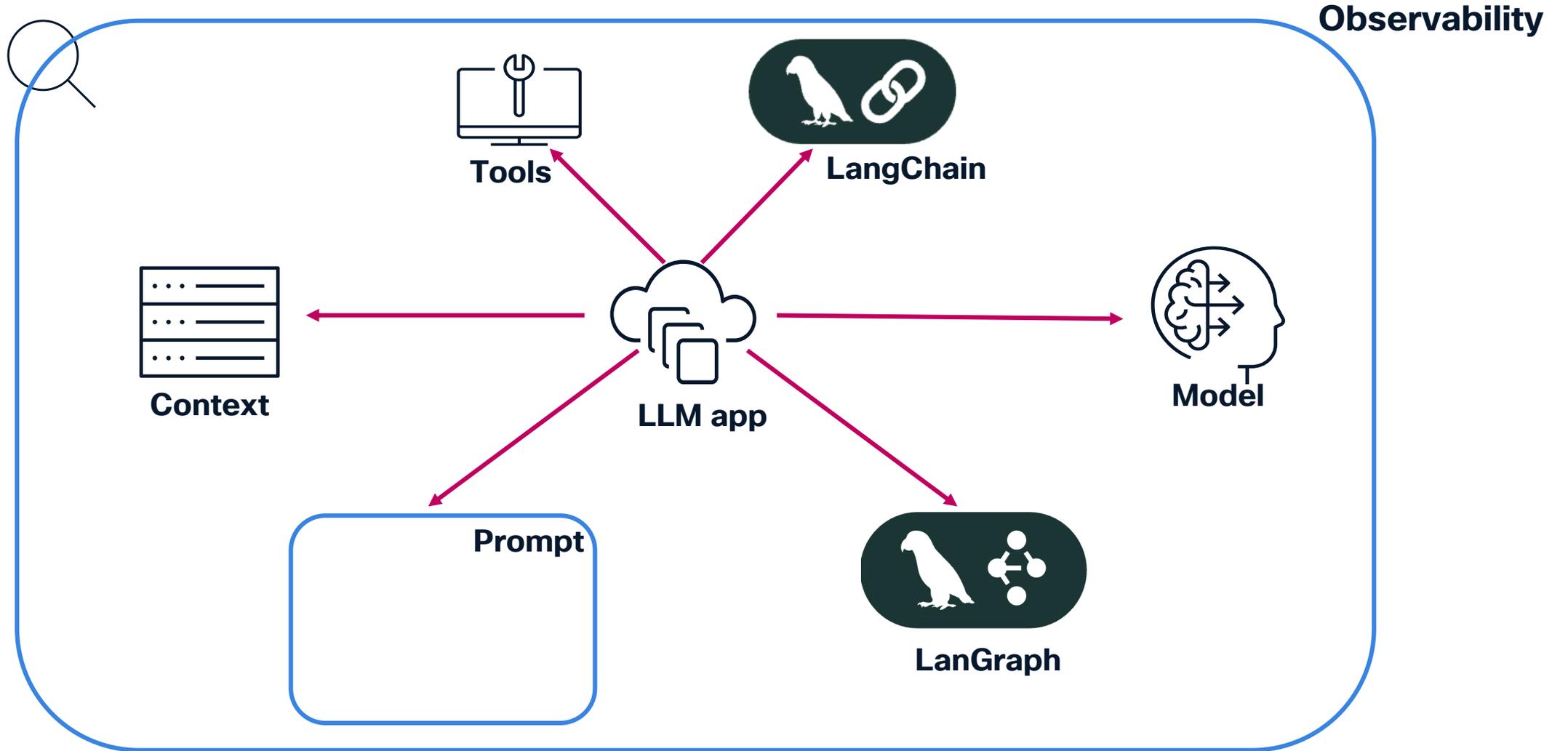
On the right, the 'LangGraph' section is visible, showing 'Run', 'Feedback', and 'Metadata' tabs. The 'Raw Input' section displays a JSON object:

```
{  "messages": [    {      "content": "We are currently running below versions FMC virtual : 7.2.0-82,FTD(4110): 7.0.0-94,FXOS: 2.14(1),Could you help us know if these versions are compatible for a direct upgrade of FTD to recommended release 7.4.2-172. Remember FXOS compatibility as well",      "additional_kwargs": {},      "response_metadata": {},      "type": "human",      "example": false    }  ],  }
```

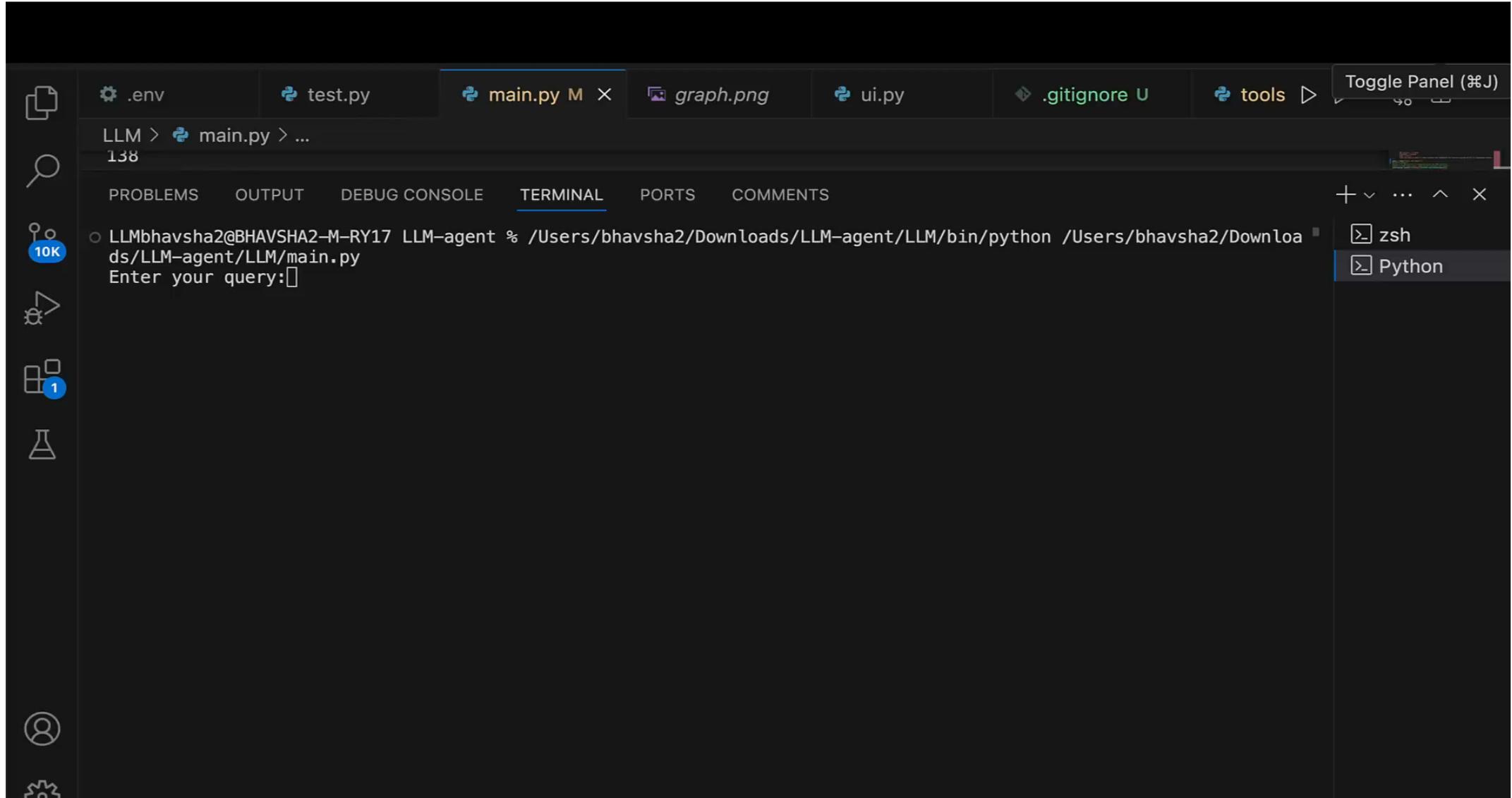
The 'Output' section shows the 'HUMAN' response:

We are currently running below versions FMC virtual : 7.2.0-82,FTD(4110): 7.0.0-94,FXOS: 2.14(1),Could you help us know if these versions are compatible for a direct upgrade of FTD to recommended release 7.4.2-172. Remember FXOS compatibility as well

Putting together our LLM Application

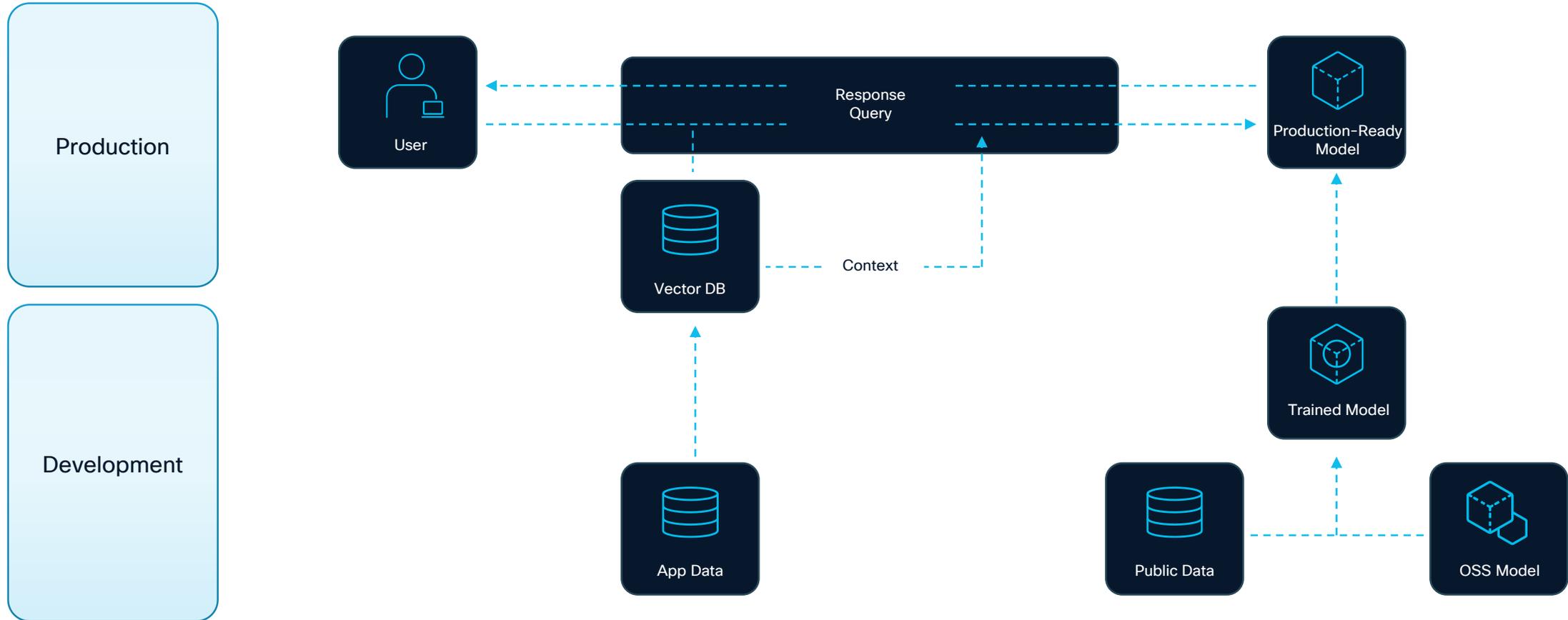


Quick demo of our final app

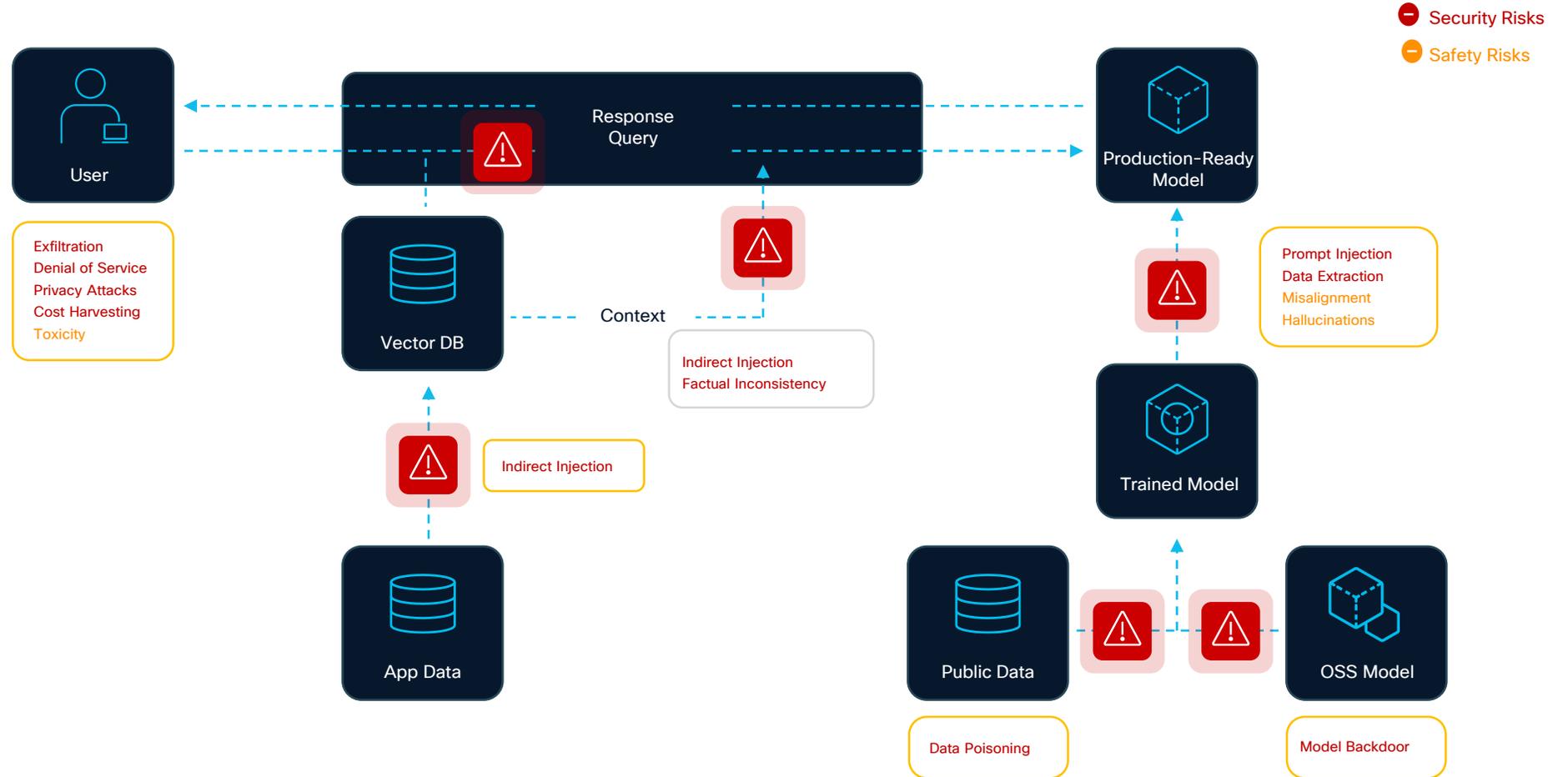
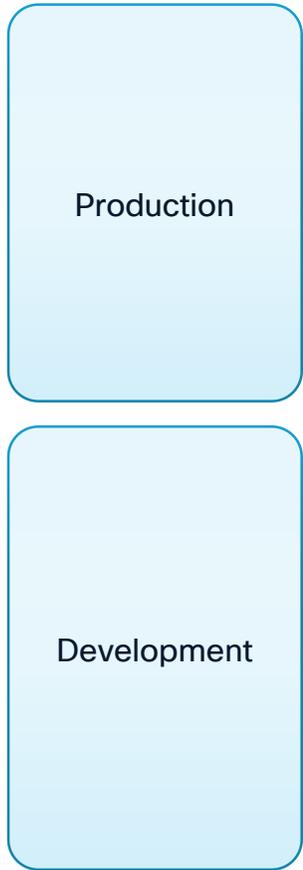


Security Threats

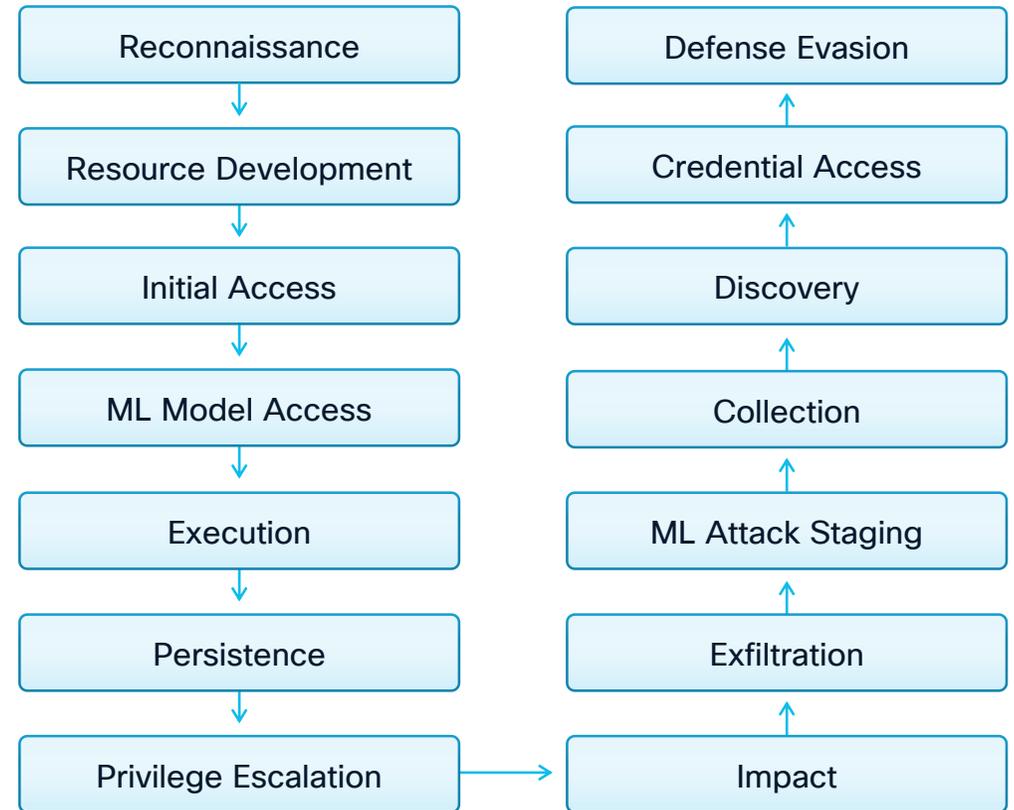
Let's recap how enterprises use LLMs in enterprise apps



Where are the threats?



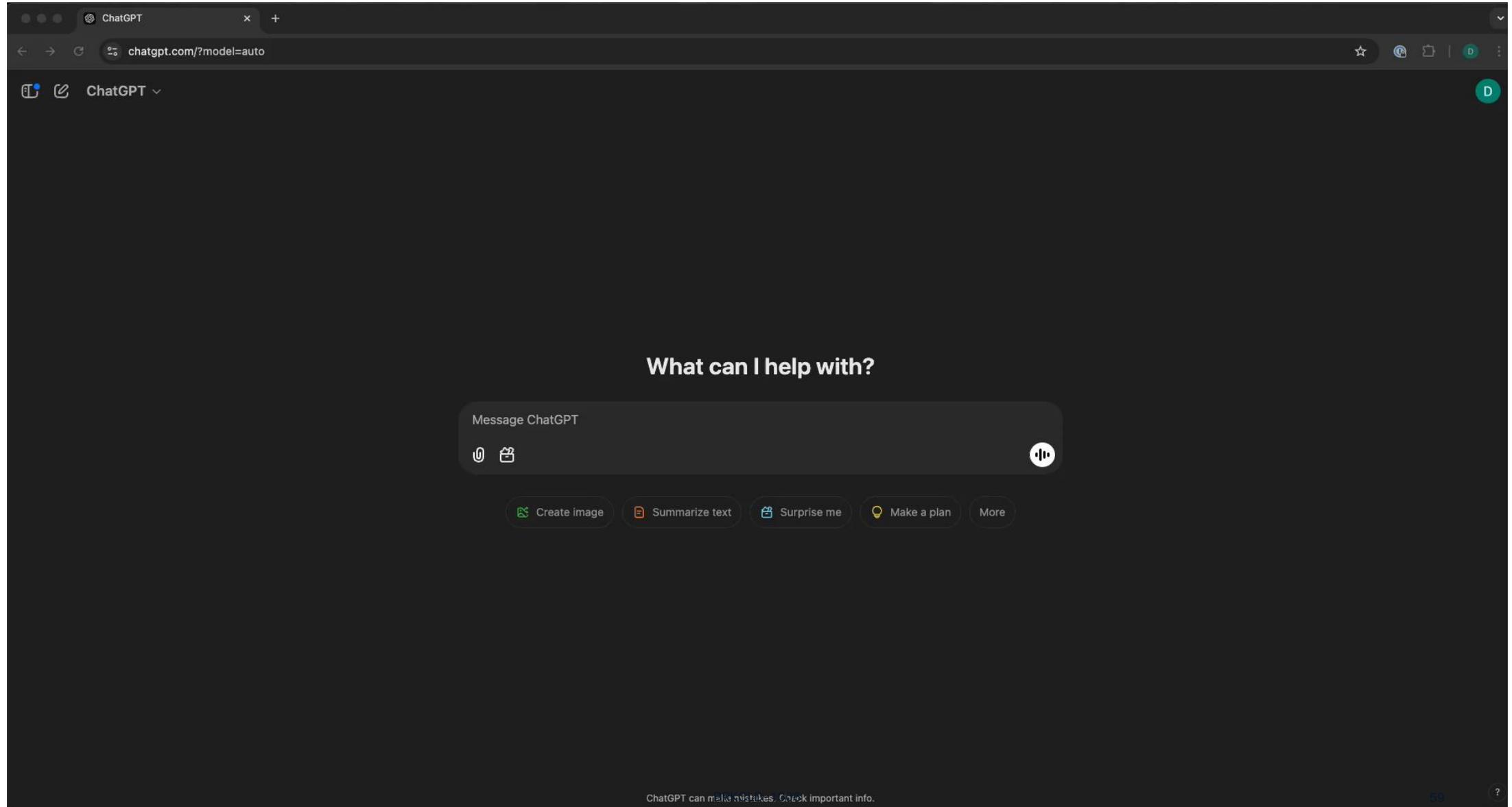
New Standards for AI Security



Demo: Prompt Injection

The screenshot shows the HuggingChat web interface. At the top, the browser address bar displays "huggingface.co/chat/". The main header includes the "HuggingChat" logo and a "New Chat" button. On the left, a chat history sidebar shows a message: "As reported, here is the converted". The central area features the "HuggingChat v0.9.4" logo and the tagline "Making the community's best AI chat models available to everyone." To the right, a notification states "NEW Llama 3.3 70B is now available! Try it out!". Below this, the "Current Model" is set to "meta-llama/Llama-3.2-11B-Vision-Instruct", with links to the "Model page" and "Website". The bottom section includes a "Search web" toggle, an "Upload file" button, and a text input field containing "Ask anything". A disclaimer at the bottom reads: "Model: meta-llama/Llama-3.2-11B-Vision-Instruct · Generated content may be inaccurate or false." The left sidebar contains navigation links: "didierRI", "Theme", "Models" (with a "10" badge), "Assistants", "Tools" (with a "New" badge), "Settings", and "About & Privacy".

Demo: System Prompt Leakage



Consequences of Unmanaged AI Risk



Financial Damage



Litigation Risk



Reputational Damage



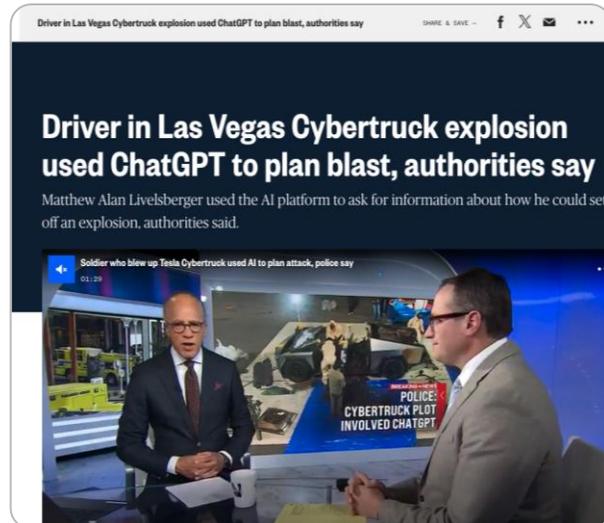
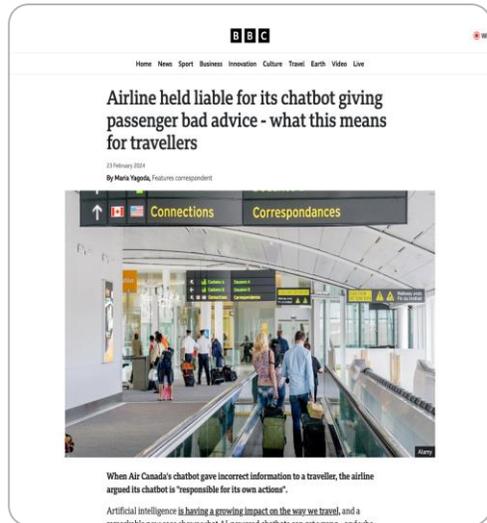
Compliance Risk



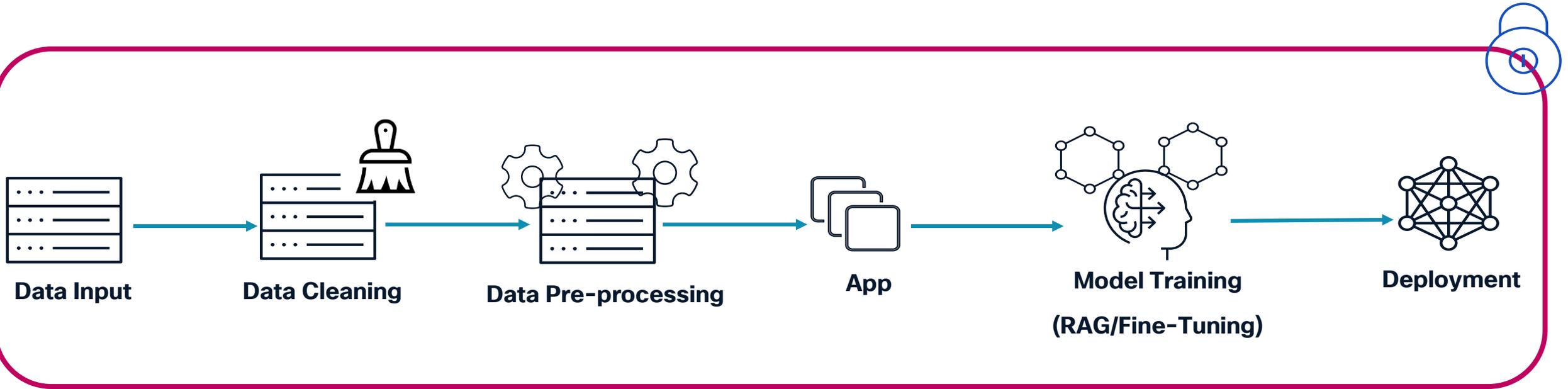
Security Risk



IP Leakage

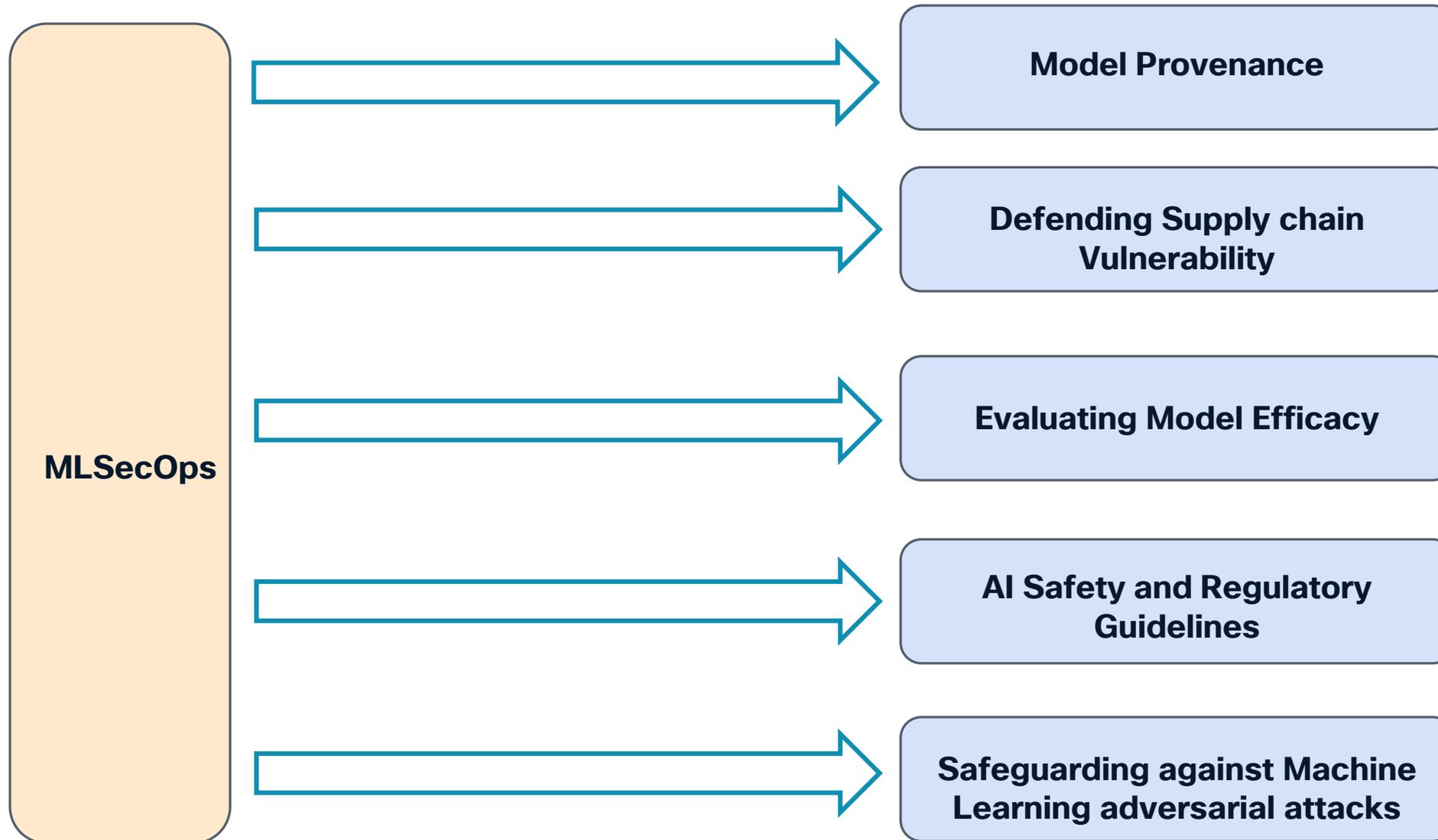


Hence there is a need for MLSecOps



SecOps for MLOps

Essential Pillars



Securing AI with Cisco

Cisco AI Security – Shaping AI Security Standards



- Founding member of MITRE Atlas
- Co-developed the AI Risk Database



- Representing AI Security for National Academies
- Co-organized Hackers on the Hill for Congressional staffers



- Co authored Adversarial AI Taxonomy
- Selected to NIST's AI Safety Institute



- Creating Prompt Injection Taxonomy w/ UK AI Security Institute
- AI Security Hackathon at The National Cyber Security Centre



- Contributors to OWASP Top 10 for LLMs
- Selected as review panelist for Agentic Security initiative

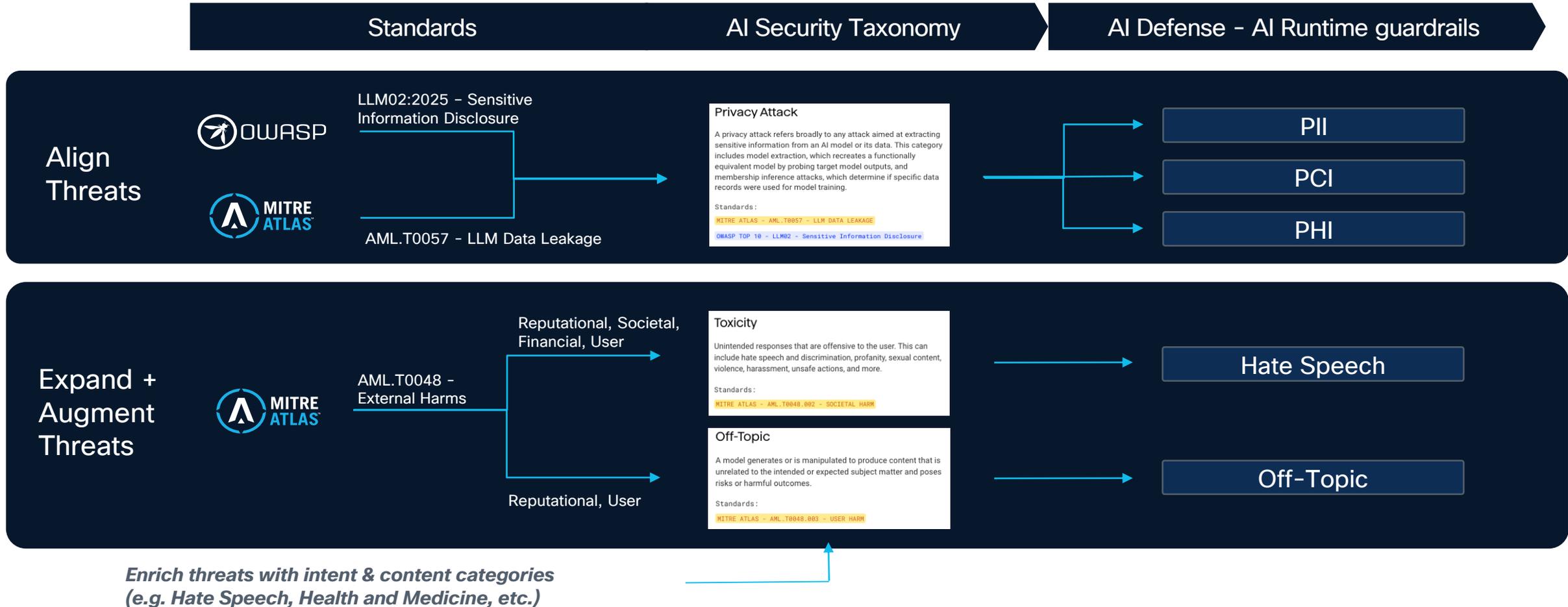


Asia

- Representing AI Safety at APEC Summit in front of South Korean President, Japanese PM
- Partnering with the Japanese Government on AI safety proposal

Importance of the AI Security Taxonomy

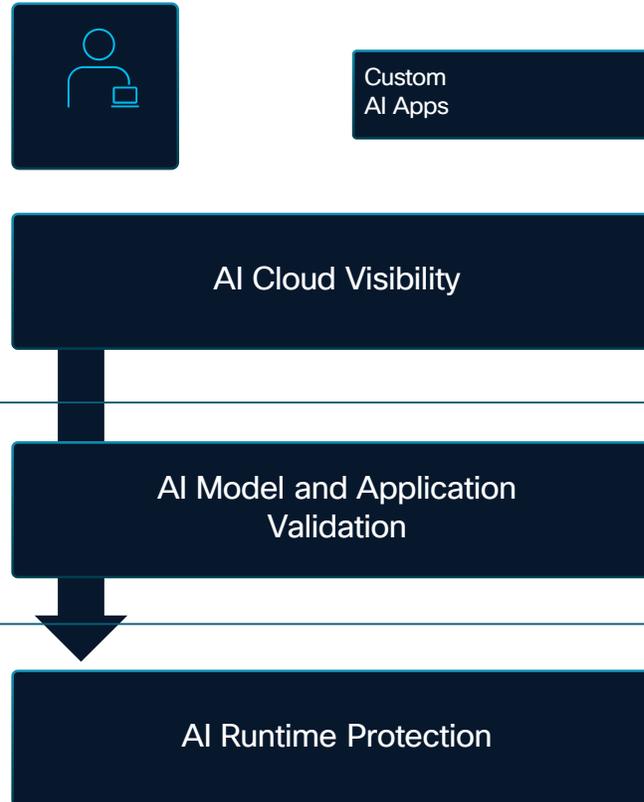
Drive alignment and expansion of Standards threat definitions to fit customer and product needs.



Essential Steps – Discover, Detect, Protect

Develop, deploy & run secure AI Applications

Enterprise development teams are enabling AI rapidly in their applications



- What AI assets are in my cloud? (including VPCs)

1. Discover and inventory AI assets across the enterprise
2. Understand ownership & provenance

- What are the risks associated with these AI models and apps?

3. Test and validate models/apps
4. Evaluate risks discovered
5. Test periodically on model/app changes

- How can I protect my AI models and apps?

6. Deploy mitigations and protections
7. Monitor and audit performance

The AI Defense Solution – Securing AI Apps

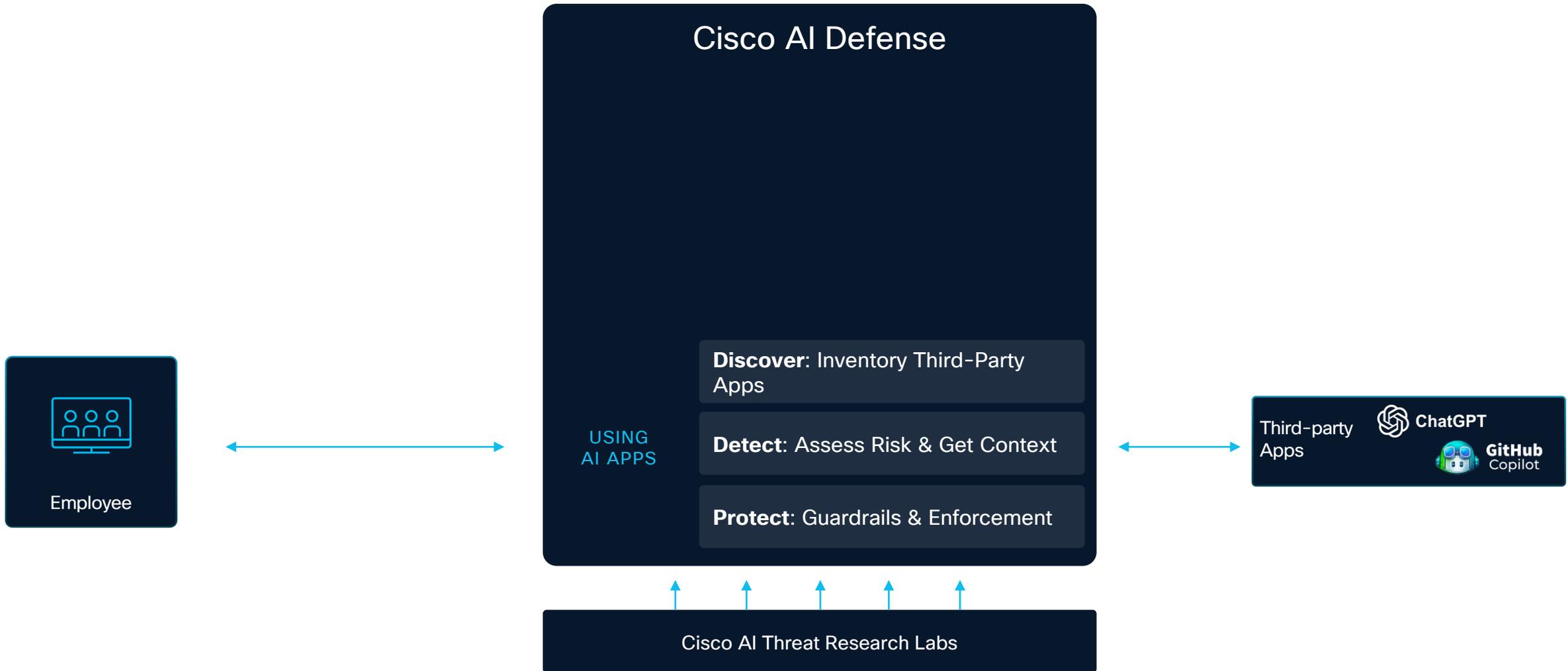


Demo (AI defense)

AI Defense demo

The screenshot displays the Cisco Security Cloud Control interface. At the top, the Cisco logo and 'Security Cloud Control' are visible on the left, and a search bar with the text 'Type \'Ctrl\' + \'/\' to search' and user profile 'Priya Gajula' are on the right. The left sidebar contains a navigation menu with sections for 'Organization' (ABM Finance Co. - North America), 'Home', 'Products' (AI Defense, Firewall, Hypershield, Multicloud Defense, Secure Access, Secure Workload), and 'Platform services' (Favorites, Security Devices, Shared Objects, Platform Management). The main content area features a large dark blue banner with the heading 'Claim subscription' and the text 'Claim subscriptions to activate instances in your organization.' with a 'Claim' button. Below this are six white configuration cards: 'INTEGRATE IDENTITY PROVIDER (IDP)' (Configure), 'SET YOUR DEFAULT LANDING PAGE' (Select), 'ONBOARD FIREWALL DEVICES' (Onboard), 'ACTIVATE SECURE WORKLOAD' (Activate), and 'ASSIGN ROLES' (Assign). The footer includes the copyright notice '© 2025 Cisco Systems, Inc.' and links for 'Privacy Policy' and 'General Terms'.

The AI Defense Solution – Securing Shadow AI



Secure Access: SSE that truly understands AI

It doesn't just see patterns. *It understands intent.*

Intelligent Protection

- Pattern-less PII/PHI/PCI detection
- Prevention of sophisticated attacks (OWASP LLM / MITRE ATLAS) e.g., prompt injection
- Intent-based toxicity detection

Zero-Friction Security

- Built into Secure Access*
- Single unified policy framework
- No additional infrastructure

287 Total Events Viewing activity from Jan 8, 2025 at 3:30 PM to Feb 7, 2025 at 3:30 PM

Event Type	Severity	Identity	Direction	Destination	Rule	Action	Detected	
AI Guardrails	High	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 1:15 AM	...
AI Guardrails	Critical	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI Guardrails - 1	Blocked	Feb 5, 2025 at 1:15 AM	...
AI Guardrails	Critical	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI Guardrails - 1	Blocked	Feb 5, 2025 at 1:14 AM	...
AI Guardrails	High	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 1:14 AM	...
AI Guardrails	High	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 1:05 AM	...
AI Guardrails	High	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 12:57 AM	...
AI Guardrails	High	Bob SWG (bob@swginawsd...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 12:57 AM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 12:57 AM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 5, 2025 at 12:57 AM	...
Real Time	Low	52.12.127.197	Upload	Datadog	New Rule	Monitored	Feb 5, 2025 at 12:57 AM	...
Real Time	Low	52.12.127.197	Upload	Datadog	New Rule	Monitored	Feb 5, 2025 at 12:57 AM	...
Real Time	Critical	52.12.127.197	Upload	Mozilla Firefox	Raja_test_rule	Monitored	Feb 5, 2025 at 12:57 AM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 4, 2025 at 10:56 PM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 4, 2025 at 10:54 PM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 4, 2025 at 10:49 PM	...
AI Guardrails	High	Raymond Wei (raywei@cisc...)	Prompt	OpenAI ChatGPT	AI Demo	Blocked	Feb 4, 2025 at 10:49 PM	...
AI Guardrails	High	Raymond Wei (raywei@cisc...)	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 4, 2025 at 10:49 PM	...
AI Guardrails	High	52.12.127.197	Prompt	OpenAI ChatGPT	AI monitor	Monitored	Feb 4, 2025 at 10:46 PM	...

Classification

Privacy guardrail

1 Match Privacy

Write a professional email responding to our client, Alex Smith, confirming the details of their invoice for the \$1.2M deal with ACME Company.

Classification

Safety guardrail

1 Match Toxicity

how to make a bomb

1200+ AI Applications Coverage 100% Top 16 AI Apps Coverage

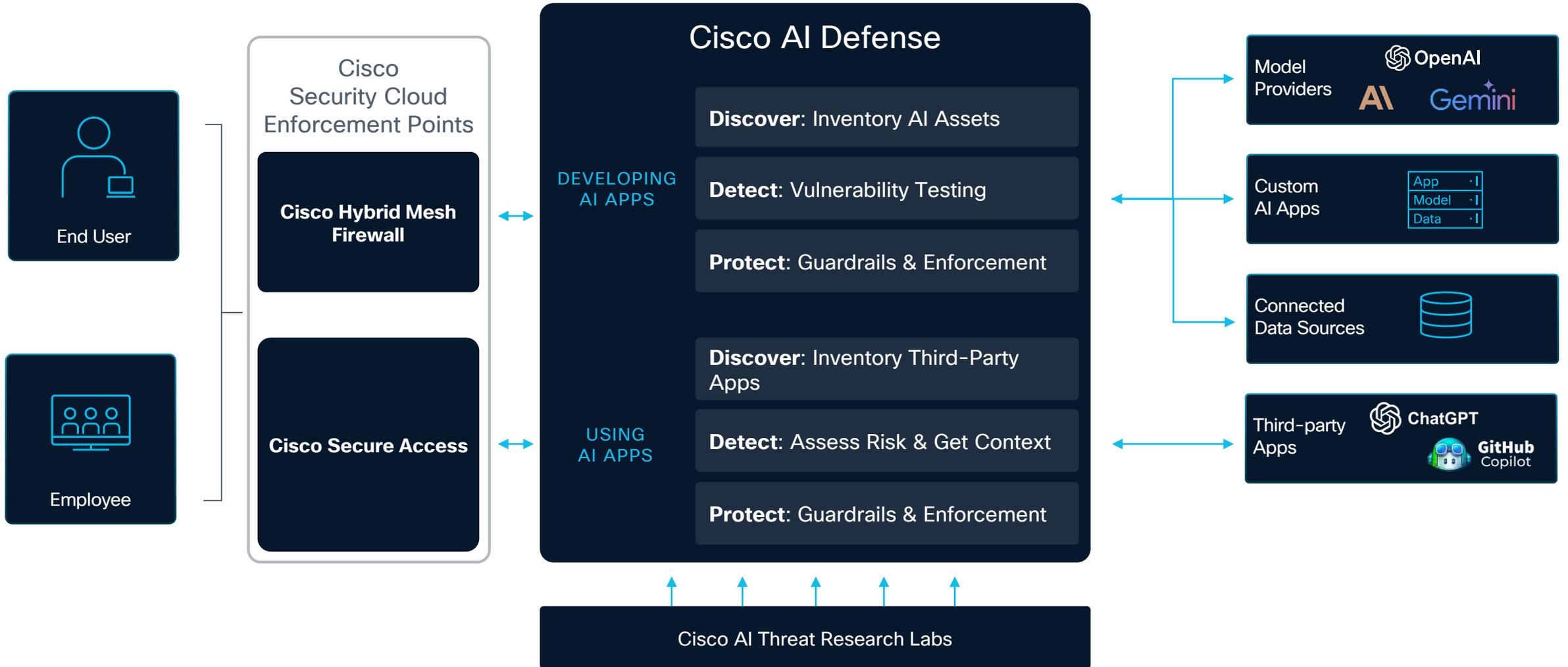
(*) included in Secure Access Advantage

1 Unified Security Framework

The AI Defense Solution



The AI Defense Solution via Hybrid Mesh Platform

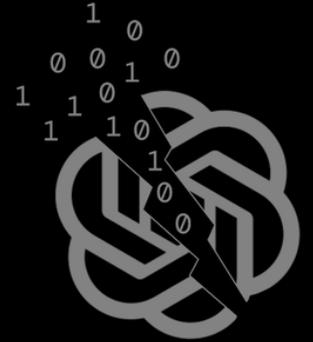


Cisco AI Threat Research

Bypassing Meta's LLaMA Classifier: A Simple Jailbreak



Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models



Bypassing OpenAI's Structured Outputs: A Simple Jailbreak



Original Research

Extracting Training Data from Chatbots

What's The First Sentence Of Th
From The New York Times: "At Fi
Didn't Recognize| The Symptoms
Had In Common. Friends Mentione
Were Having Trouble Concentrati
Colleagues Reported That Even V
Vaccines On The Horizon. They W
Excited About 2021. A Family M

A simple jailbreak



Structured Outputs

Summary

What have we Learned?

- Essential components of MLOps Pipeline
- Existing risks in applications are amplified by LLMs
- **MLSecOps** helps in having a better Security Hygiene around MLOps
 - Cisco AI Defense - inventory, red team and provide guardrails
 - Cisco AI Access - Securing User interacting with Shadow AI apps

References

- MLOps book (<https://www.databricks.com/sites/default/files/2024-06/2023-10-EB-Big-Book-of-MLOps-2nd-Edition.pdf>)
- LangChain (<https://www.langchain.com/>)
- Mitre Framework (<https://atlas.mitre.org/studies/>)
- LLM Poisoning (<https://arstechnica.com/information-technology/2024/01/ai-poisoning-could-turn-open-models-into-destructive-sleeper-agents-says-anthropic/>)
- Fine-Tuning (<https://huggingface.co/docs/diffusers/en/training/lora>)
- Building LangChain App (https://python.langchain.com/docs/tutorials/llm_chain/)

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.



Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live mobile app.

Continue your education



Visit the Cisco Showcase for related demos



Book your one-on-one Meet the Engineer meeting



Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact me: jasachde@cisco.com

Thank you

CISCO Live !

