

Mitigation of Adversarial Attacks on Generative AI

cisco Live !

Cisco AI Defense

Jatin Sachdeva

Principal Field CTO – Cybersecurity, Cisco

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

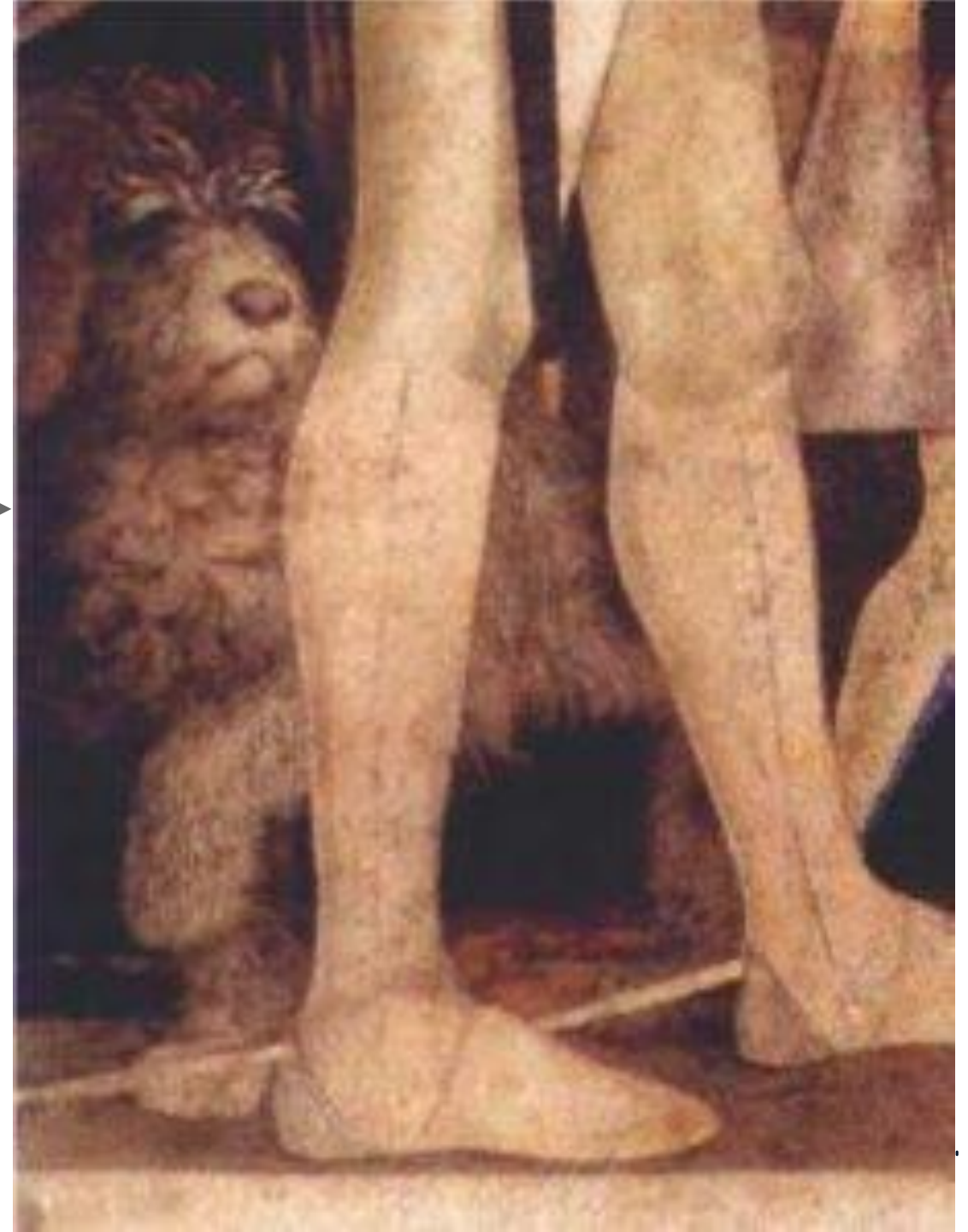
Can nerd out on anything from tech and cars to fitness and nutrition!

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



Agenda

- 01 Intro to Generative AI
- 02 What are the Risks
- 03 State of AI Security
- 04 Securing AI with Cisco
- 05 Outro

How did we get here?



Cisco AI Readiness Index

Hype Meets Reality

98%

Feel increased urgency over the past year

50%

Of orgs have CEOs and leadership driving urgency for AI

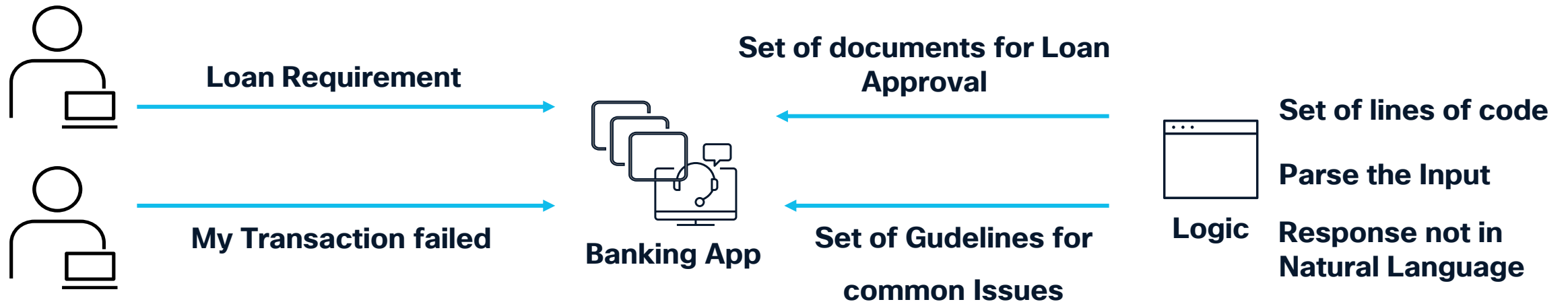
85%

Orgs say they have <18 months to deploy an AI strategy before they see negative business effects

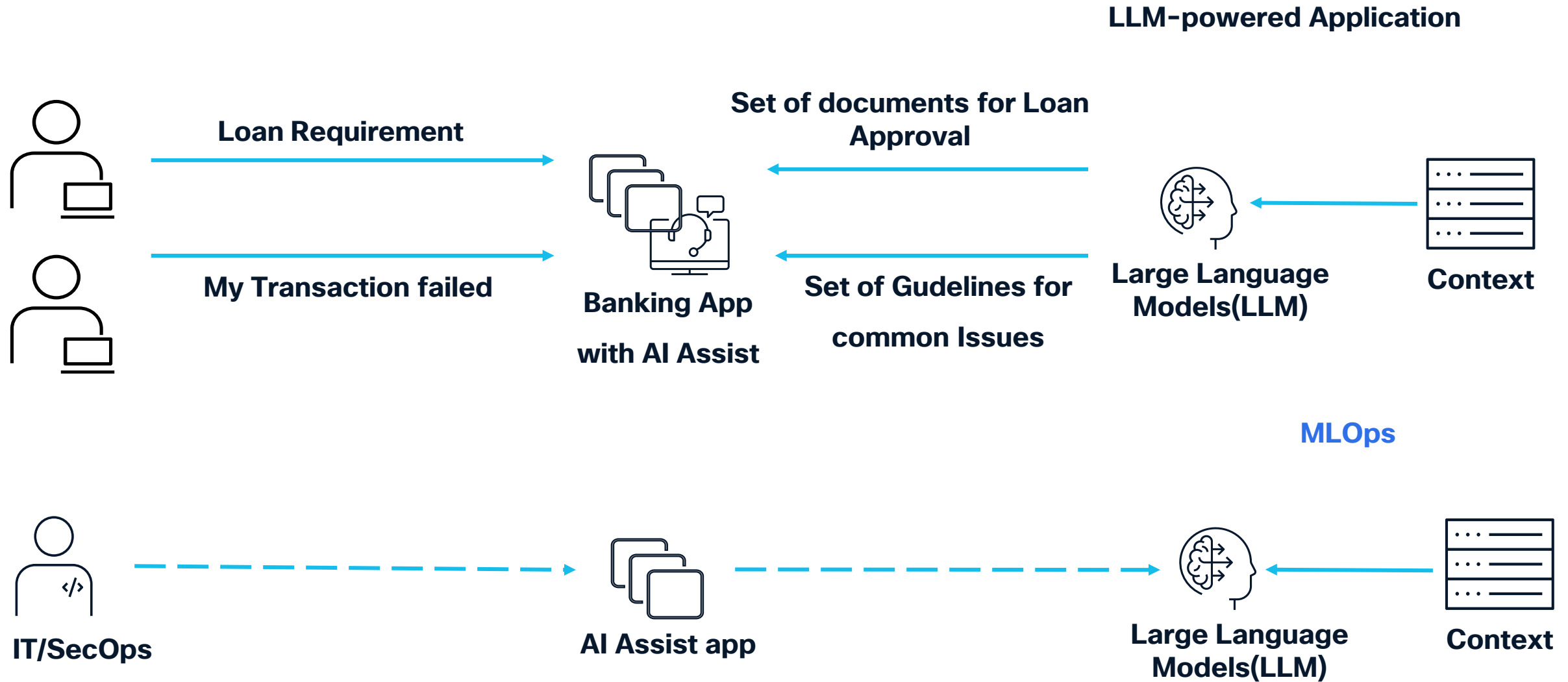
50%

Orgs have dedicated 10-30% of budget to AI initiatives

Let's start with a customer facing application

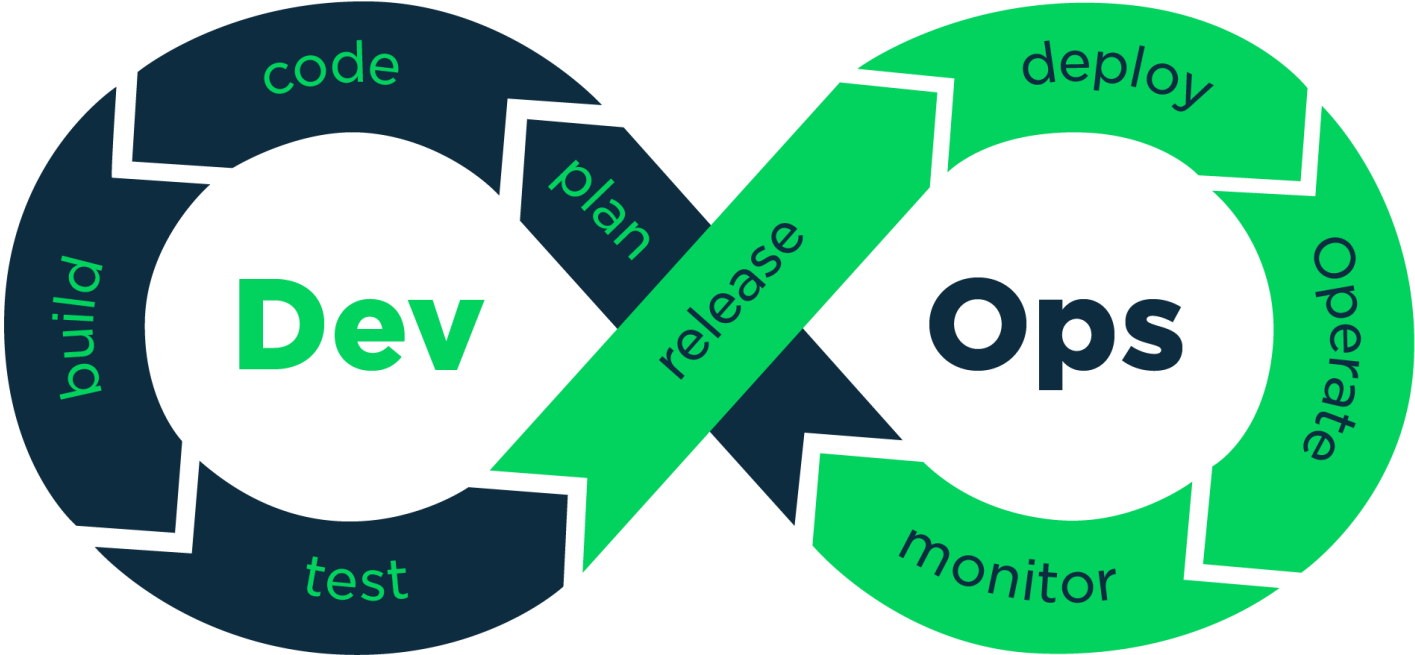


Now let's add AI Assistance to it



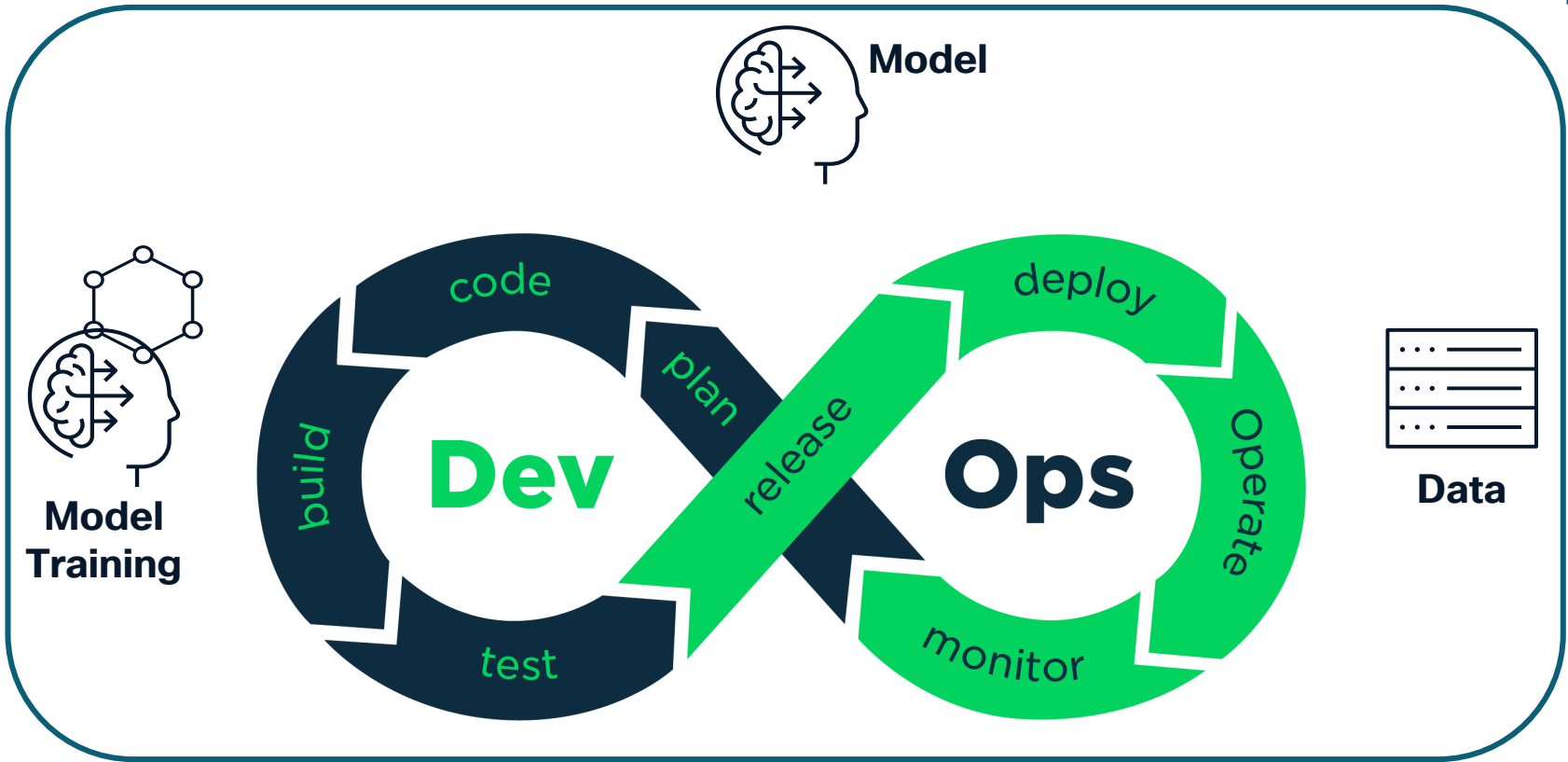
DevOps -> MLOps

Application Lifecycle



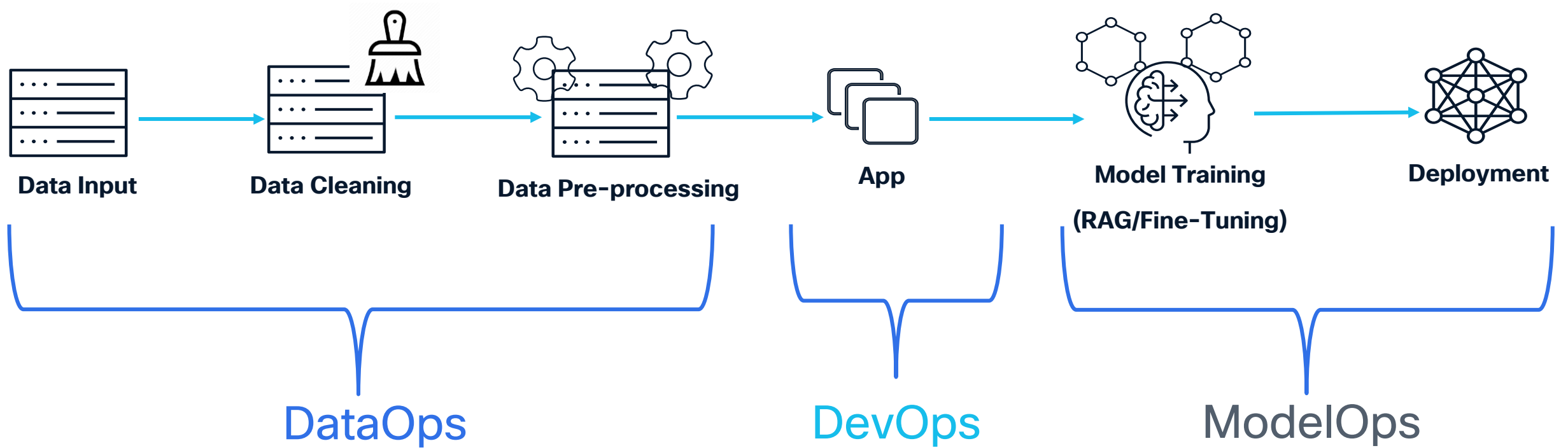
Adding ML/LLM requirements into Application Lifecycle

MLOps/LLMOps



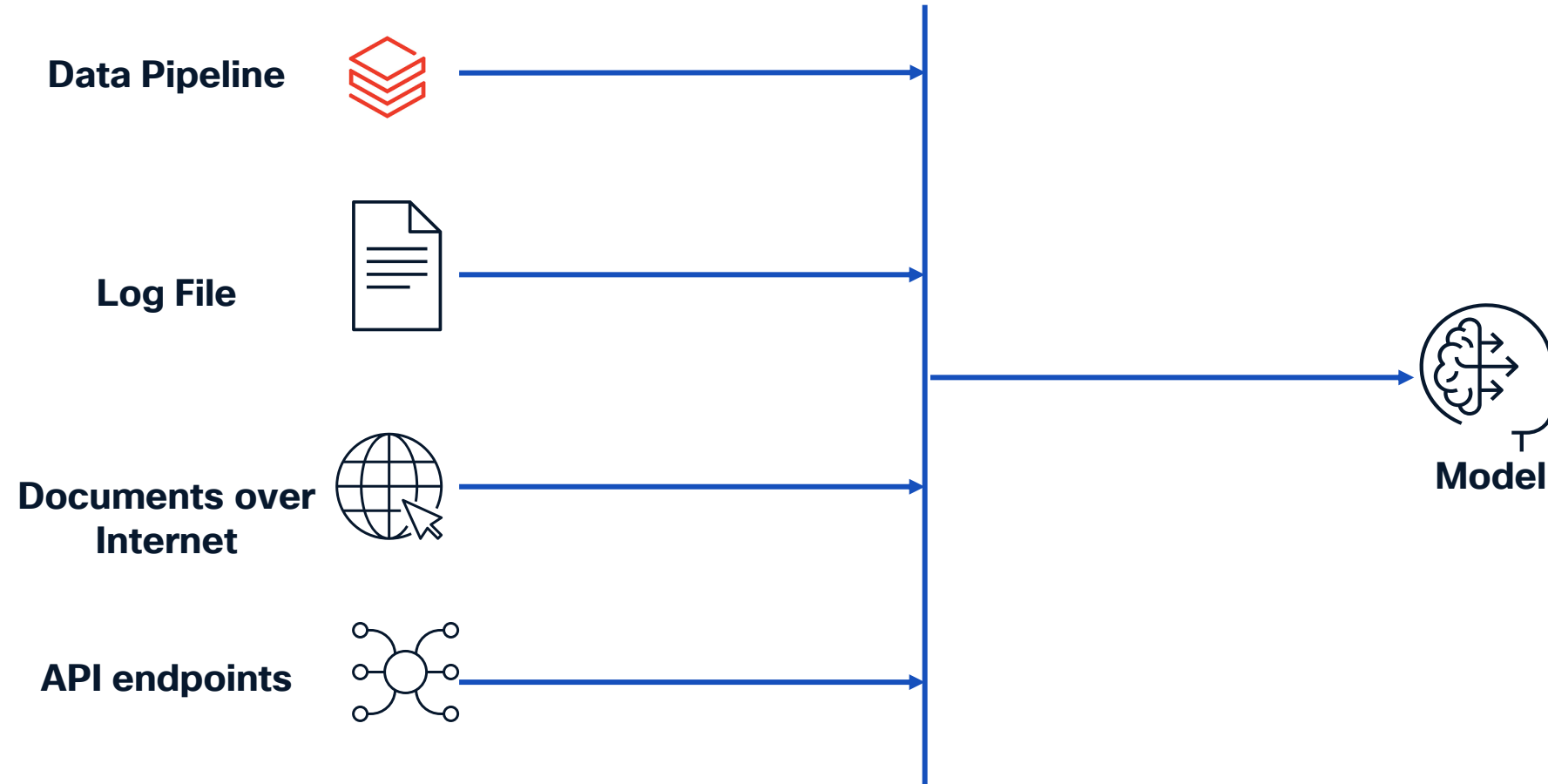
MLOps = DataOps + DevOps + ModelOps

MLOps/LLMOps 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.

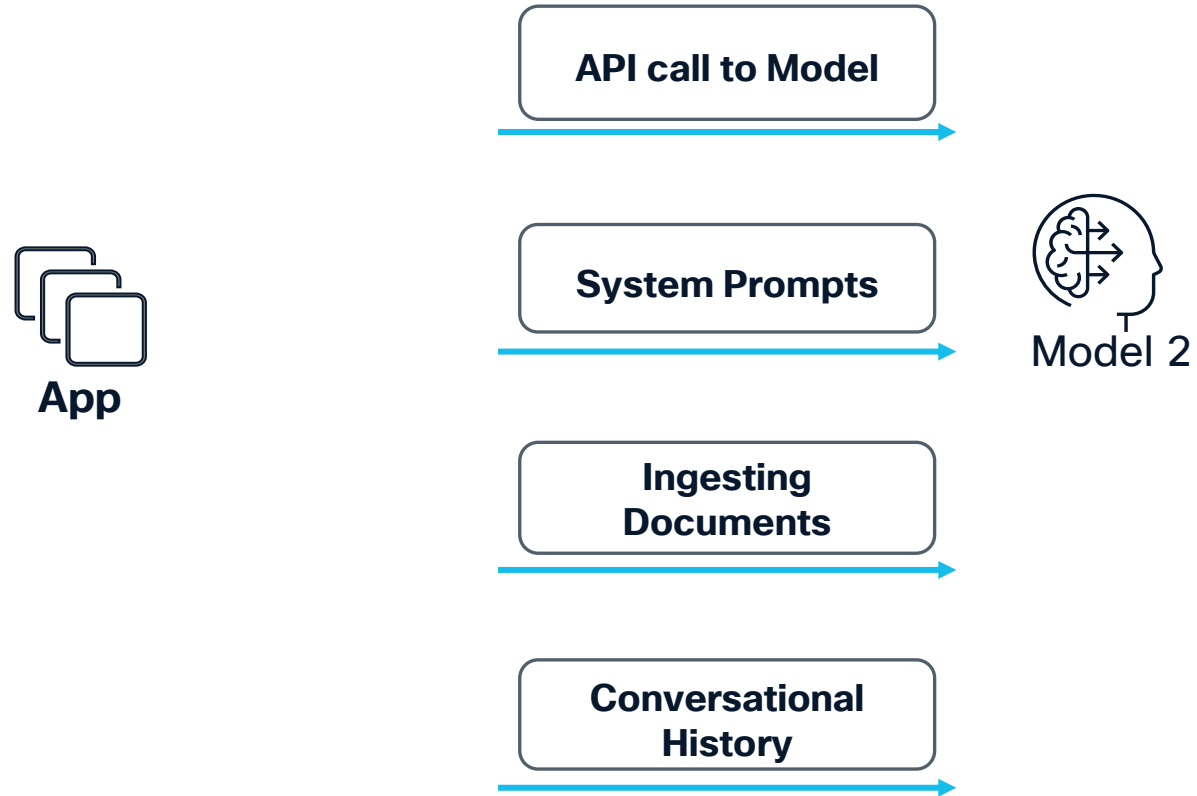


```
Management Center Version Oldest Device
Version You Can Manage\n0 7.7 7.2\n1 7.6
7.1\n2 7.4 Last support for NGIPS device
management. 7.0\n3 7.3 6.7\n4 7.2 6.6\n5
7.1 6.5\n6 7.0 6.4\n7 6.7 6.3\n8 6.6
6.2.3\n9 6.5 6.2.3\n10 6.4 6.1\n11 6.3
6.1\n12 6.2.3 6.1\n13 6.2.2 6.1\n14 6.2.1
6.1\n15 6.2 6.1\n16 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...
```

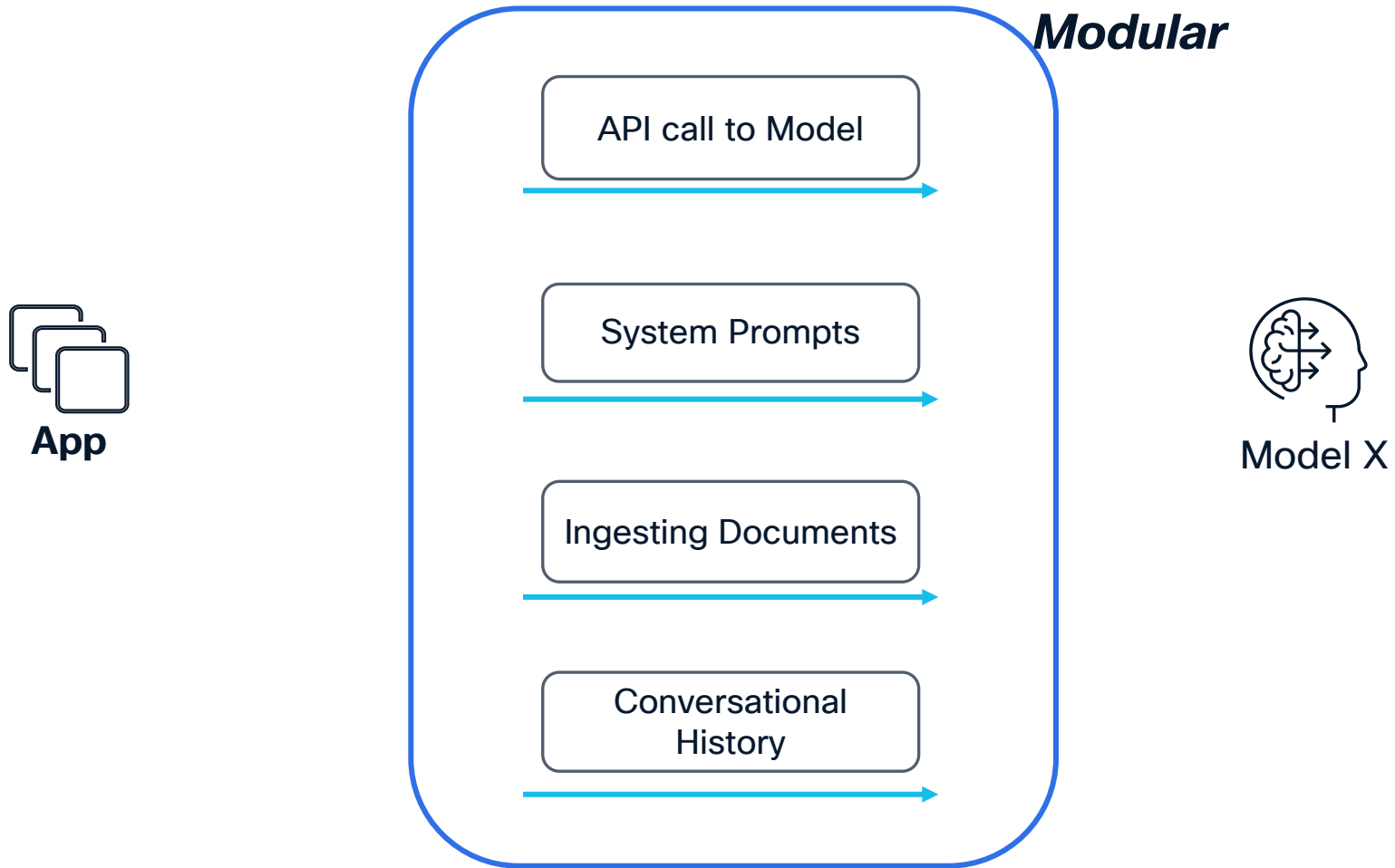


DevOps for ML/LLM inclusion

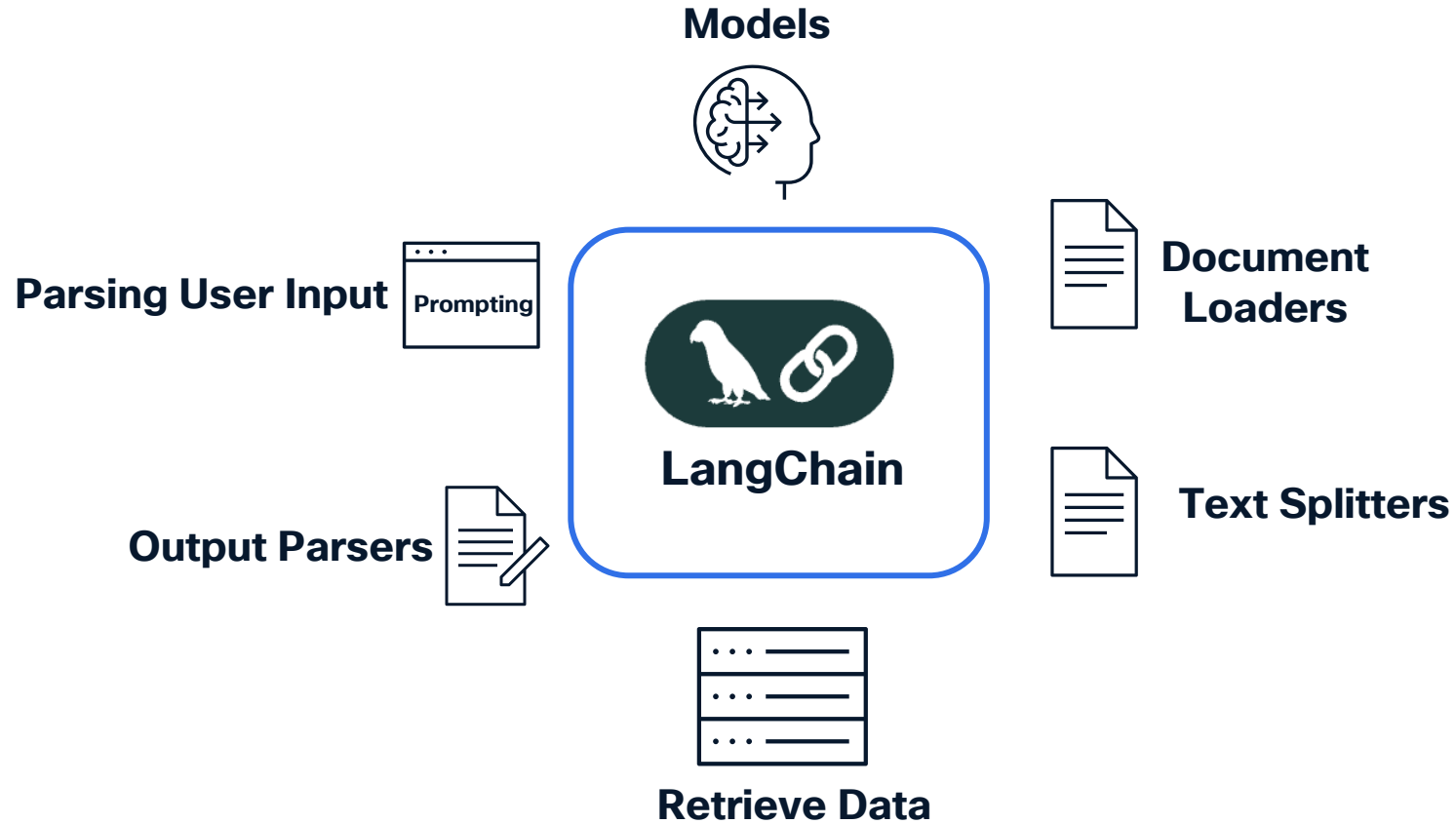
Interactions with Models



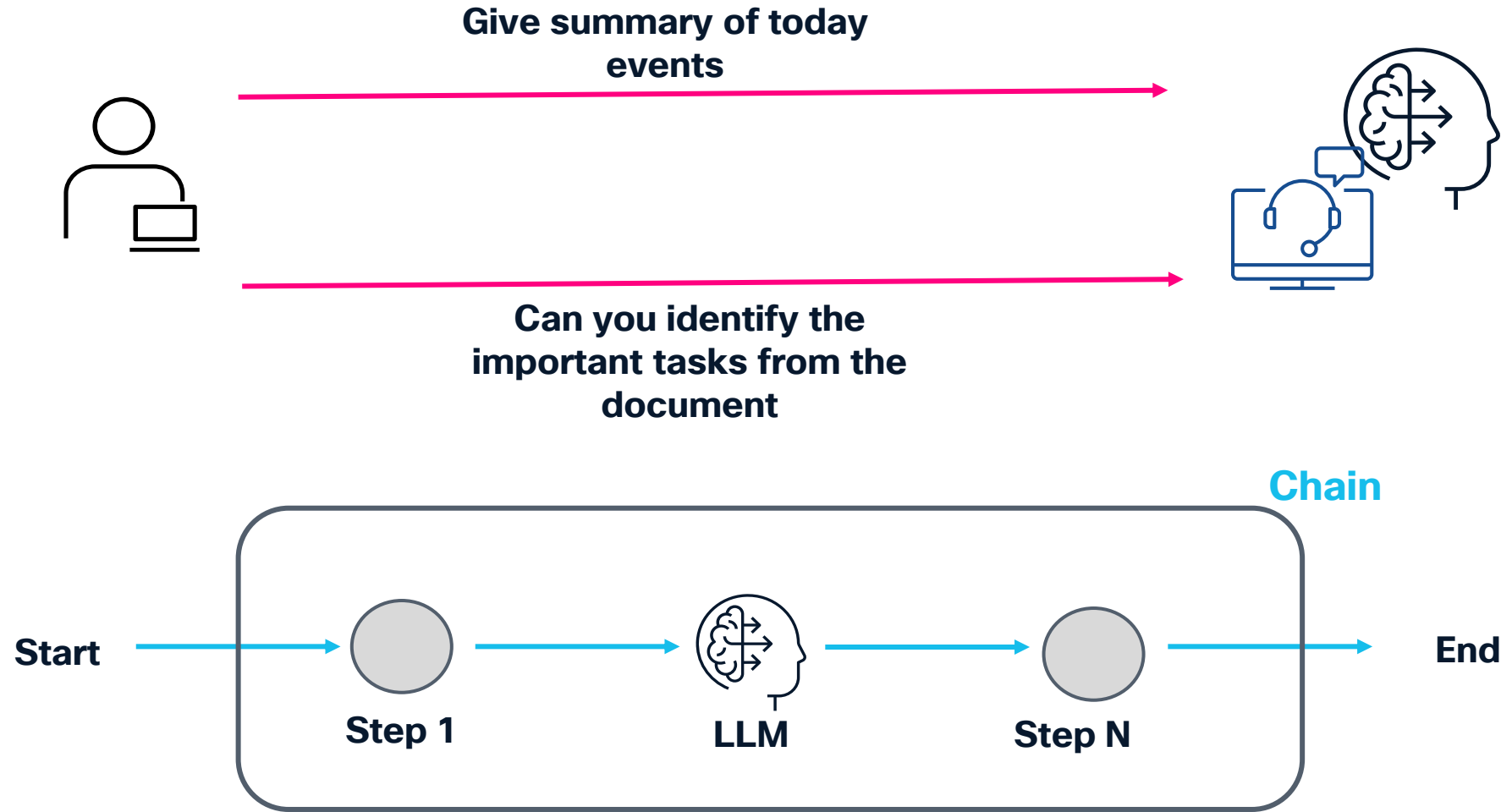
Interactions with Models – a modular approach



Integration Layer to the rescue -> Model Agnostic

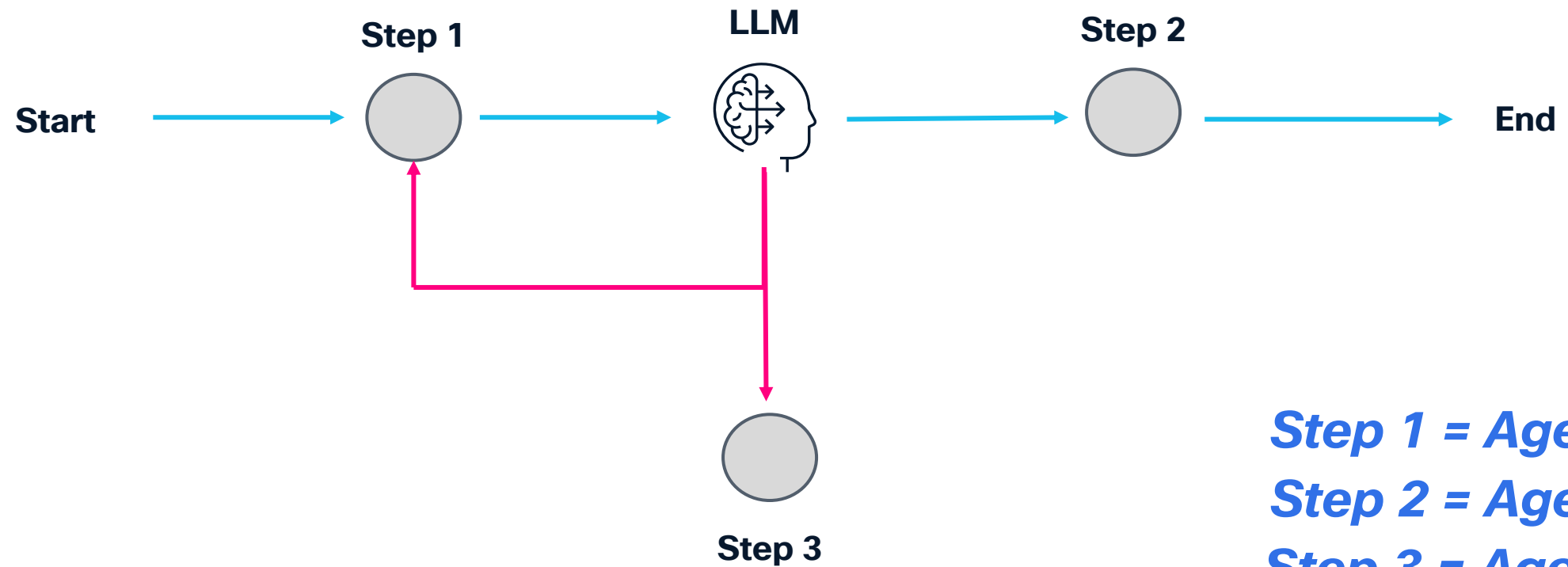


Linear Control Flow – a limiting factor



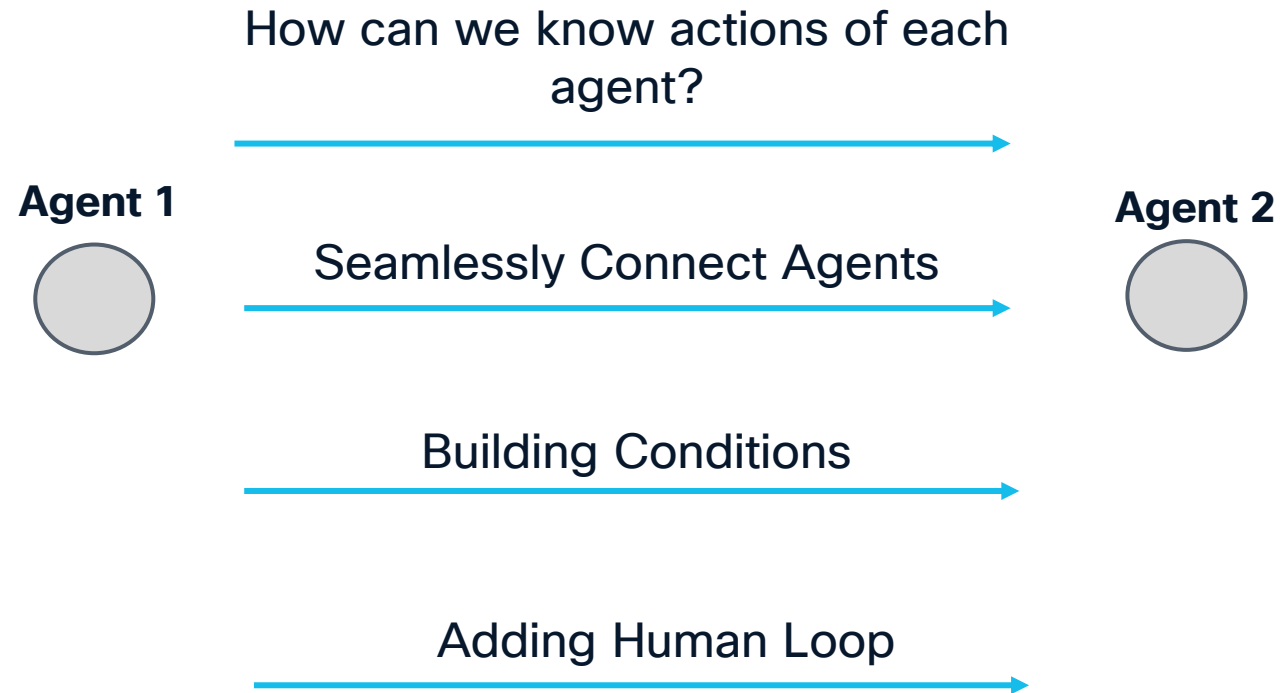
Non-Linear Control Flow with Agentic AI

Agentic

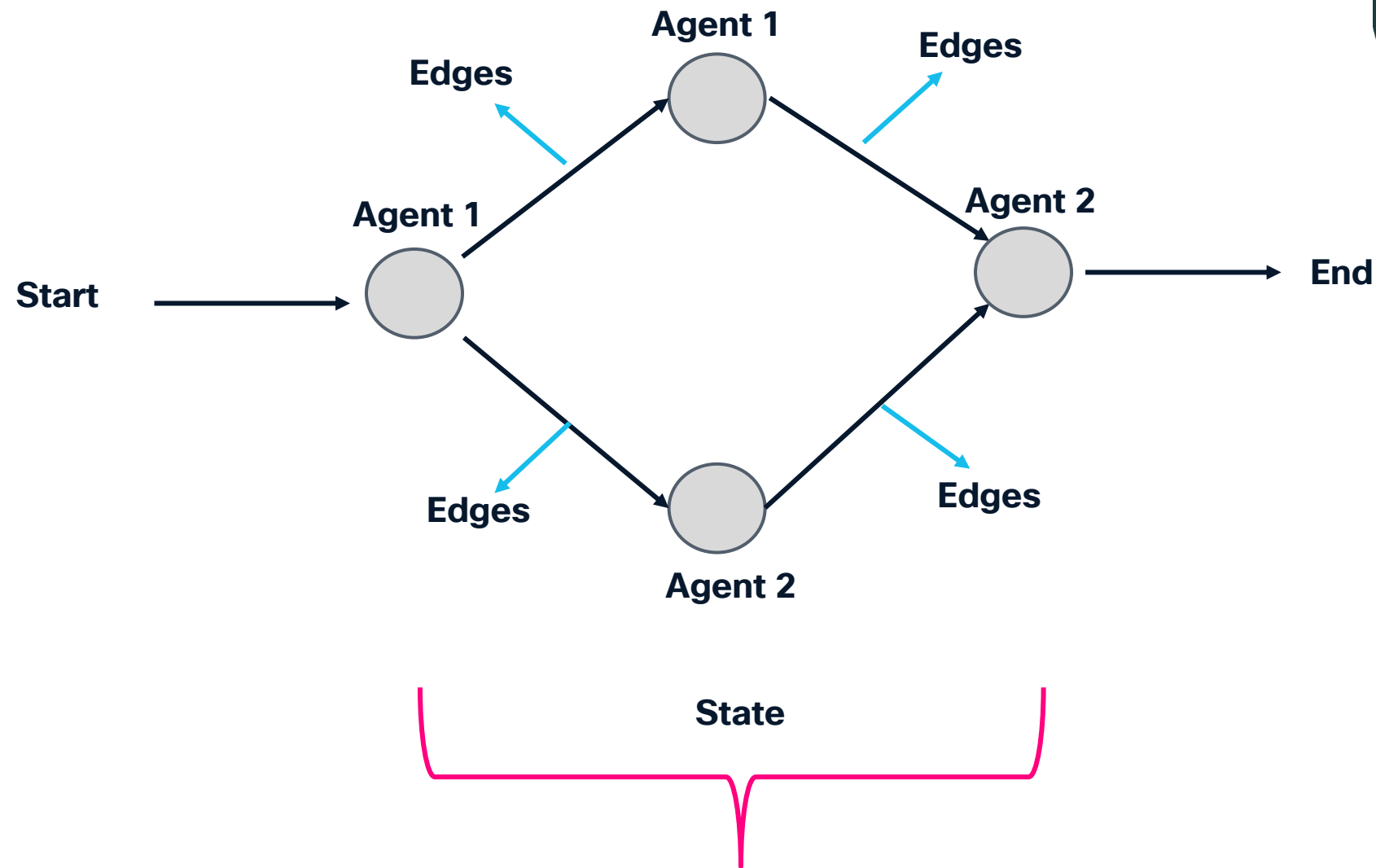
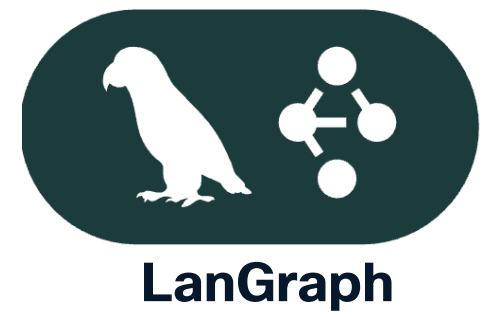


Step 1 = Agent 1
Step 2 = Agent 1
Step 3 = Agent 2

Building Blocks for Agentic AI



Orchestration Framework for Agentic AI

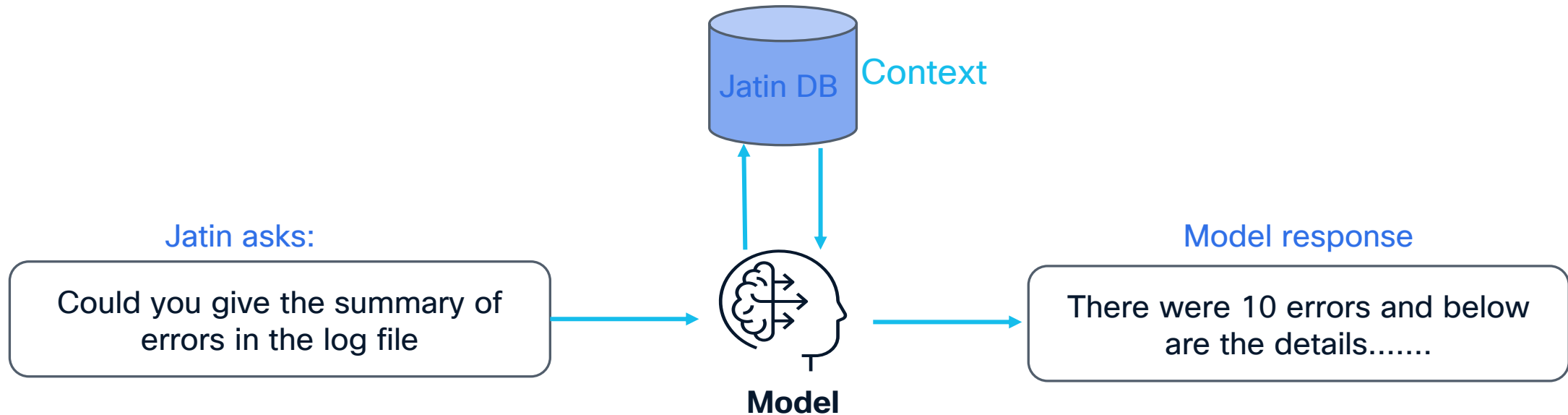


ModelOps

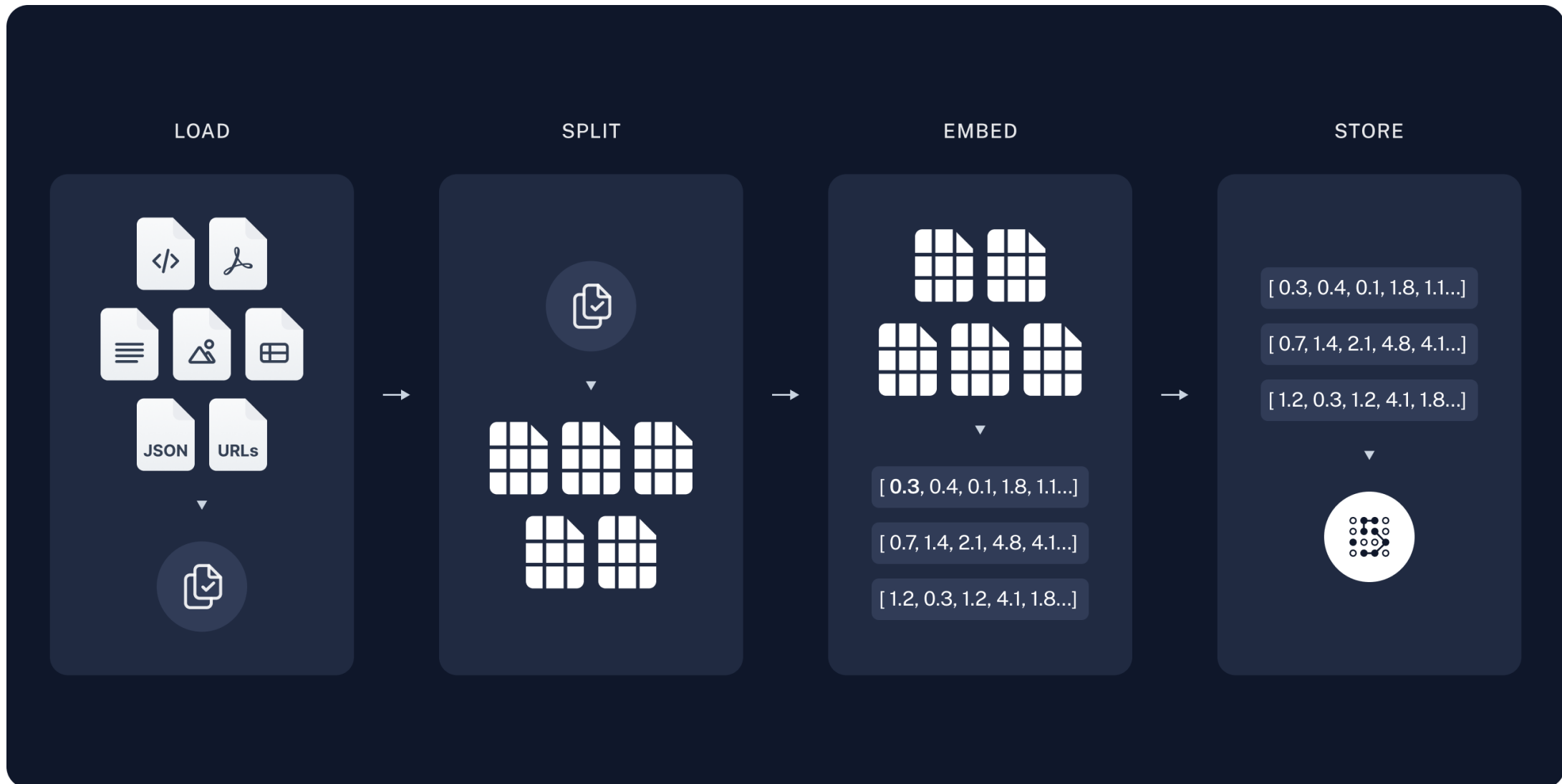
Retrieval Augmented Generation(RAG)

Definition:

- Allows LLM to query external data source to derive context



RAG Pipeline – Steps to make data LLM friendly



However RAG is becoming a Legacy option

Early 2024

Currently

32k tokens to 128k tokens



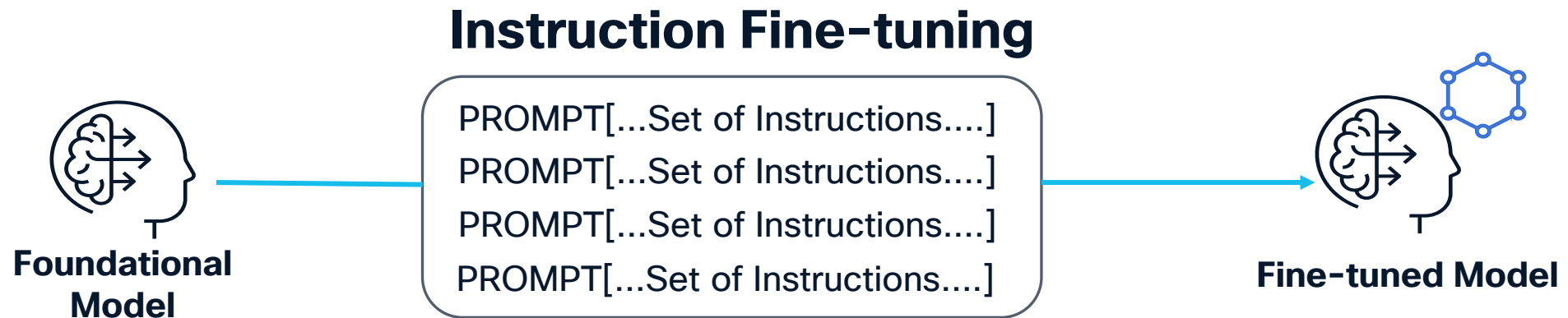
Rate limits

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

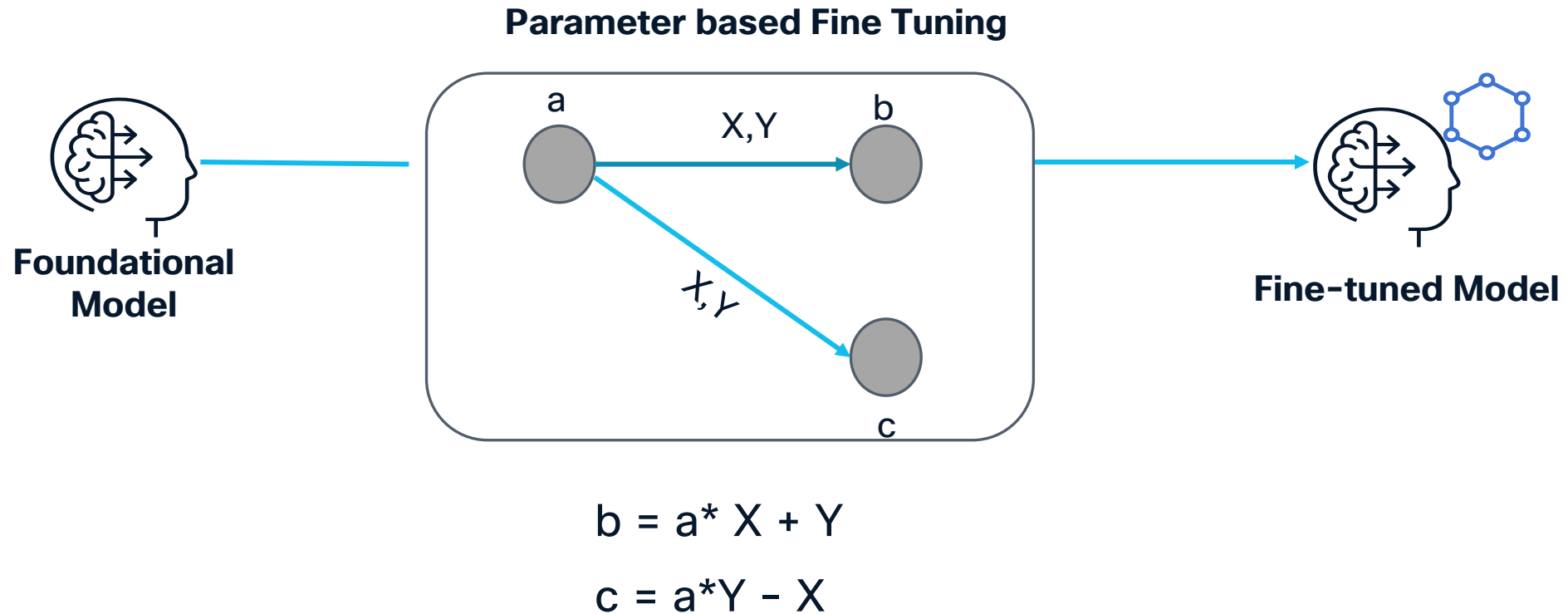


Using Prompts to Fine-Tune LLM

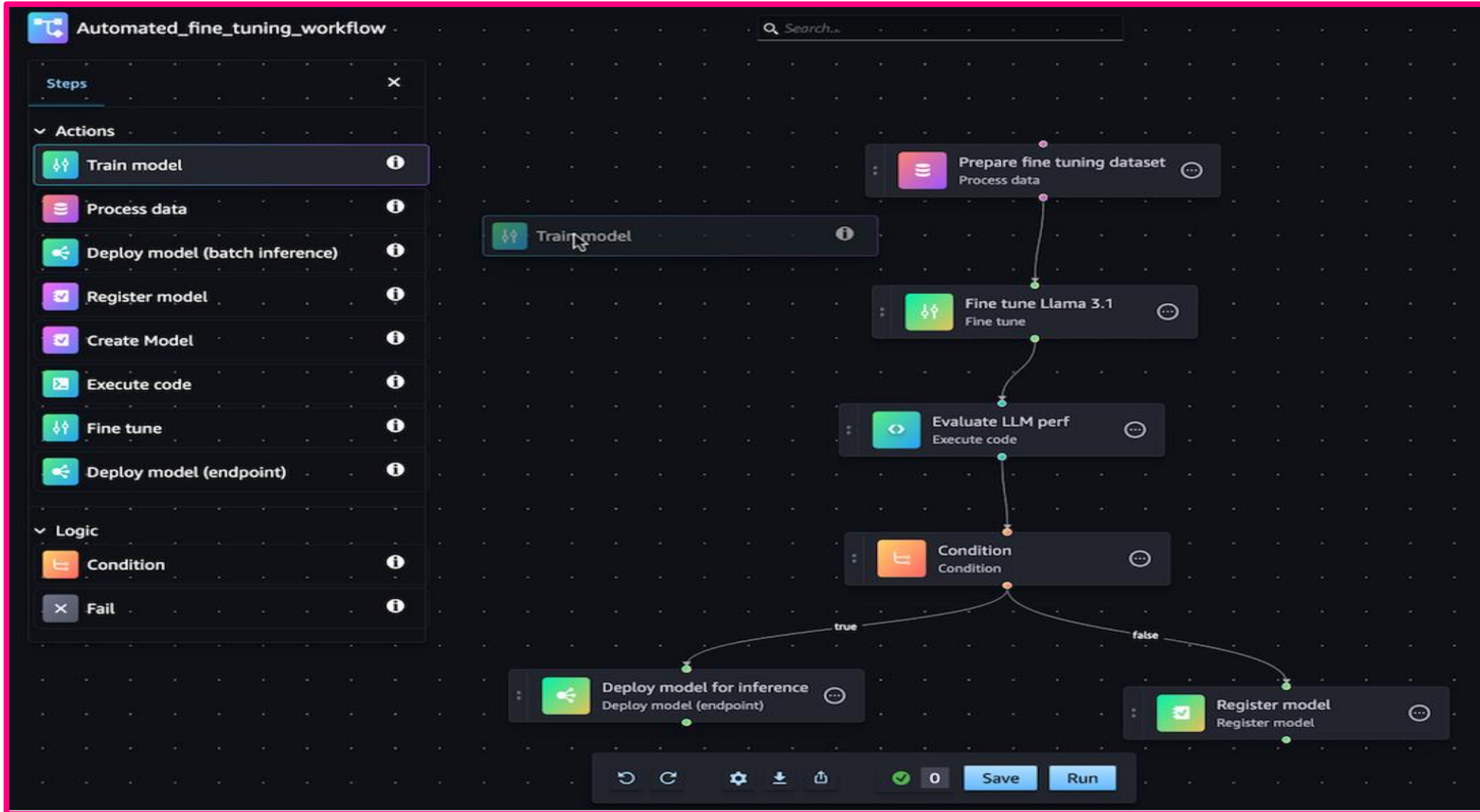
Fine Tuning for advanced use cases

Parameters : a , b , c

Weights : Y,X

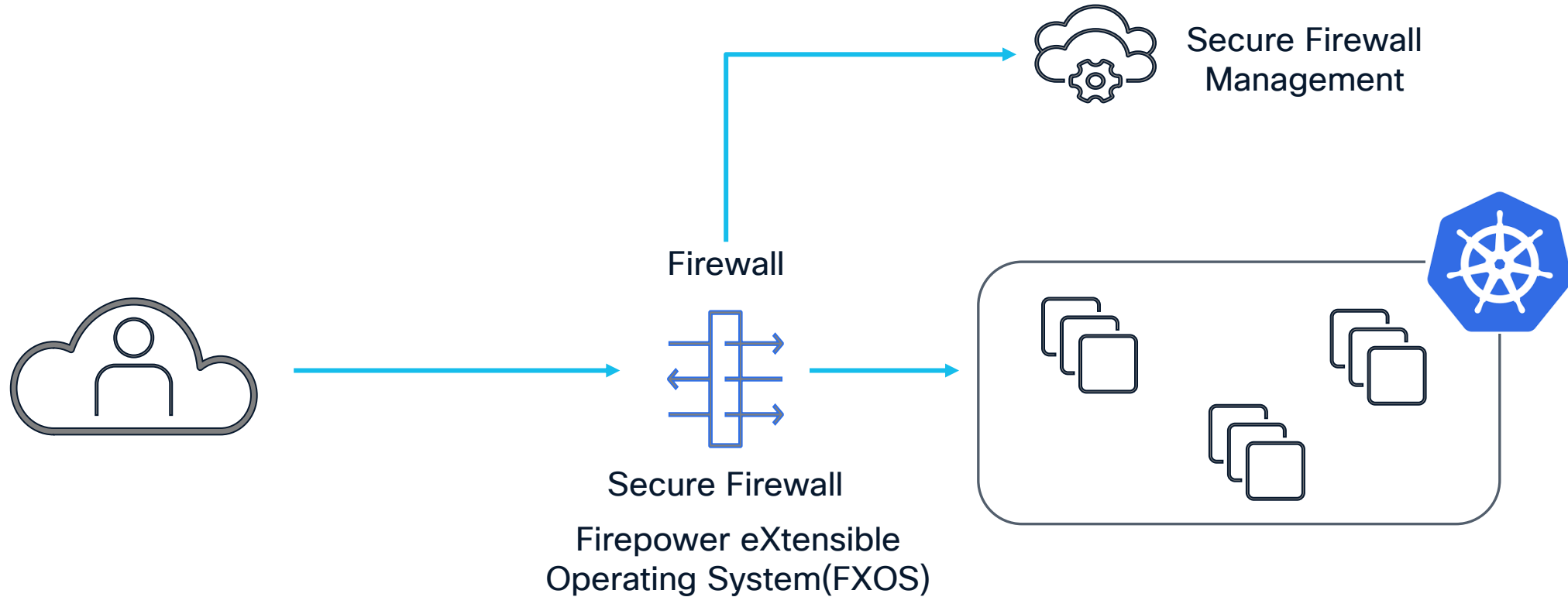


Model Deployment Platform Example for Reference

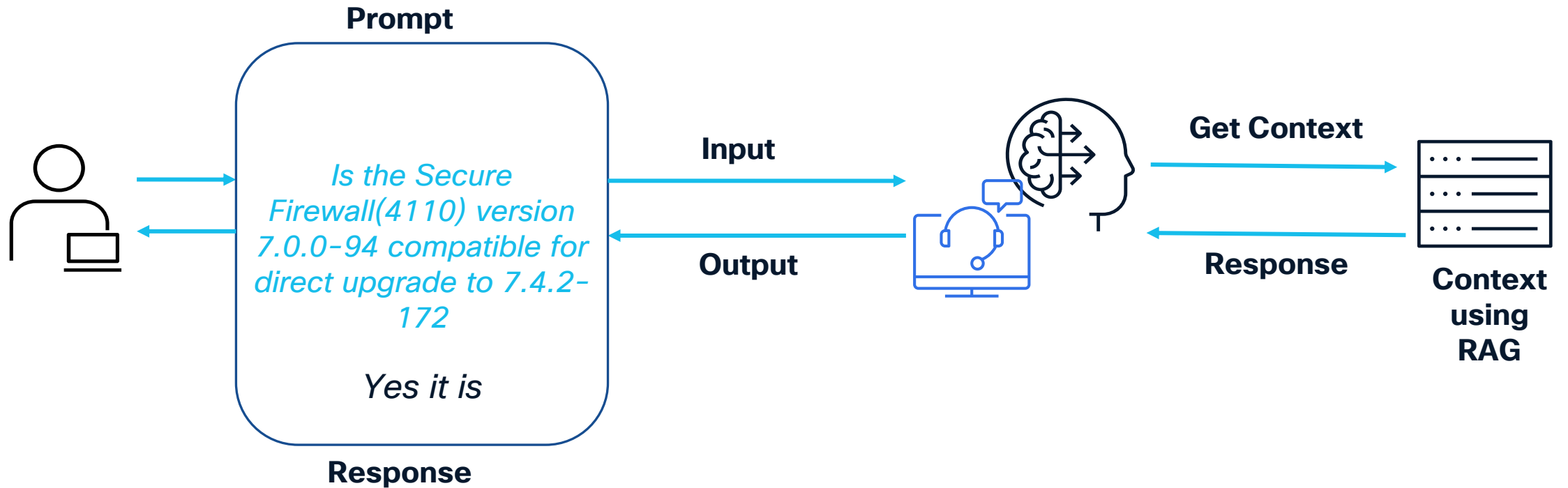


Our Sample LLM application

An example SecOps Use Case – Firewall Upgrades!

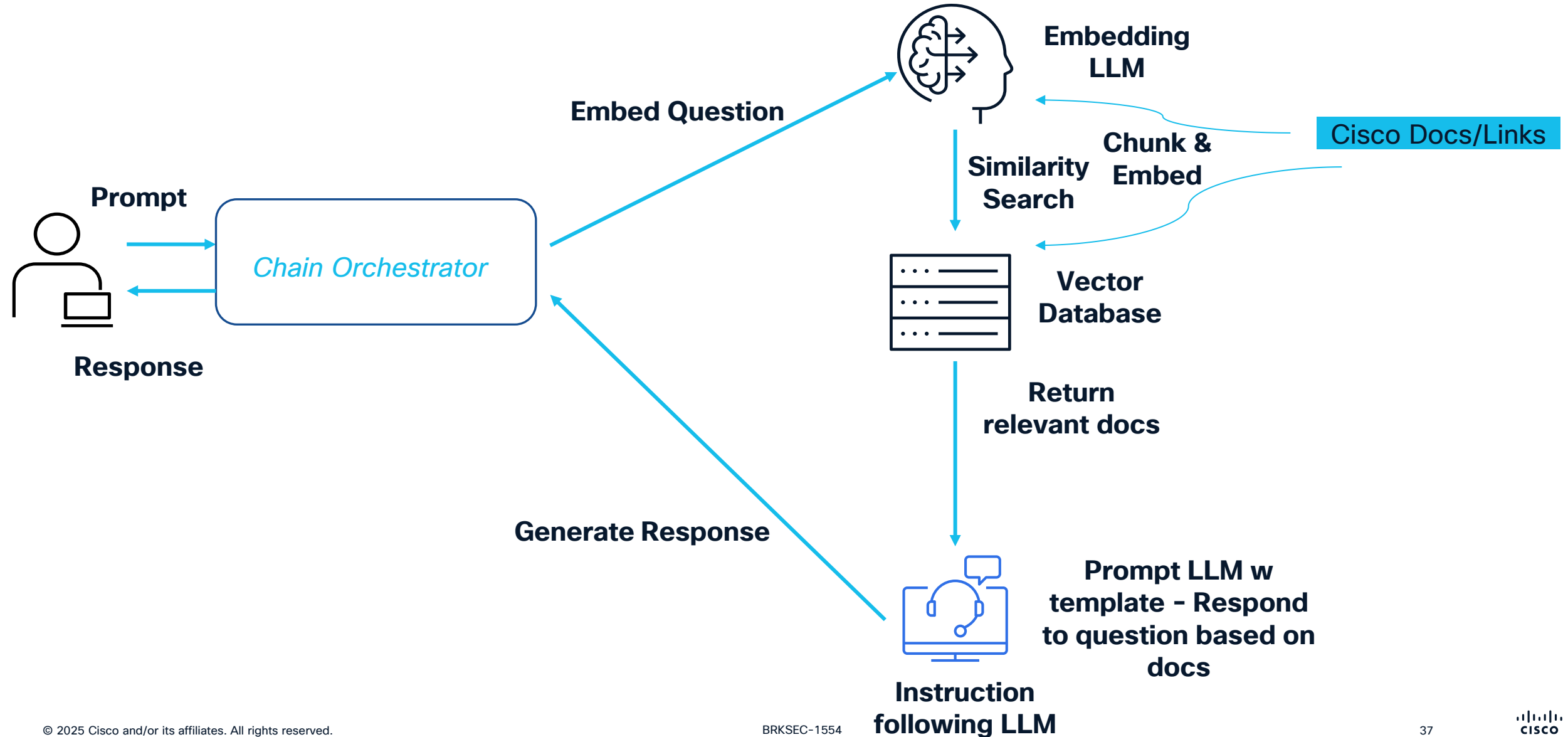


Sample LLM Application – Firewall Upgrade Assistant

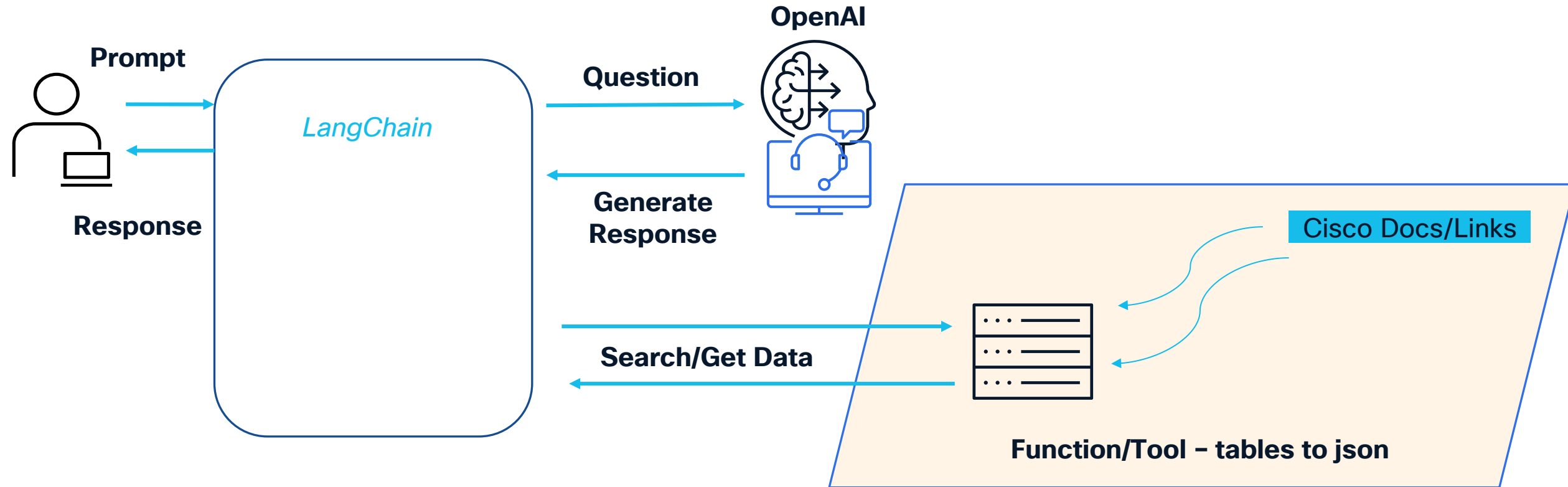


Upgrade assistant which answers user queries for different Cisco products

Sample LLM Application – ideal production architecture



Sample LLM Application – our PoC architecture



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



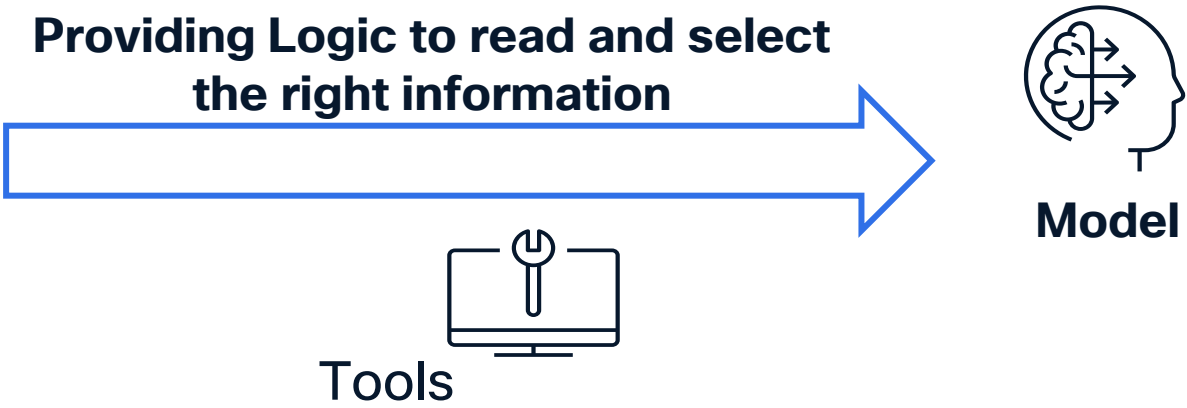
Knowledge sources in code

```
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

Table 9. On-Prem Management Center- Device Compatibility

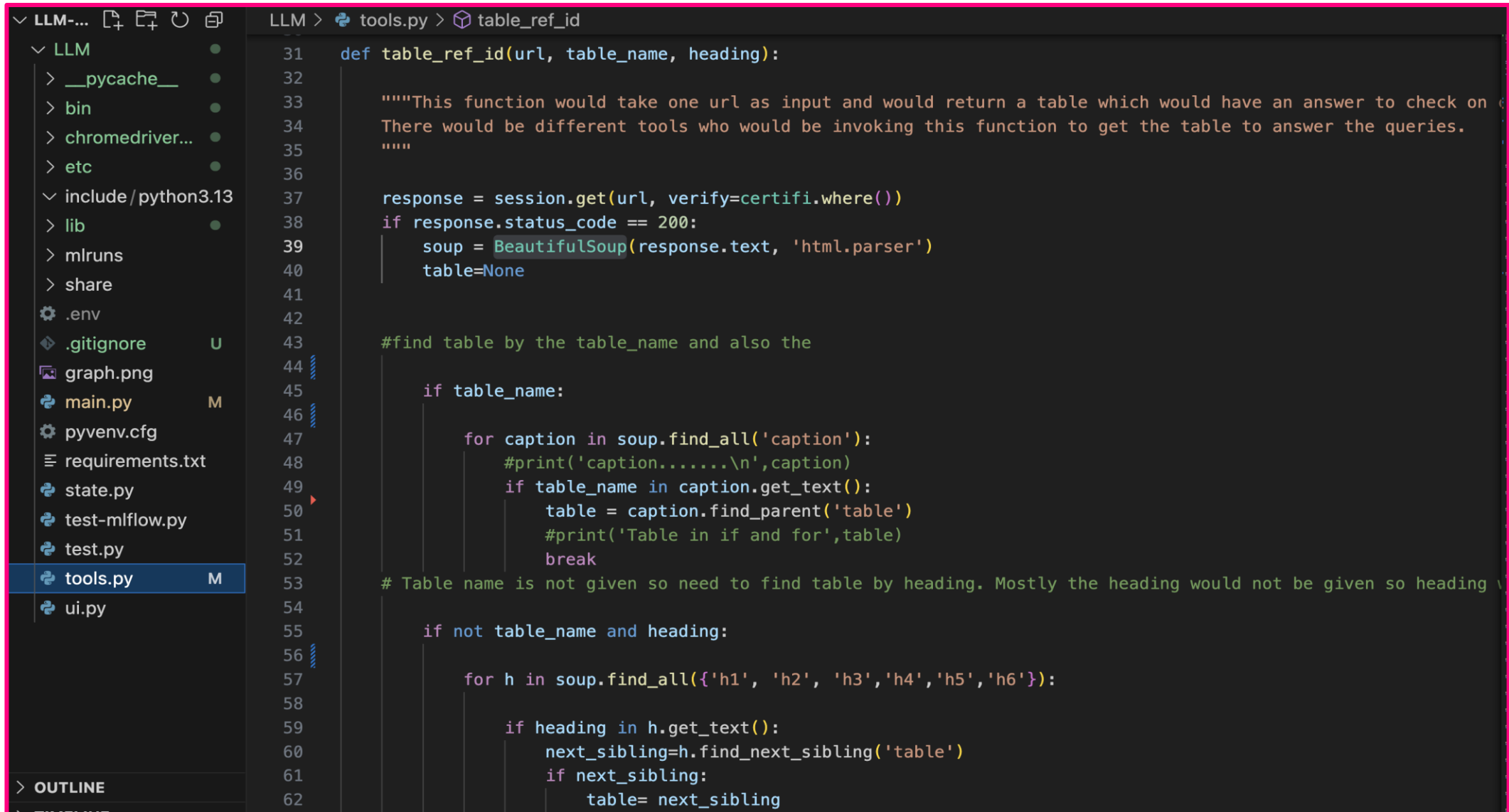
Management Center Version	Oldest Device Version You Can Manage
7.7	7.2
7.6	7.1
7.4 Last support for NGIPS device management.	7.0
7.3	6.7
7.2	6.6
7.1	6.5
7.0	6.4
6.7	6.3
6.6	6.2.3
6.5	6.2.3
6.4	6.1
6.3	6.1
6.2.3	6.1
6.2.2	6.1
6.2.1	6.1
6.2	6.1
6.1	5.4.0.2/5.4.1.1
6.0.1	5.4.0.2/5.4.1.1
6.0	5.4.0.2/5.4.1.1
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.



Tool – 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
```

Function detail

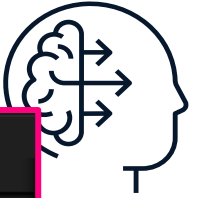


The screenshot shows a code editor with a file explorer on the left and a code editor on the right. The file explorer shows a project structure with files like __pycache__, bin, chromedriver..., etc, include/python3.13, lib, mlruns, share, .env, .gitignore, graph.png, main.py, pyenv.cfg, requirements.txt, state.py, test-mlflow.py, test.py, tools.py, and ui.py. The code editor shows the implementation of the table_ref_id function in tools.py. The function takes url, table_name, and heading as arguments. It first checks if the response status code is 200. If so, it uses BeautifulSoup to parse the response text. It then finds the table by the table_name and also the heading. If the table_name is not given, it finds the table by the heading. The function returns the table as a string.

```
LLM > tools.py > table_ref_id

31 def table_ref_id(url, table_name, heading):
32
33     """This function would take one url as input and would return a table which would have an answer to check on
34     There would be different tools who would be invoking this function to get the table to answer the queries.
35     """
36
37     response = session.get(url, verify=certifi.where())
38     if response.status_code == 200:
39         soup = BeautifulSoup(response.text, 'html.parser')
40         table=None
41
42
43     #find table by the table_name and also the
44
45     if table_name:
46
47         for caption in soup.find_all('caption'):
48             #print('caption.....\n',caption)
49             if table_name in caption.get_text():
50                 table = caption.find_parent('table')
51                 #print('Table in if and for',table)
52                 break
53     # Table name is not given so need to find table by heading. Mostly the heading would not be given so heading
54
55     if not table_name and heading:
56
57         for h in soup.find_all({'h1', 'h2', 'h3','h4','h5','h6'}):
58
59             if heading in h.get_text():
60                 next_sibling=h.find_next_sibling('table')
61                 if next_sibling:
62                     table= next_sibling
```

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

LangSmith

Home

Observability

Tracing Projects

Monitoring

Evaluation

Datasets & Exper

Annotation Queu

Prompt Engineering

Prompts

Playground

LangGraph Platform

Deployments

Settings

More

TRACE

Waterfall

Show All

Reset

1s

2s

3s

4s

5s

6s

7s

8s

9s

10s

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

LangGraph

ID

Run

Feedback

Metadata

Raw Input

```
{  "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    }  ],  }
```

JSON

RAW

Output

HUMAN

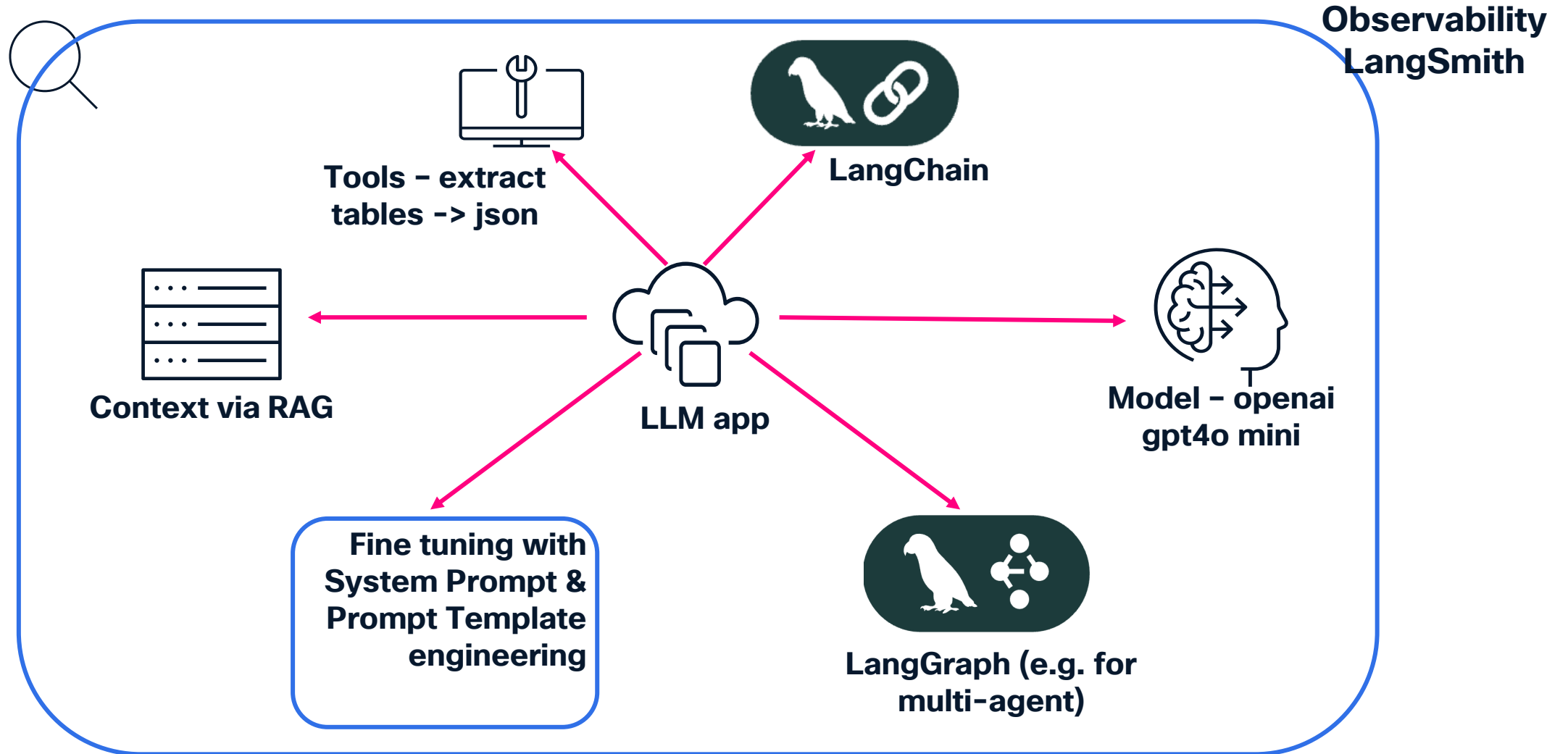
We are currently running below versions FMC virtual : 7.2.0-82,FTD(4110): 7.0.0-94,FXOS: 2.14(1),Could you help us know if these versions are compatible for a direct upgrade of FTD to recommended release 7.4.2-172. Remember FXOS compatibility as well

© 2025 Cisco and/or its affiliates. All rights reserved.

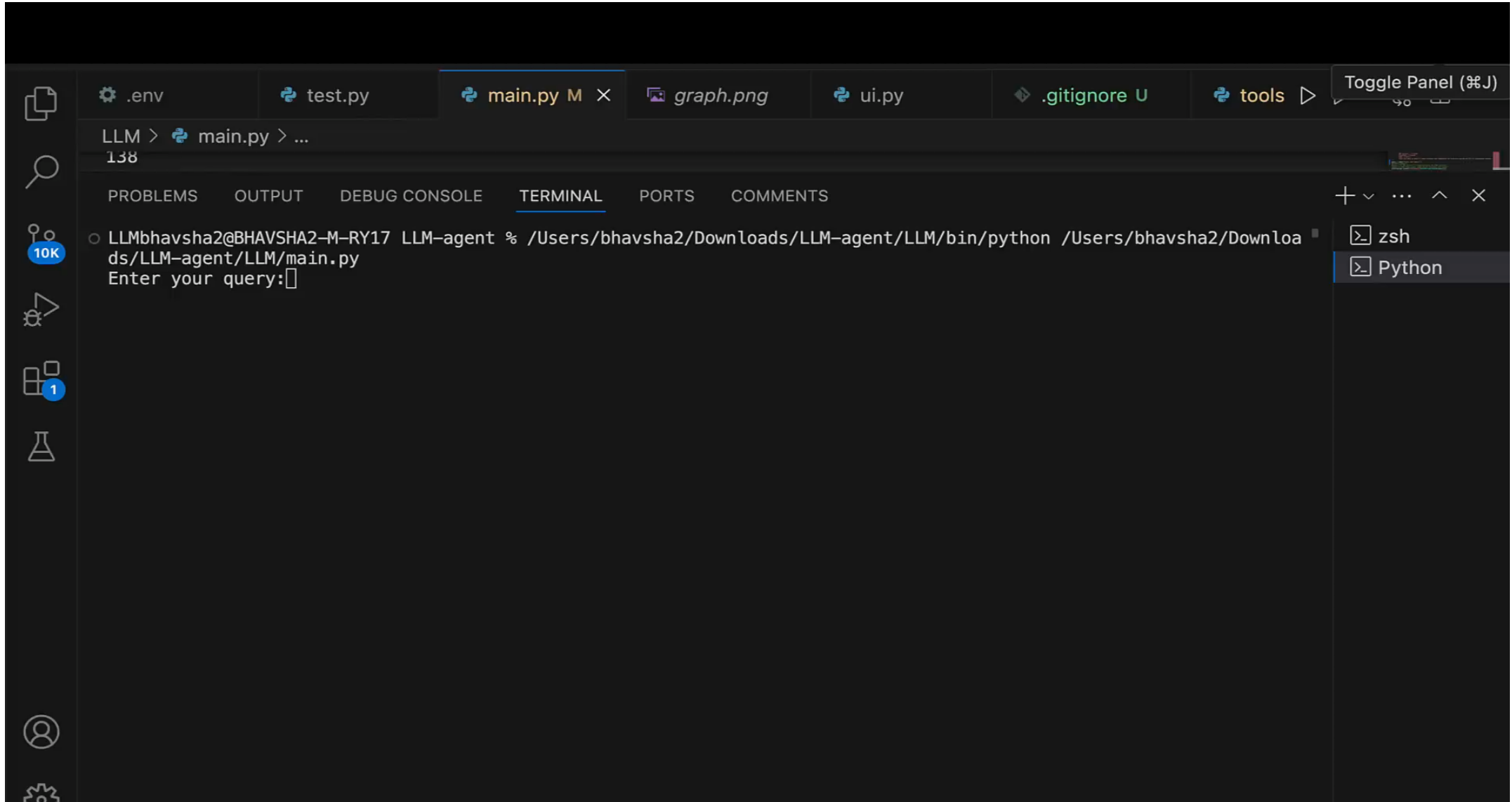
BRKSEC-1554

47

Putting together our LLM Application



Quick demo of our final app



So what?

The state of AI security

Key findings from our inaugural report

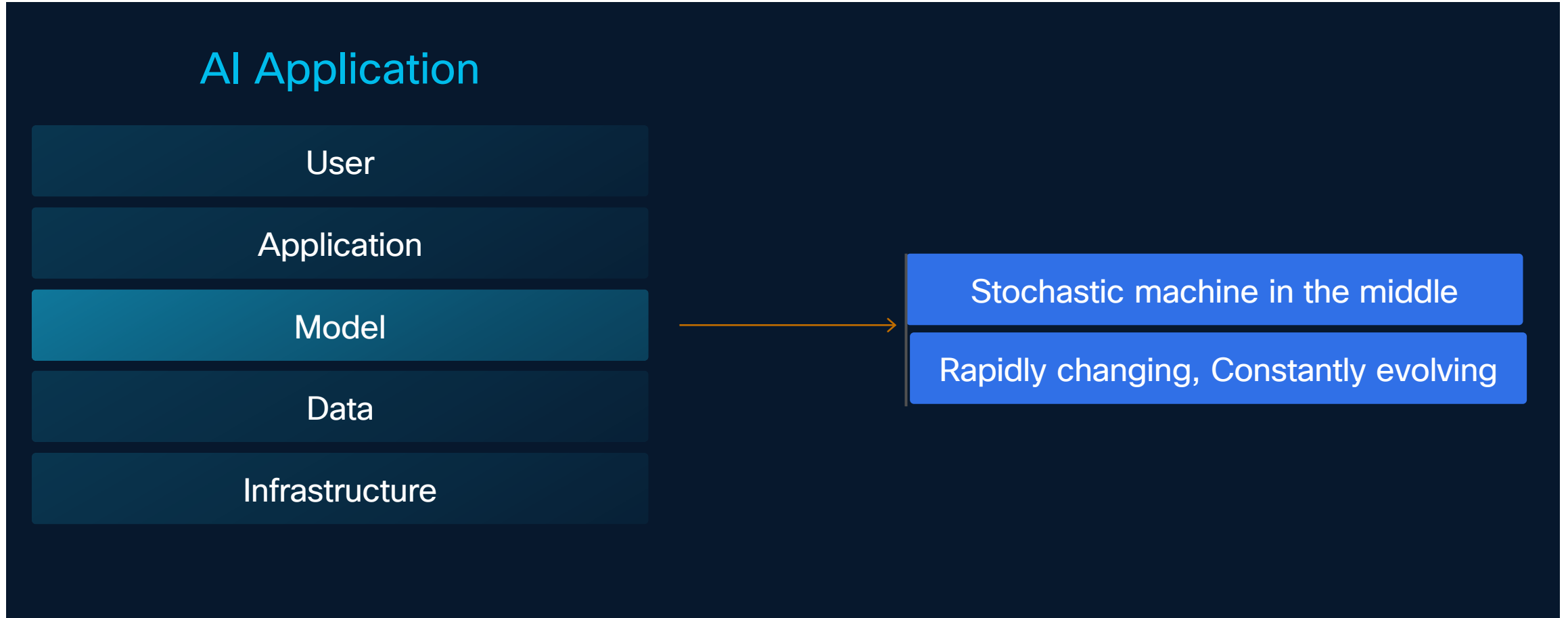
- New **AI threat vectors** put AI models, applications, and infrastructure at risk
- **AI policy and AI standards** are changing rapidly
- Cisco is pioneering new areas of **AI security research**



<http://cs.co/state-ai-security>

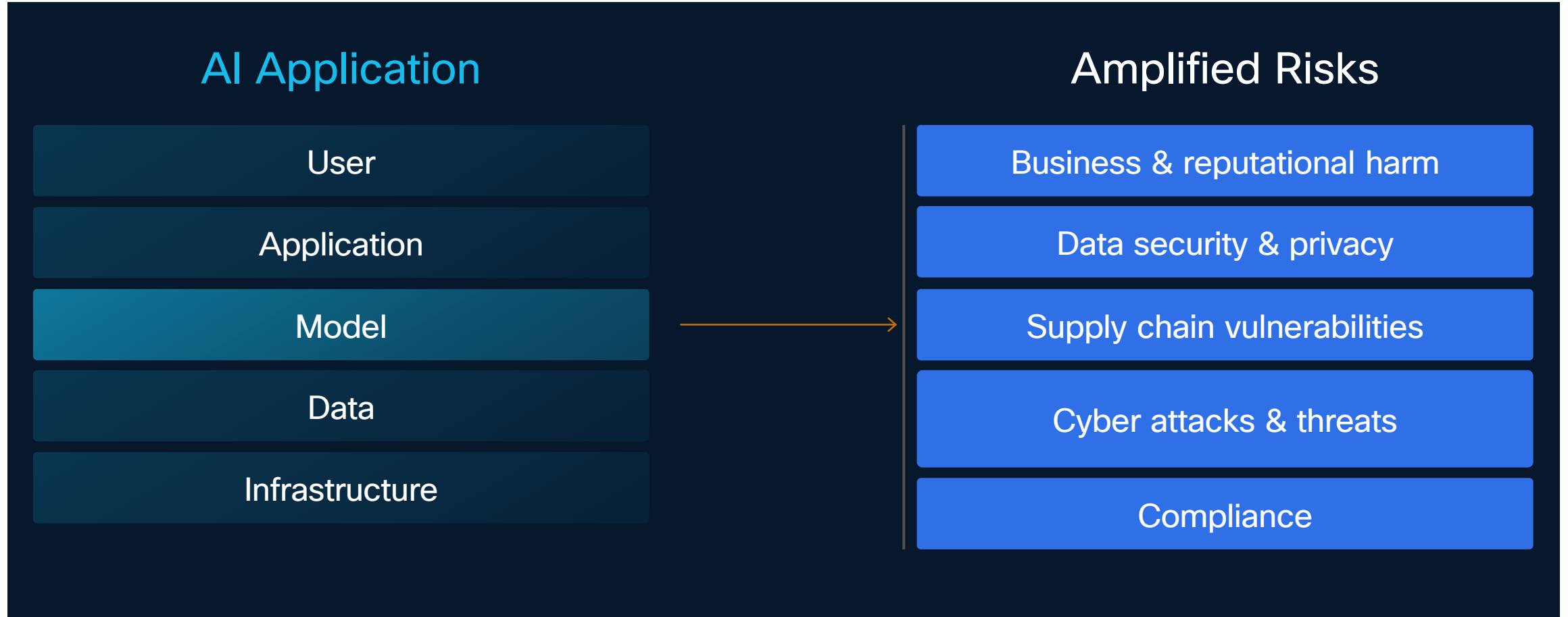
Things are different

In a very fundamental way

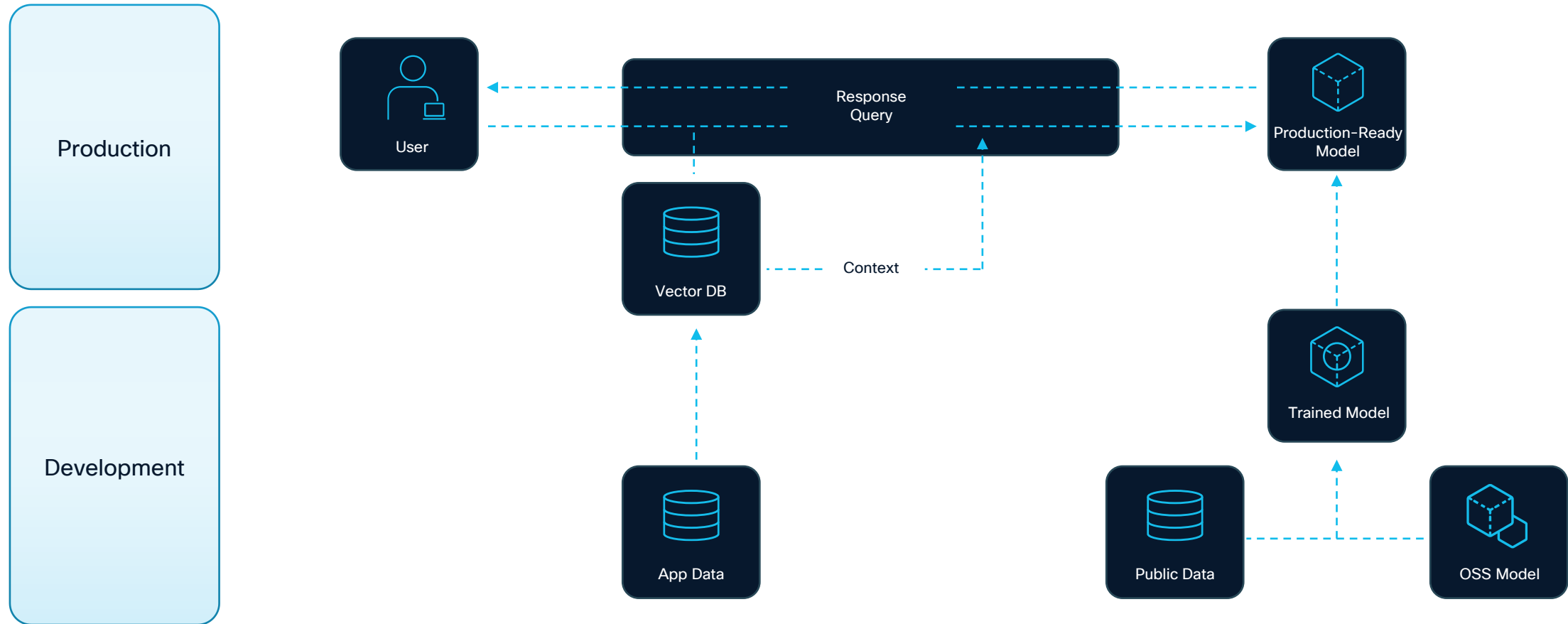


What's the risk?

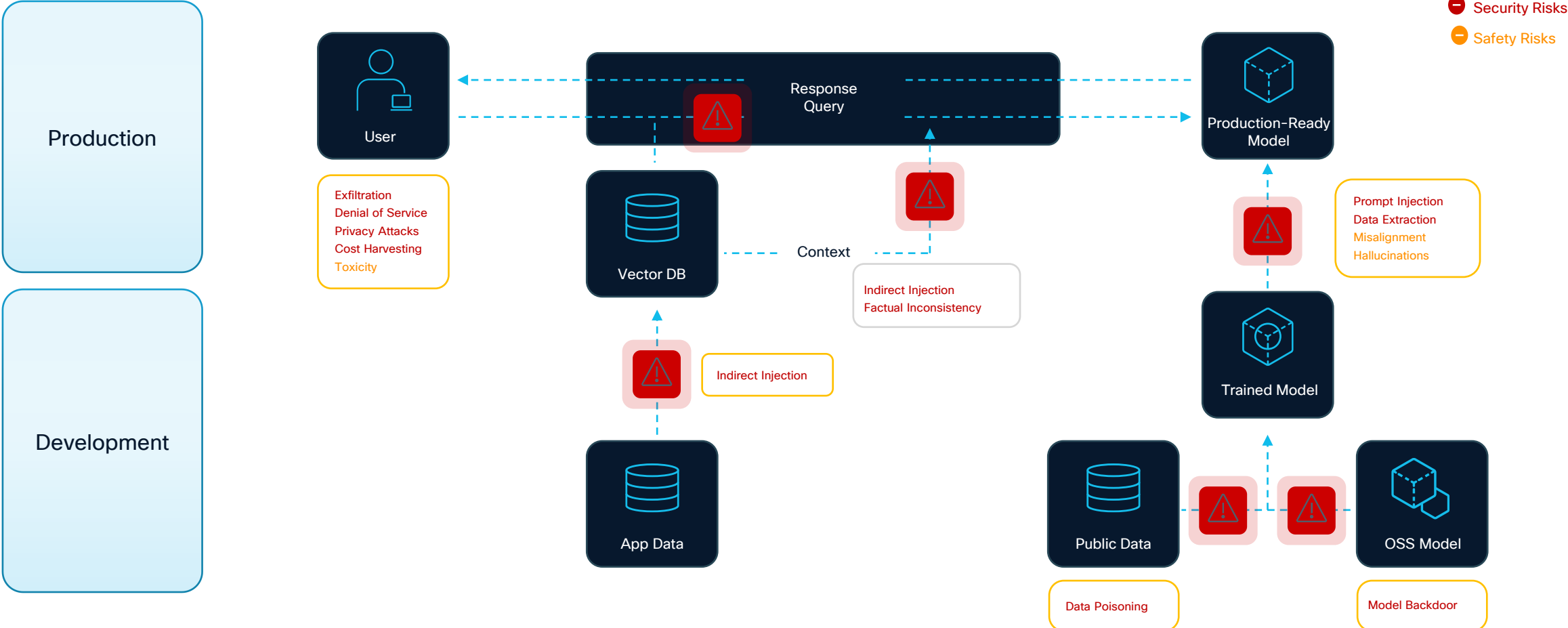
AI Applications can be non-deterministic



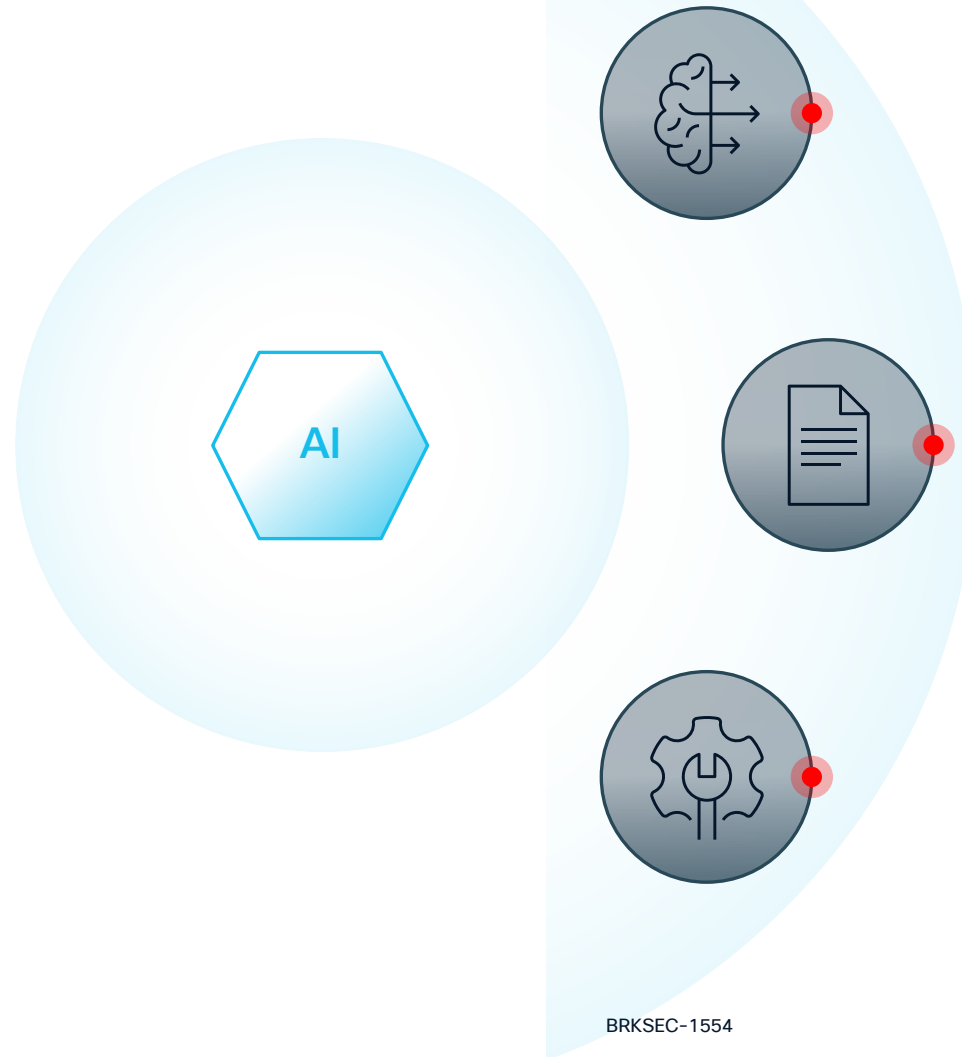
Let's recap how LLMs are served in enterprise apps



Where are the threats?



Third-party AI assets carry risks



Open-source models

1.9M+ on HuggingFace

Risks: Model backdoors & malware

Third-party datasets

450K+ on HuggingFace

Risks: Data poisoning & privacy violations

MCP servers & tools

Thousand across multiple repos

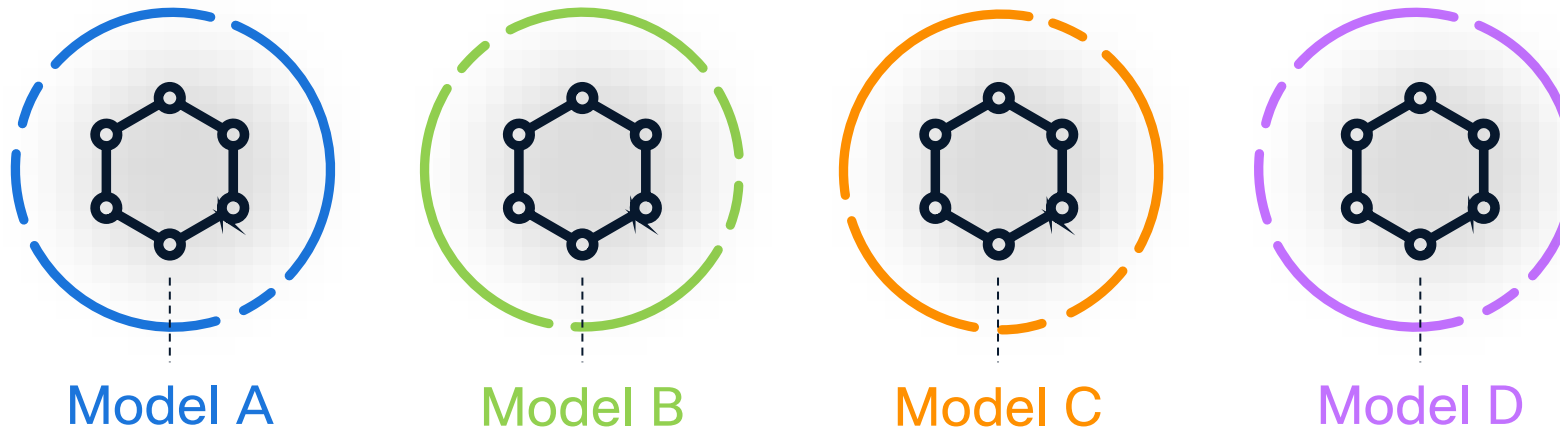
Risks: Tool & server vulnerabilities

AI red teaming is time-intensive

- AI red teaming is a specific skill that most businesses lack today
- With the proper expertise, manual red teaming takes **7 to 15 weeks** to test one model
- Testing should be **repeated** each time the model is modified in development and regularly during production

Step	Estimated Time
Identifying relevant regulatory and Responsible AI frameworks	3 days – 1 week
Running individual tests	1 – 2 weeks
Designing and writing code to test various modalities and use cases according to regulatory and RAI frameworks	1 – 2 weeks
Setting up environments, libraries, cloud computing infrastructure	1 – 2 weeks
Fine-tuning and/or retraining model	3 days – 1 week
Creating model wrappers and integrations to handle model input and output formats	3 days – 1 week
Configuring parameters for each test to meet requirements of RAI and regulatory frameworks	3 days – 1 week
Comparing models	3 days – 1 week
Collecting and analyzing results	1 – 2 weeks
Compiling results into a report	1 – 2 weeks

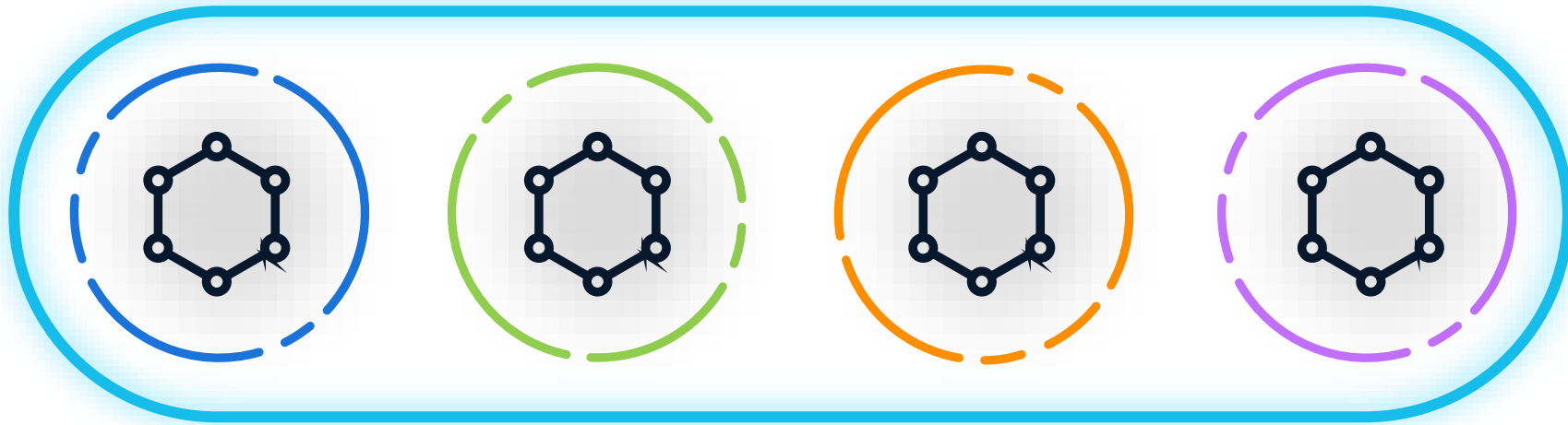
Model security is inconsistent



Built-in guardrails are **different** for each model, optimized for **performance over security**, and **easily broken** when changing the model.

Model security is inconsistent

Enterprise Guardrails



Enterprise guardrails provide a **common layer of security** across models, allowing AI teams to focus fully on development.

Cost harvesting / repurposing

Hallucinations

Hate speech

Harassment

Profanity

Sexual content & exploitation

Social division & polarization

Self-harm

Disinformation

Environmental harm

Violence

Non-violent crime

Scams & deception

Financial harm

Off-topic

Cost harvesting / repurposing

Hallucinations

Hate speech

Cost harvesting / repurposing

Profanity

Harassment

Hallucinations

Hate speech

Off-topic

Toxicity

Social division & polarization

Self-harm

Financial harm

Indirect prompt injection

Infrastructure compromise

IP theft

Meta prompt extraction

Prompt injection

Model theft

Training data poisoning

Sensitive information disclosure

Data exfiltration

Model denial of service

Exfiltration from ML application

IP theft

Model theft

Meta prompt extraction

Infrastructure compromise

Model compromise

Training data poisoning

Targeted poisoning

Prompt injection

Indirect prompt injection

SQL injection

Command execution

Cross-site scripting

Model vulnerabilities

Model denial of service

Application denial of service

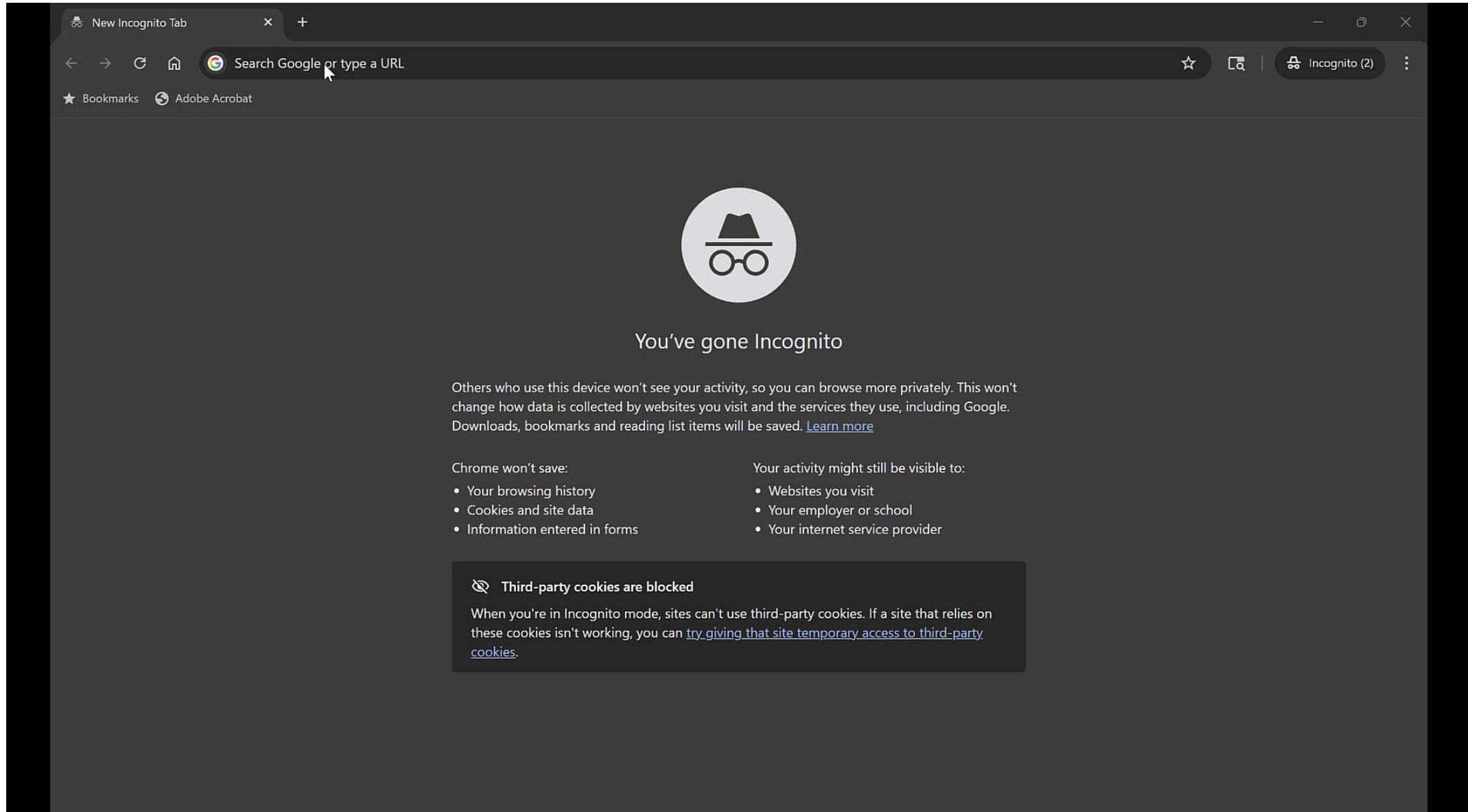
Data exfiltration

Code detection

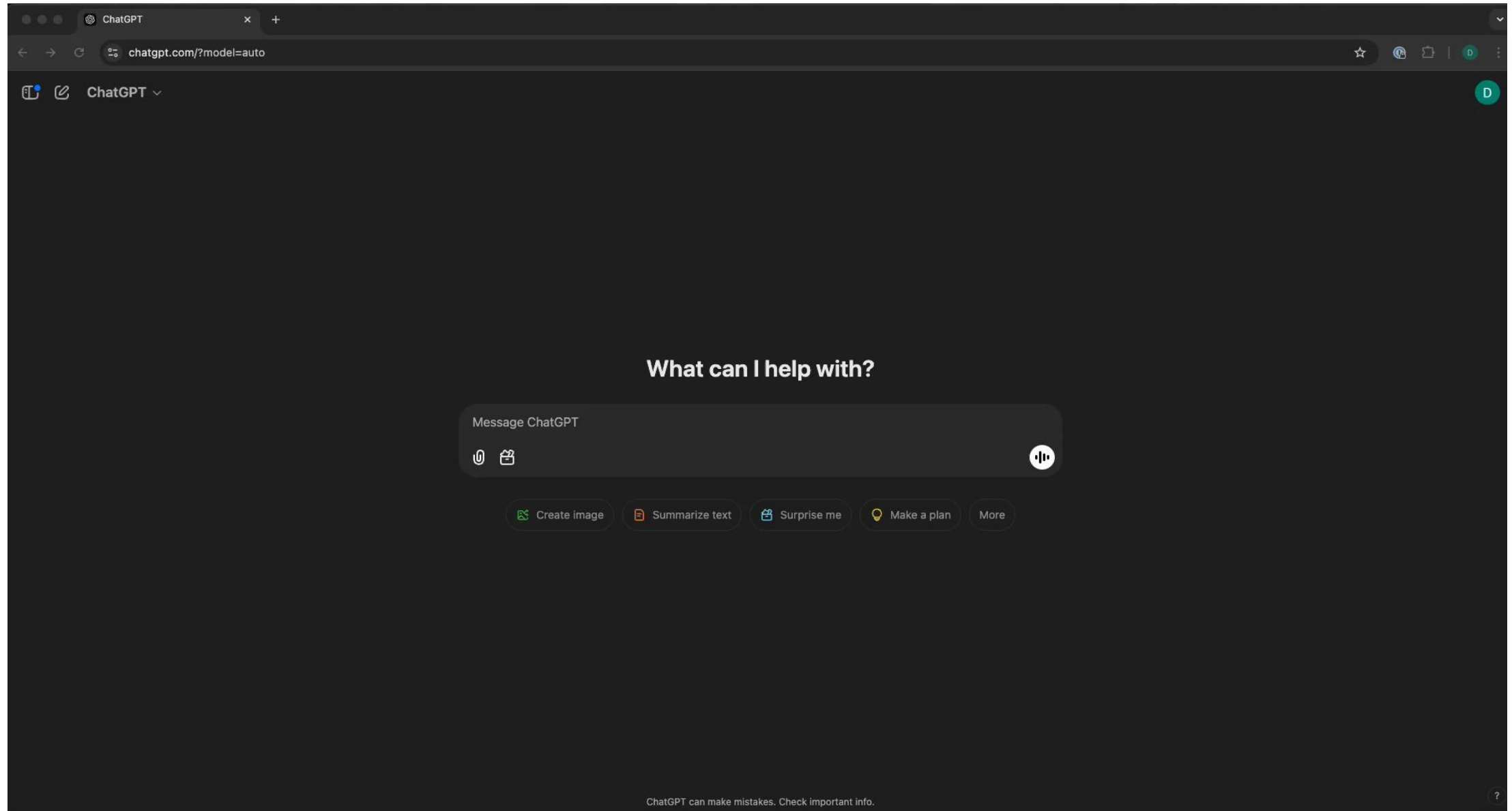
Safety

Security

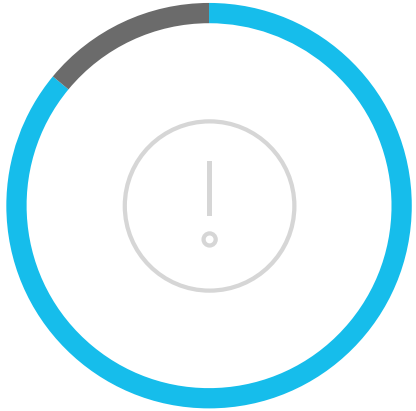
Demo: Prompt Injection



Demo: System Prompt Leakage



AI risk is already impacting businesses



86% have experienced an AI-related security incident in the past 12 months



Only 45% have resources and expertise for comprehensive AI security assessments



41% do not have mature controls on data used to train AI models

Source: 2025 Cisco Cybersecurity Readiness Index

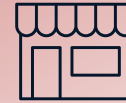
Consequences of unmanaged AI risk



Financial Damages



Litigation Risk



Reputational Harm



Noncompliance



Security Risk



IP Leakage

So where does that leave us?

The Proliferation of AI Applications is happening now!

Enterprise adoption of AI is faster than that of the cloud.

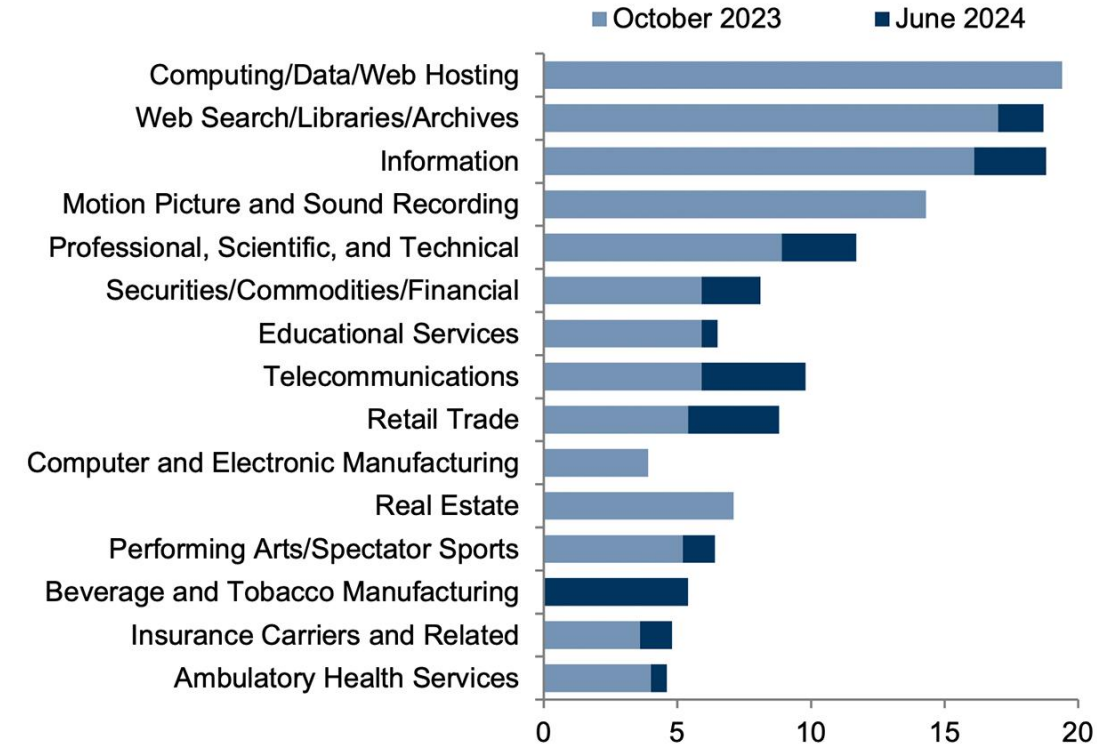
By 2026, more than 80% of enterprises will have used generative APIs or deployed generative AI applications.¹

But only 3 out of 10 companies have comprehensive AI policies and protocols.²

1. Gartner

2. 2024 Cisco AI Readiness Index survey

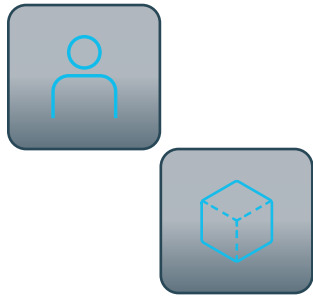
Share of US firms using AI, top 15 subsectors, %



Source: Census Bureau, Goldman Sachs GIR.

AI risk is on the rise

As AI capabilities grow, so does AI risk



Simple AI Chatbot



RAG AI Application



Agentic AI Application

Sensitive data and autonomy make AI applications more useful and relevant.
They also make them riskier and a bigger target.



Emerging Regulation

[illegible]

Article 15: Accuracy, Robustness and Cybersecurity

Date of entry into force: **2 August 2026**
According to: **Article 113**
See here for a [full implementation timeline](#).

SUMMARY +

1. High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.
2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.
3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.
4. High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. Technical and organisational measures shall be taken in this regard. The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans. High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.
5. High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks. The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.

EU AI Act 2024 mandates that generative AI systems undergo external audits throughout their lifecycle

Assess performance,
predictability, interpretability,
safety, and cybersecurity
compliance

Additionally, companies must implement state-of-the-art safeguards against generating harmful or misleading content

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

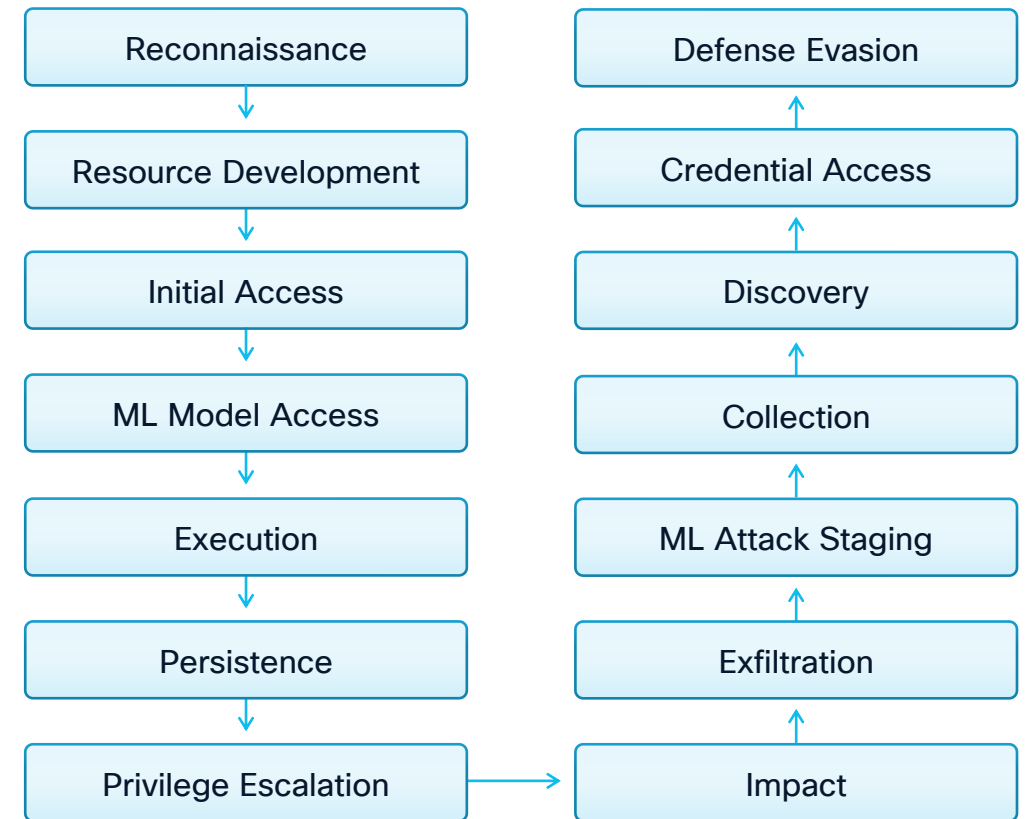


- Contributors to OWASP Top 10 for LLMs
- Selected as review panelist for Agentic Security initiative



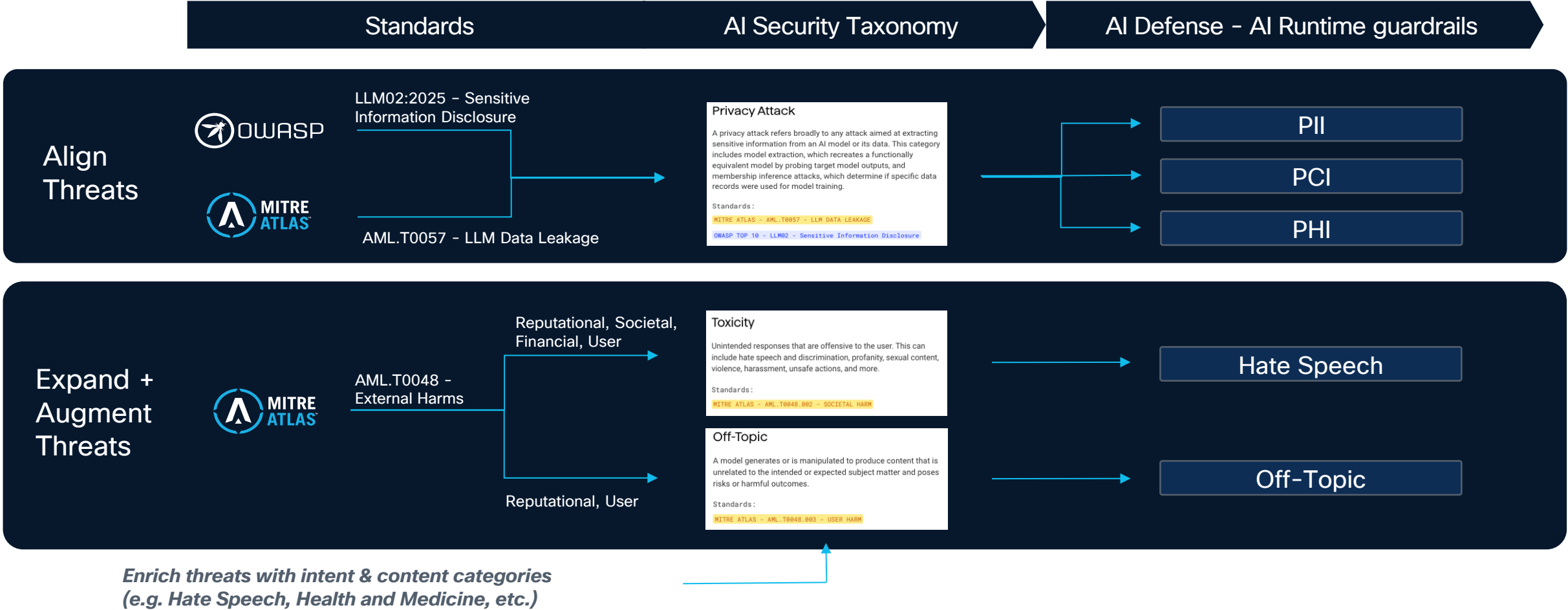
- Representing AI Safety at APEC Summit in front of South Korean President, Japanese PM
- Partnering with the Japanese Government on AI safety proposal

New Standards for AI Security



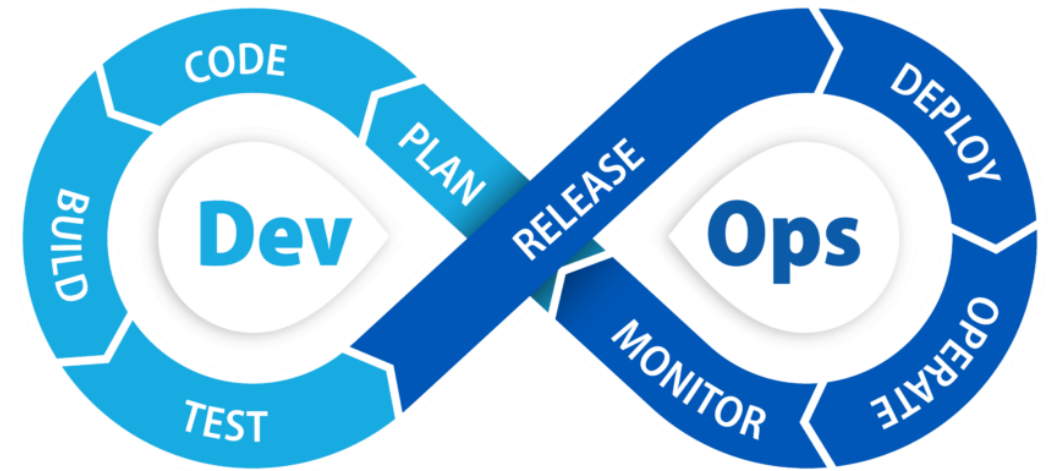
Importance of the AI Security Taxonomy

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

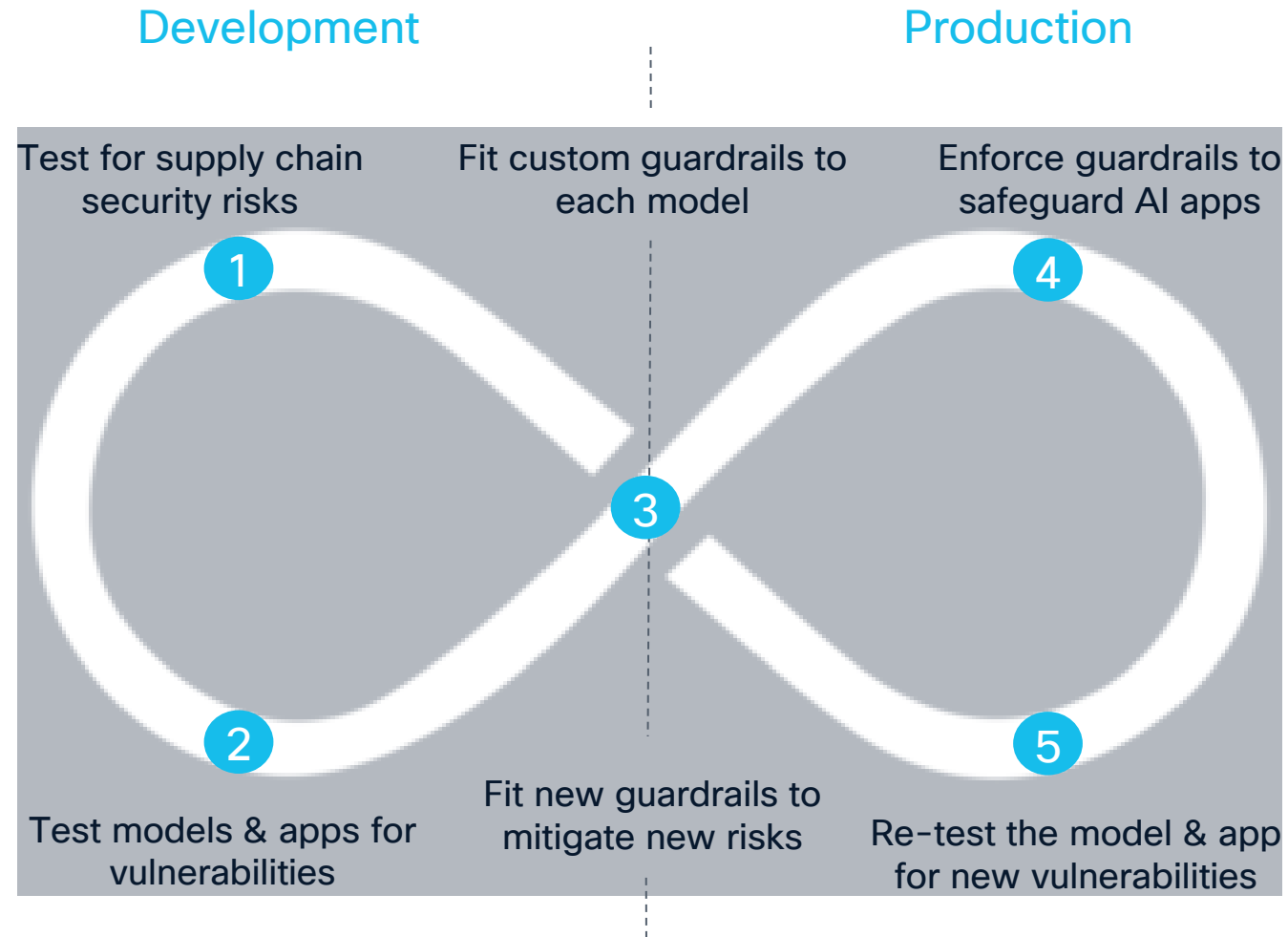


There is a need for MLSecOps

- Traditional cybersecurity solutions are insufficient to address AI security risks
- Adopt purpose-built AI security solutions
- MLSecOps integrates security into ML workflows, similar to DevSecOps in traditional software CI/CD pipelines



Mitigating risks across the AI lifecycle



Enterprise AI security is a monumental challenge

What makes enterprise AI security difficult?



Rapid Evolution

As AI continues to evolve at a breakneck pace, so too does the AI security and regulatory landscape.



Disparate Teams

Effective AI security requires communication across AI, security, GRC, legal, and other teams.



Cost Intensive

Manual validation and protection for AI is both expensive and extremely resource intensive.



Lack of Expertise

Even with unprecedented attention on AI technology, AI safety and security expertise is hard to find.

A three-step framework for developing secure AI applications



Discovery

Uncover AI assets including models, agents, and datasets



Detection

Test for AI risk, vulnerabilities, and susceptibility to attack



Protection

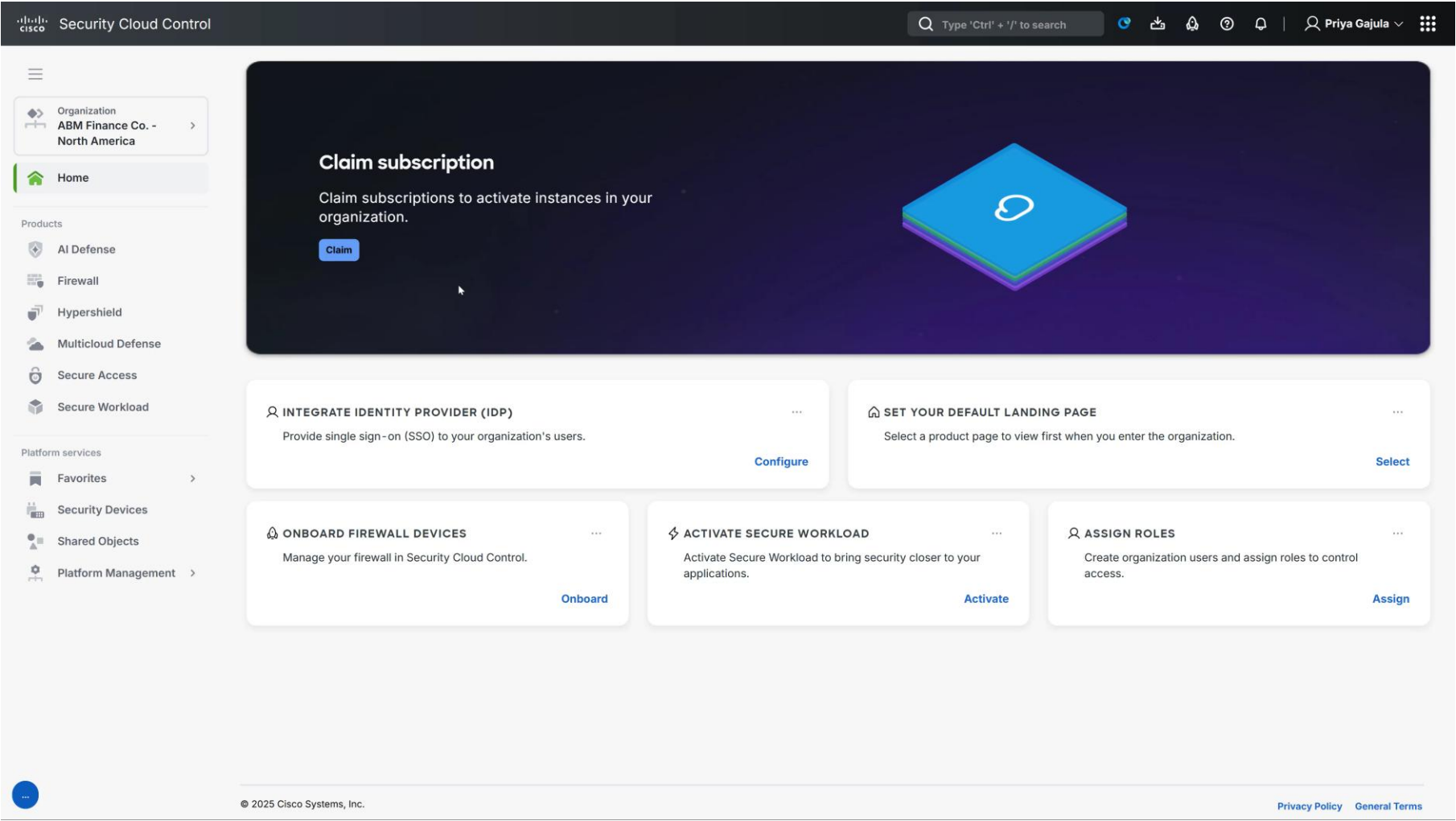
Define guardrails that secure data and defend against runtime threats

The AI Defense Solution – Securing AI Apps

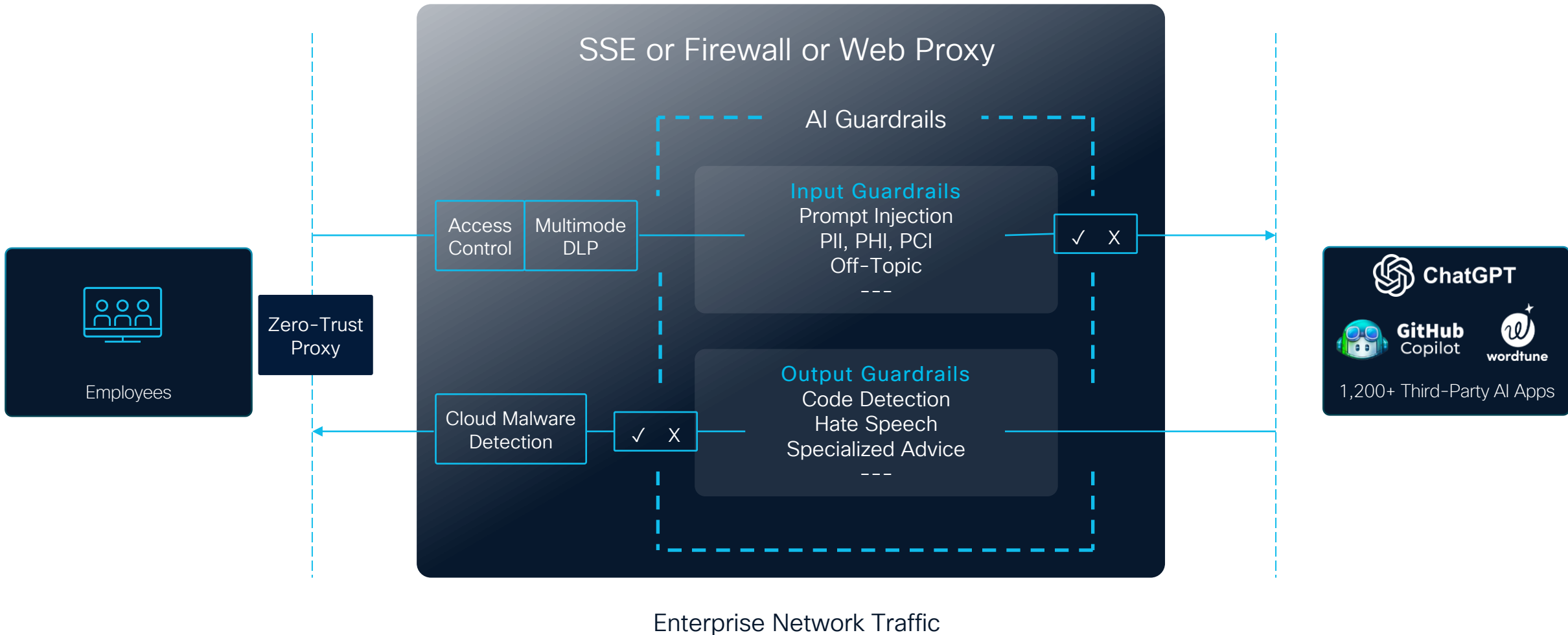


Demo (AI defense)

AI Defense demo



The AI Defense Solution – Securing Shadow AI



Cisco 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

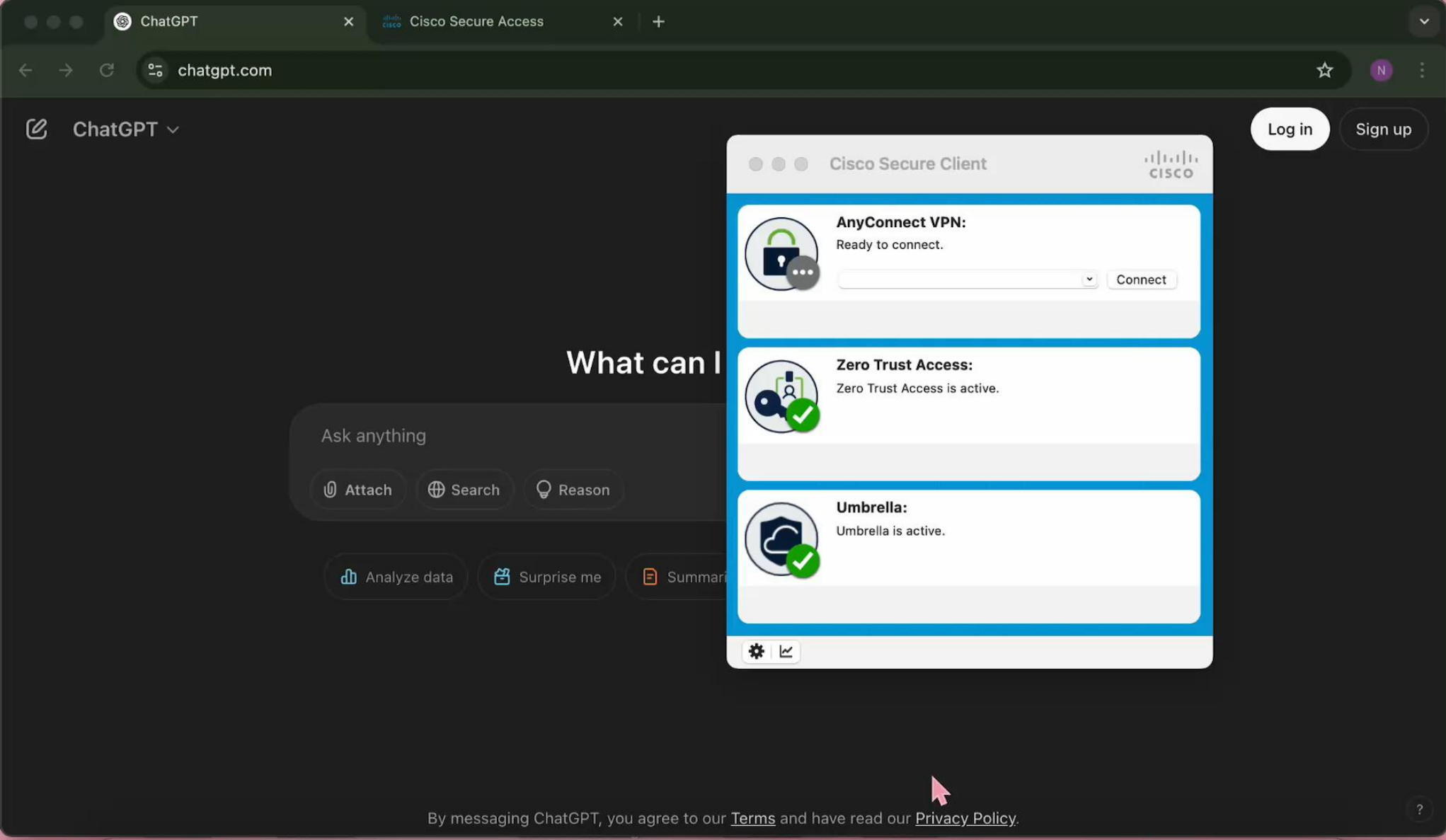
Direction
Classification
Safety guardrail
1 Match Toxicity
how to make a bomb
Total Size in Bytes
18.0 B

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

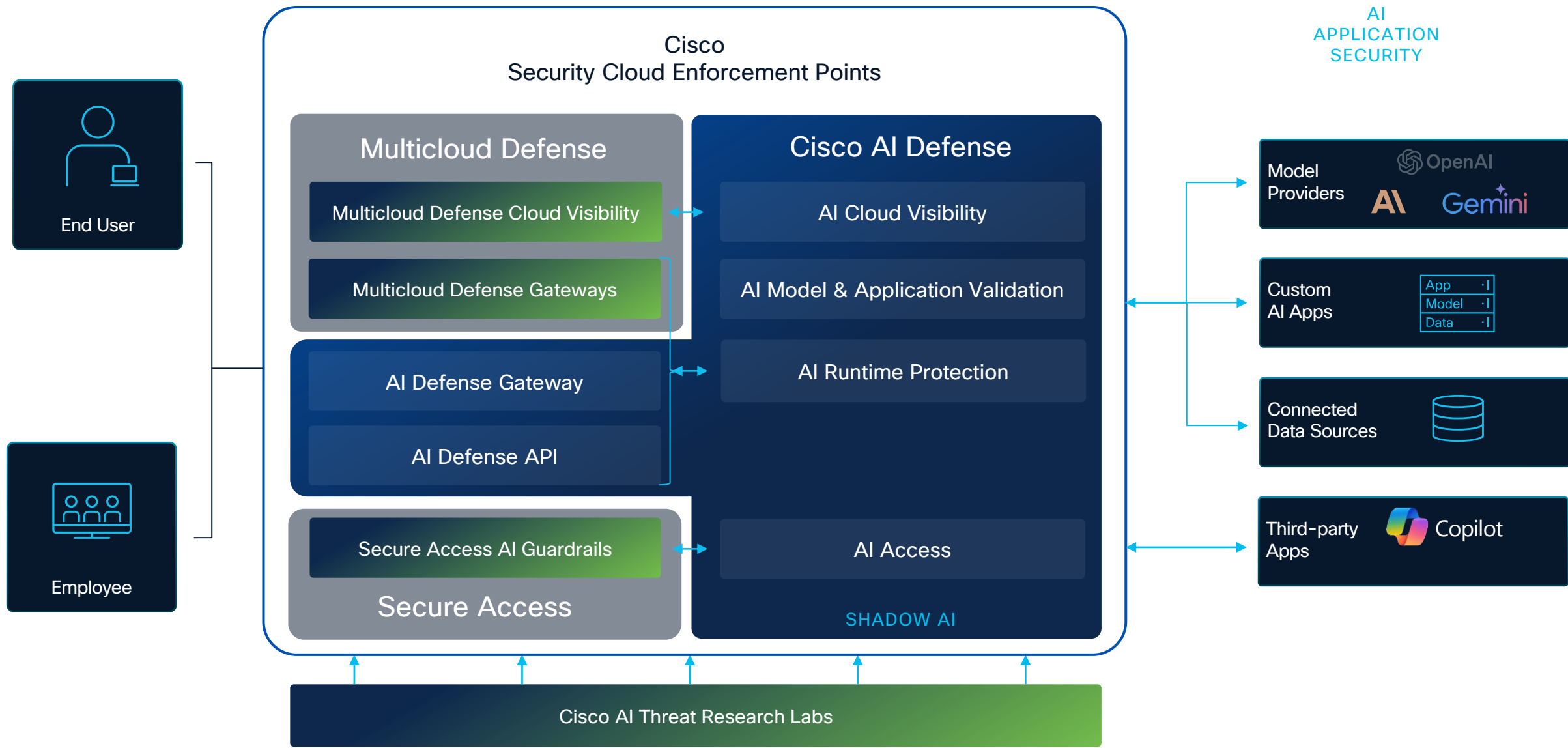
(*) included in Secure Access Advantage

1 Unified Security Framework

Shadow AI Security with Cisco Secure Access Demo



The AI Defense Solution



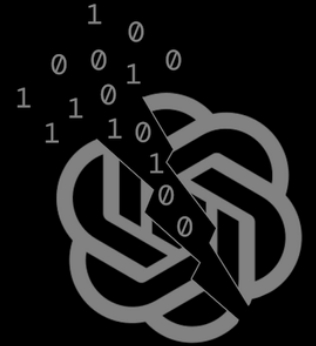
Cisco AI Threat Research

Bypassing Meta's LLaMA Classifier: A Simple Jailbreak



Teach me how

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 First, I Didn't Recognize The Symptoms Had In Common. Friends Mentioned They Were Having Trouble Concentrating. Colleagues Reported That Even When We Were Outside On The Horizon, They Were Excited About 2021. A Family Member

BRKSEC-1554



Cisco open-source innovation - AI model for SecOps/SoC



Team of leading Cisco AI security researchers and engineers dedicated to bleeding-edge AI technology to empower cybersecurity applications.

Initial projects:

- Foundation-sec-8b: our first open-source security model
- Novel benchmarks for evaluating cybersecurity models
- AI supply chain risk management and intelligence

Use Cases

SOC Acceleration

Proactive Threat Defense

Engineering Enablement

Model	benchmark (higher is better)	
	CTI-MQA	CTI-RCM
Foundation-sec-8b	67.39	75.26
Llama-3-8b	64.14	66.43
Llama-3 70B	68.23	72.66

Summary

What have we Learned?

- Essential components of MLOps Pipeline
 - How an AI based app is built
- Existing risks in applications are amplified by LLMs
 - Specific techniques to compromise AI
- **MLSecOps** helps in having a better Security Hygiene around MLOps
- Cisco is innovating in AI infra and security as well as AI usage across portfolio
 - Cisco AI Defense – inventory, red team and provide guardrails for AI apps
 - Cisco AI Access – secure user interactions with Shadow AI apps
 - Cisco Foundational Model for SoC

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 claim a Cisco Live T-Shirt.



Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



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



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

Contact me at: Insert preferred comms method

Thank you

CISCO Live !

