

The background image shows the Auckland city skyline at sunset, with the sky transitioning from blue to orange and red. The iconic Sky Tower is prominent in the center-right. The city lights begin to glow as the sun sets over the Waitemata Harbour.

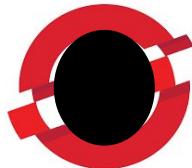
Welcome! Auckland OpenShift Meetup

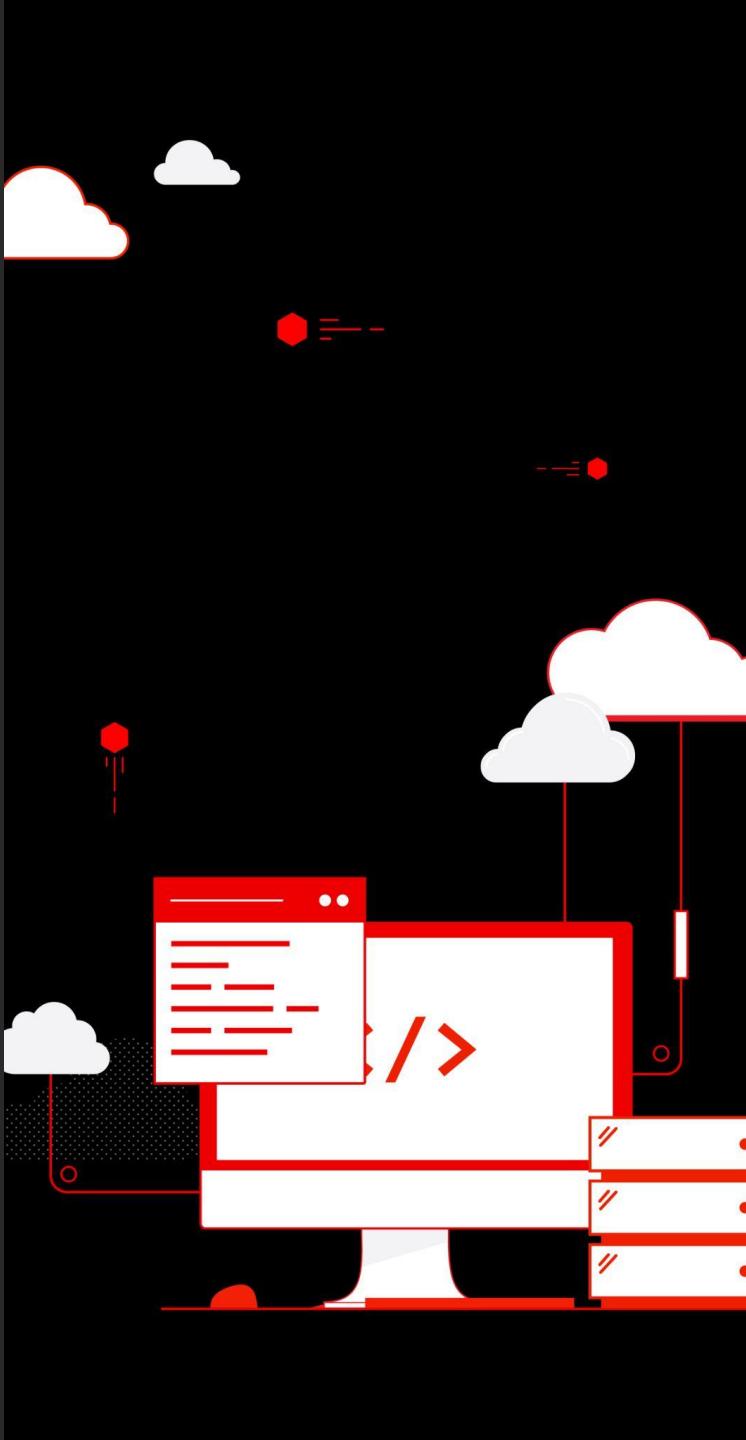


19 November – 1:00 – 5:00 pm



Mākora Toru, Generator
PWC TOWER, L2, 15 CUSTOMS STREET WEST





AI & AI Ops

Benjamin Farr
Principal Solution Architect

Wilson Toh
Principal Sales Specialist

Introductions



Benjamin Farr

- Specialist in App Dev, OpenShift, Ansible & learning AI
- My foray into AI was developing a ML PoC for Gartner in 2019

A screenshot of a mobile game titled "TRAIN OUR ROBOT TO DANCE!". The game features a bar chart on the right showing "Playing!" and "Your Dance Moves" with various colored bars representing different frequency ranges. On the left, there are buttons for "SPIN LEFT!" and "SPIN RIGHT!". Below the buttons are blue arrows indicating the direction of movement. At the bottom, there is a small grid where the robot's moves are plotted. The Red Hat logo is visible in the top left corner.

Wilson Toh

- OpenShift AI, Security
- Telco Background
- ISC2 CISSP member
- ISC2 ISSMP certificate

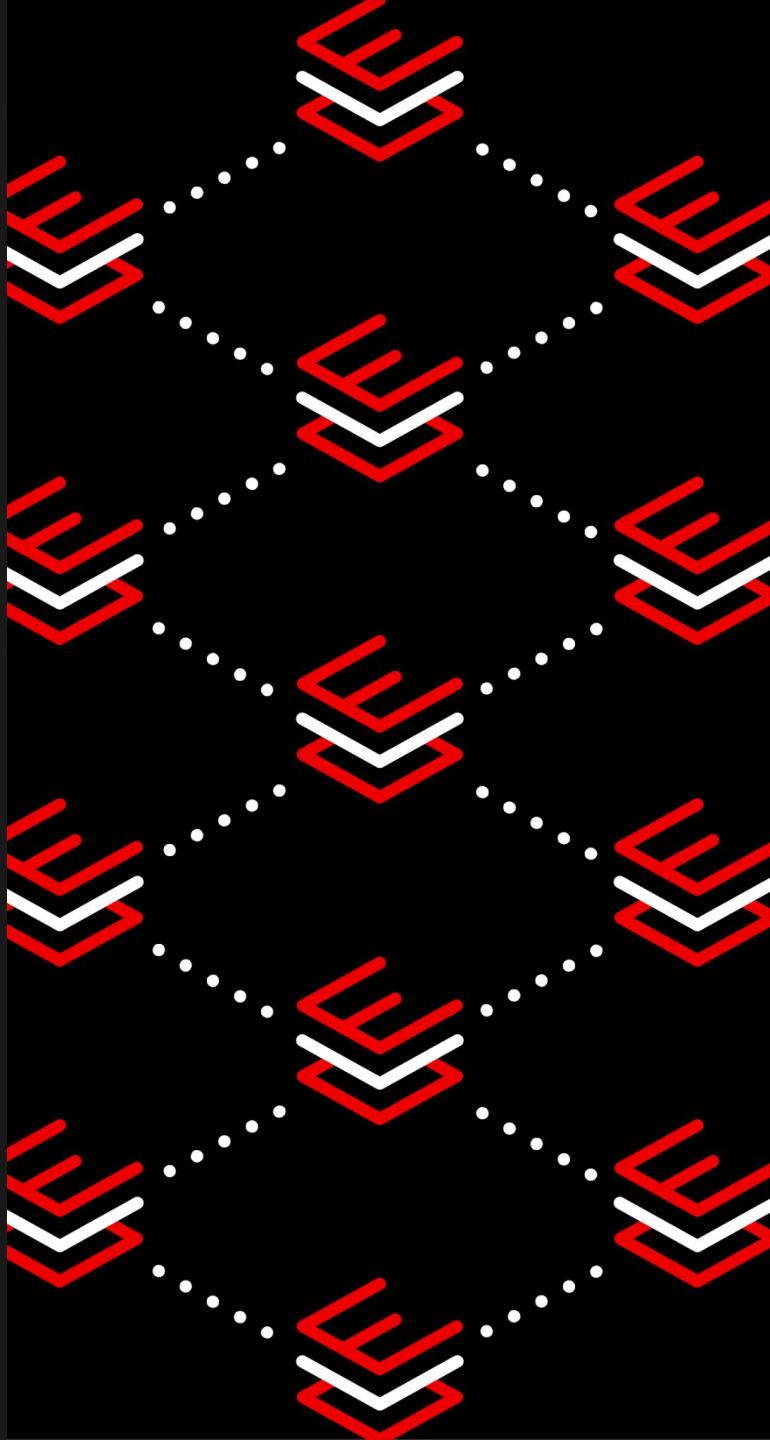


Linkedin Security Articles:
AI Security: Navigating the Future of Cybersecurity –
Leveraging AI to Manage AI Security

BPFDoor: Mitigating Advanced Kernel-Level Threats

Agenda

- ▶ AI Trends
- ▶ AI Security
- ▶ Introduction to AI ecosystem
- ▶ AI Ops Demo
- ▶ How can you get started



AI Trends

What is AI Native?

If Cloud Native was about *how* we deploy, is AI Native
about **who decides, what they decide, and who controls
the decision?**

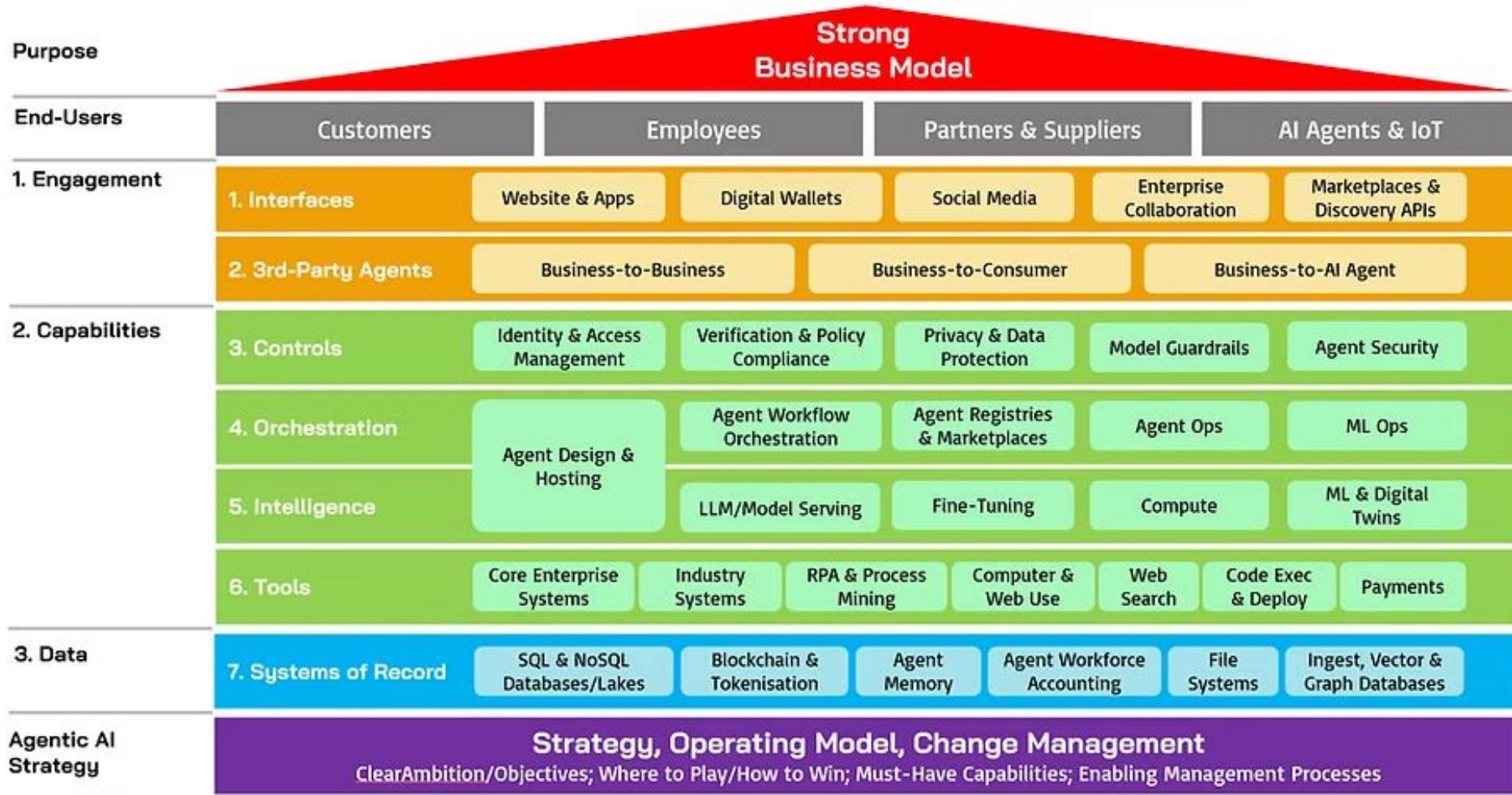


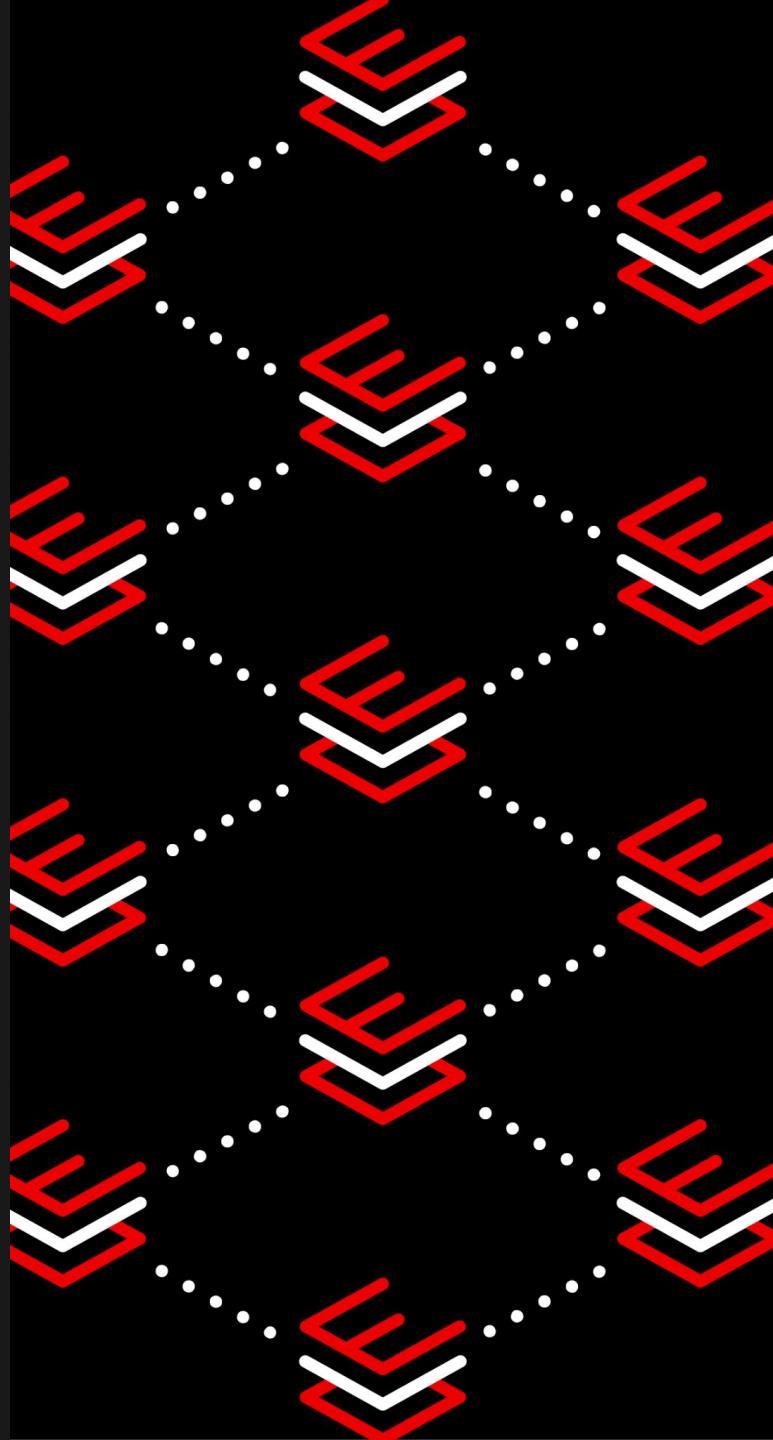
AI Trends

- Increased AI Spending & Generative AI Boom
- Open-Source AI Adoption
- Specialized LLMs
- AI Risks & Regulation
- Data Governance & Literacy
- Prompt Engineering & Data Professions
- Cloud & Agentic AI



Agentic AI Stack for Enterprises





AI Security



- AI Security: Navigating the Future of Cybersecurity – Leveraging AI to Manage AI Security

- BPFDoor: Mitigating Advanced Kernel-Level Threats

What is AI Security

Is AI Security the same or similar as *AI Safety*?

Security vs Safety

Safety is the broader topic. Safety controls mitigate harm. Harm will vary depending on how the output is used. Ranges from embarrassment and offensive content, all the way through to dangerous, illegal or physical harm.

Security is closely related and often overlapping topic. Security controls protect confidentiality, integrity or availability of the system.

Red Hat - AI Security and Safety

- Currently there is no consensus in these industry about these terms
- Red Hat Product Security is looking at both of these.
- While we are experts in security, safety is entirely new to us.
- Many teams in Red Hat looking at AI safety.

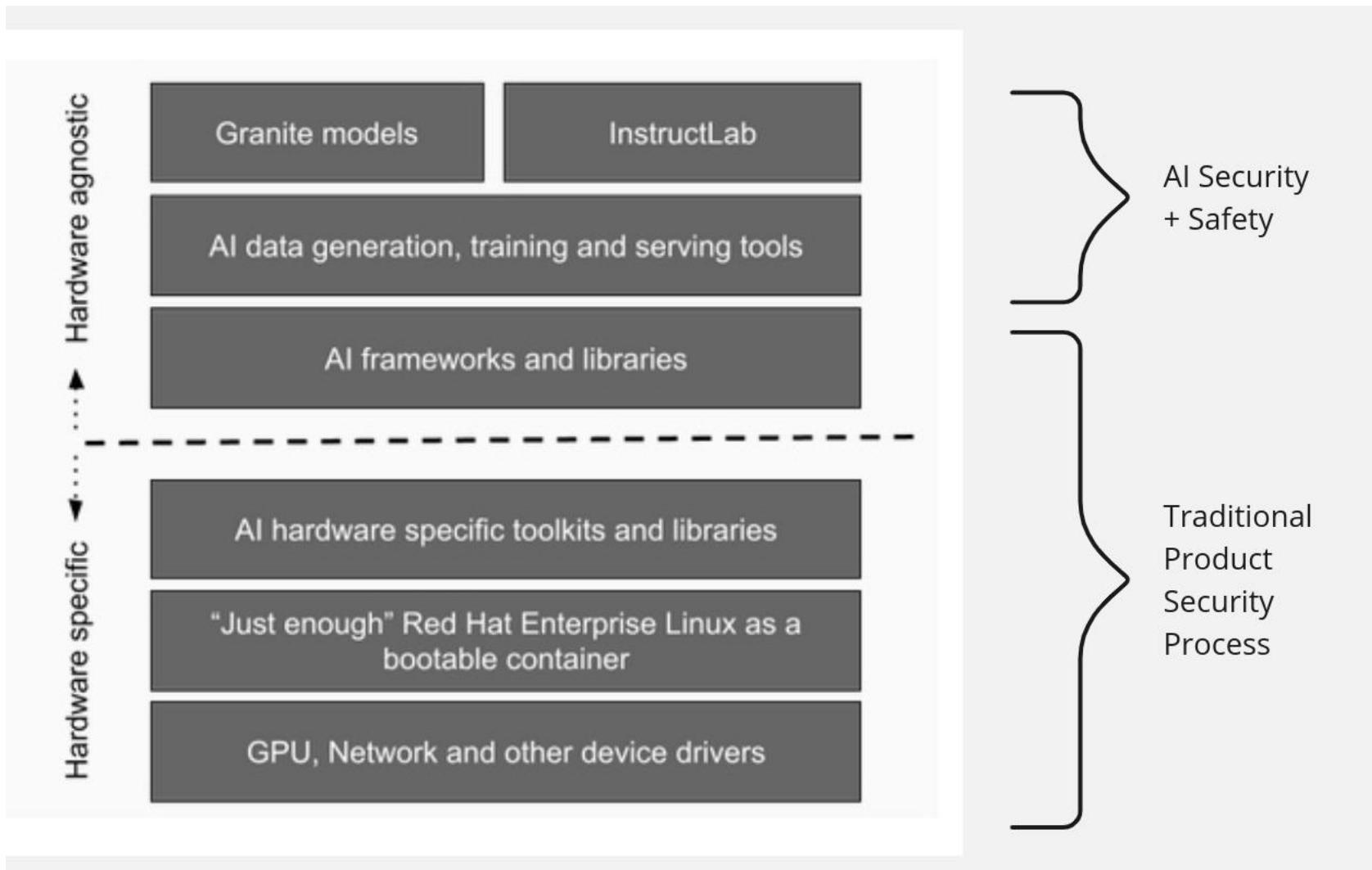
AI Security flaw

- Any flaw which affects the Confidentiality, Integrity, Availability of the AI model is called a security flaw.
- Our definitions is different from other vendors who have closed/closed-weight models.
- Our definition is in line with how we classify other security flaws, but its open and evolving.
- Common flaws include, jailbreak, adversarial fine-tuning etc

AI Safety hazards

- Toxic and harmful content produced by the model.
- Safety Taxonomy trained into the granite model.
- Industry view of safety depends on culture, ethics and popular beliefs –
Corporate and/or legal requirements may be different
- We are inclining towards the MLCommons taxonomy, which is generally accepted in the industry.

Red Hat AI - Clear distinction between platform and AI model(s)



The introduction of agentic AI requires organizations to update their risk taxonomies.

AI risks by enterprise risk category, illustrative (not exhaustive)

Financial	Operational	People	Regulatory	Reputational	Strategic
<ul style="list-style-type: none"> ● AI cost overrun ● Algorithmic financial exposure ● Synthetic fraud and transaction risk 	<ul style="list-style-type: none"> ● Data corruption/model poisoning ● System drift/misbehavior ● Systemic dependency/lack of fallback 	<ul style="list-style-type: none"> ● Accountability ambiguity/loss of human oversight ● Deskilling ● Skill gaps ● Stress and resistance ● Workforce displacement 	<ul style="list-style-type: none"> ● Bias/discrimination ● Lack of transparency/explainability ● Noncompliance ● Unauthorized data use or disclosure 	<ul style="list-style-type: none"> ● Controversial or misled AI decisions ● Stakeholder distrust ● Perceived ethical violations 	<ul style="list-style-type: none"> ● Opaque decision influence ● Overreliance ● Strategic misalignment
<p>Acceleration because of agentic AI</p> <ul style="list-style-type: none"> ● Chained vulnerabilities: Strategies built on fragile multiagent architectures ● Cross-agent task escalation: Agents expand decision scope or delegate tasks beyond intent ● Data corruption propagation: Impact of low data quality is amplified by decision chains across agents ● Synthetic identity risk: Use of agents to simulate identities, generate fraud, or manipulate transactions ● Untraceable data leakage: Exchange of data between agents without oversight obscures data leaks 					
<p>● Gen AI risks not linked to novel agentic AI risk types</p>					

AI software security threats

MITRE ATLAS Matrix

Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Prompt Injection	LLM Plugin Compromise	LLM Jailbreak			Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction	Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity	Cost Harvesting
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection										External Harms	
	Poison Training Data	Phishing &											

<https://atlas.mitre.org/matrices/ATLAS>

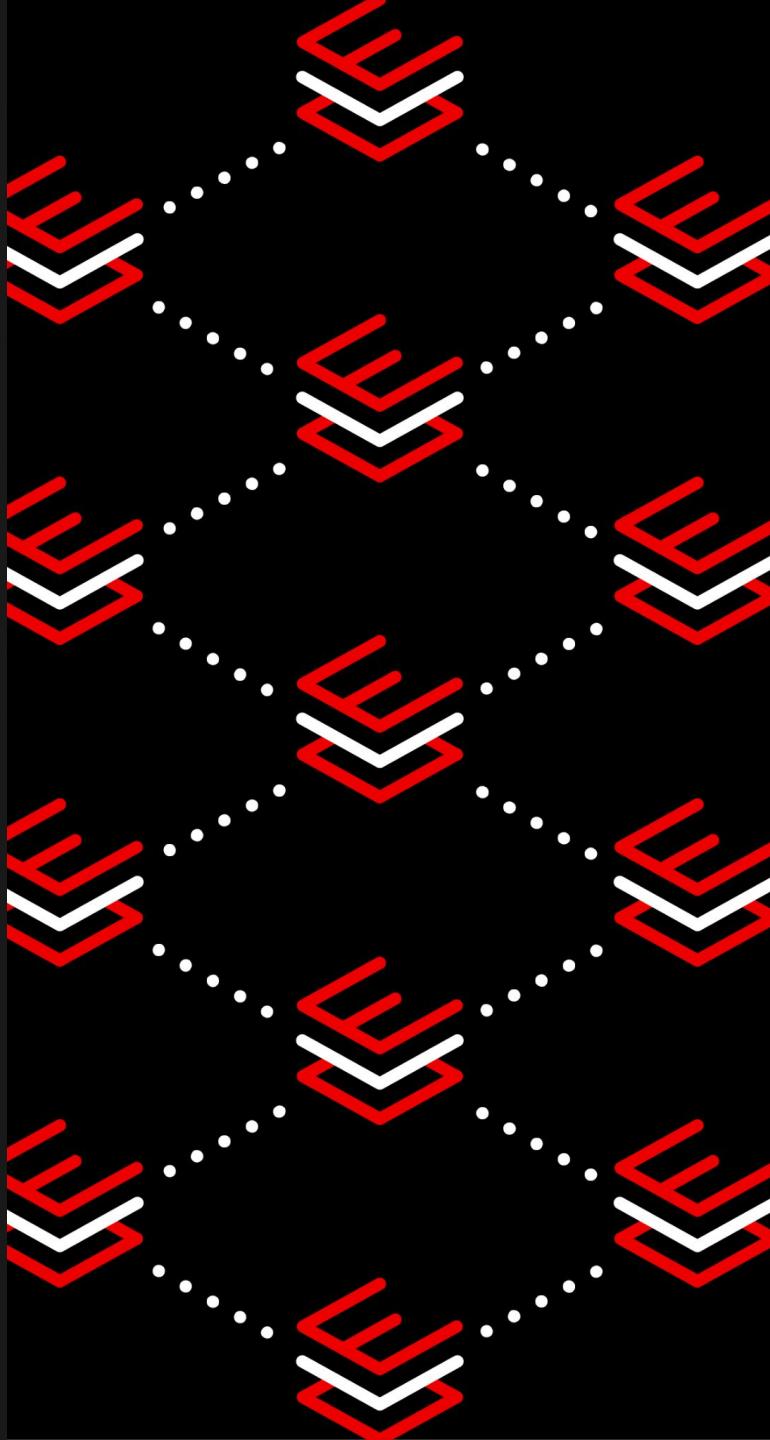
AI Security, Governance and Compliance Framework

Stage 0: Data Acquisition: Management and Training



Wilson Toh - Linkedin Security Articles

- AI Security: Navigating the Future of Cybersecurity – Leveraging AI to Manage AI Security
- BPFDoor: Mitigating Advanced Kernel-Level Threats



Introduction to AI ecosystem

The world changed in November 2022

ChatGPT woke the world up to the power of generative AI

The collage includes the following elements:

- TechFinative:** A screenshot of a website featuring a red robot arm interacting with a classical statue. Headlines include "Midjourney 5.2 delivers infinite zoom alongside powerful tools to add extra style and variety" and "OpenAI's ChatGPT Plus launches for \$24 per month".
- BUSINESS INSIDER:** A large image of a circuit board with the text "HOW AI IS CHANGING EVERYTHING" overlaid.
- CTECH by CALCALIST:** A news article with the headline "The AI revolution, especially generative AI, will completely change the world as we know it". It features a photo of two men, Eyal Agmoni and Takenaka Heizo, speaking at a conference.
- VentureBeat:** An article titled "The language revolution: How LLMs could transform the world" featuring a photo of a digital face.
- Harvard Business Review:** An article titled "Generative AI Will Change Your Business. Here's How to Adapt" by David C. Edelman and Mark Abraham. It includes a photo of a laptop and a sidebar for "Welcome to HBR!".

Text from the CTECH by CALCALIST article:

"The AI revolution, especially generative AI, will completely change the world as we know it."

Written by James O'Malley on Aug 09 2023

It's strange how fast technology changes from something that feels magical, to utterly mundane. Think about the last time you went on a plane. Did you take a moment to appreciate the extraordinary fact that somehow, a flying metal cube can project you across the planet of hours – something that would have been unimaginable not so long ago? Or, perhaps more likely, did you grumble about the ridiculousness of dropping a crumpled seat and bristle at the crying toddler kicking the center console?

Of course, like everyone, you did the latter. And now this same transition is also dropping advances in generative AI we've witnessed over the last few months.

Takenaka Heizo, Professor Emeritus at Keio University, was speaking with Eyal Agmoni, Founder of Chartered Group, during the Israel-Japan Conference in Tokyo

Meir Orbach 1102, 30.05.23

TAGS: Generative AI | AI | Israel-Japan Conference

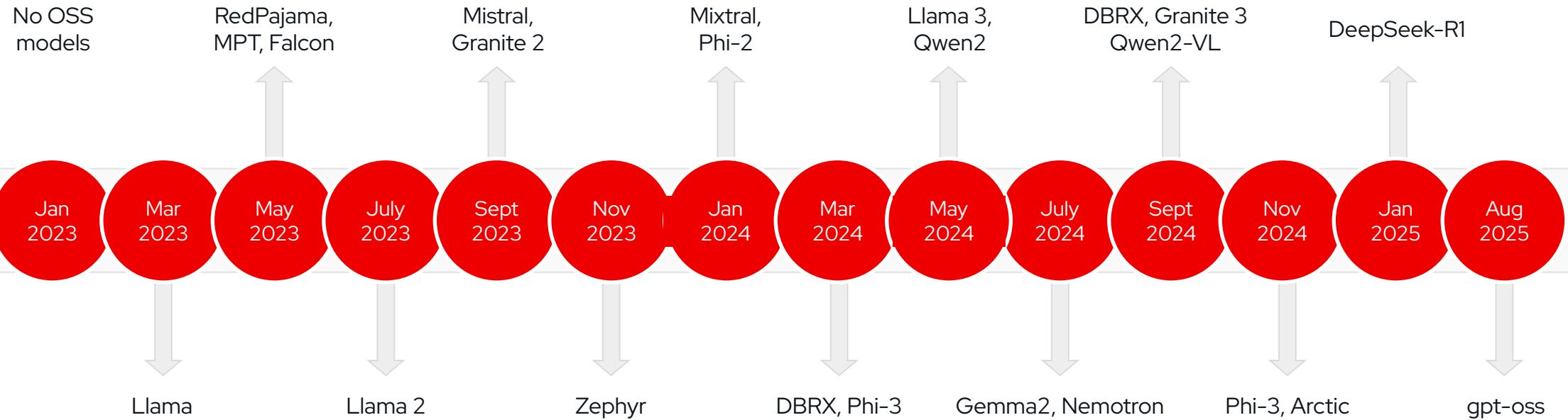
Eyal Agmoni and Takenaka Heizo (Shmuli Dudur)

"The AI revolution, and especially generative AI, will completely change the world as we know it," said Takenaka Heizo, Professor Emeritus at Keio University, during a talk with Eyal Agmoni, Founder of Chartered Group during the Israel-Japan Conference in Tokyo



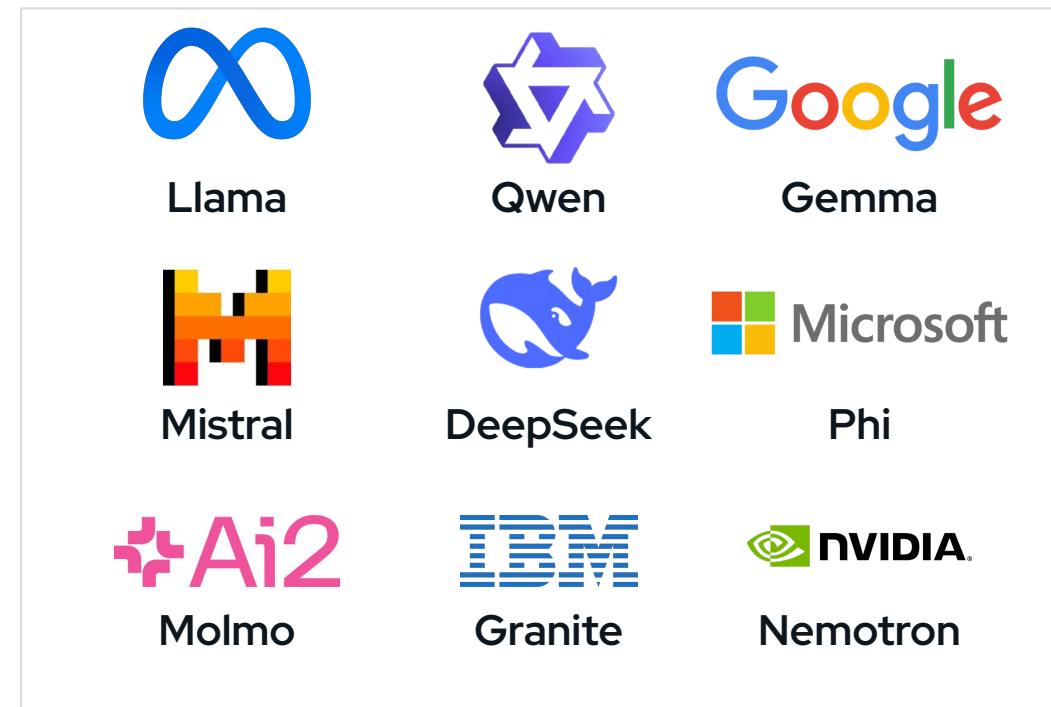
The power of open

There has been an explosion of capability in open models over the last 2 years

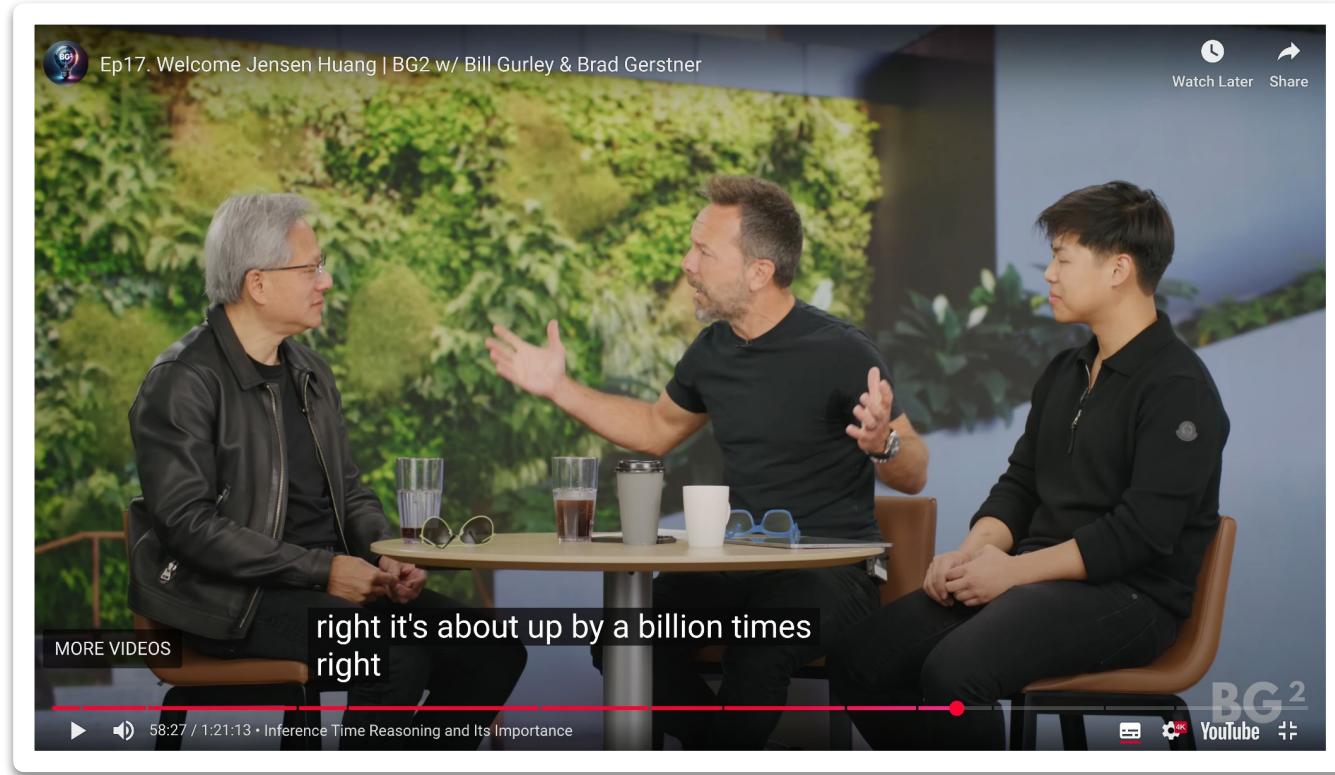


OpenSource LLMs

LLM	Strengths
LLama	Excels at advanced reasoning, complex problem-solving, and deep content generation.
Qwen	A strong contender for creative writing, fact-based queries, and information retrieval, showing high performance in these areas.
DeepSeek	Specifically designed for logical problem-solving and coding, making it a prime choice for complex programming tasks.
Mistral	Ideal for applications requiring speed, lower costs, and efficient deployment, making it suitable for resource-constrained environments.
Phi	Represents a class of smaller models that aim to deliver strong general capabilities - a balance of performance and efficiency is needed.
Granite	Trusted, built on enterprise data spanning legal, code, academia, finance.



In the beginning... It was all about
training and fine tuning



"[Inference] is about to go up by
a billion times."

Reasoning models are driving a 20x increase in inferencing

"I need to seat 7 people around a table at my wedding reception, but my parents and in-laws should not sit next to each other. Also, my wife insists we look better in pictures when she's on my left, but I need to sit next to my best man. How do I seat us on a round table? But then, what happens if we invited our pastor to sit with us?"

Traditional LLM



Reasoning Off

Tokens: 439

Reasoning Model



Reasoning models are driving a 20x increase in inferencing

"I need to seat 7 people around a table at my wedding reception, but my parents and in-laws should not sit next to each other. Also, my wife insists we look better in pictures when she's on my left, but I need to sit next to my best man. How do I seat us on a round table? But then, what happens if we invited our pastor to sit with us?"

Traditional LLM



Reasoning Off

Tokens: 439

Reasoning Model



Reasoning On

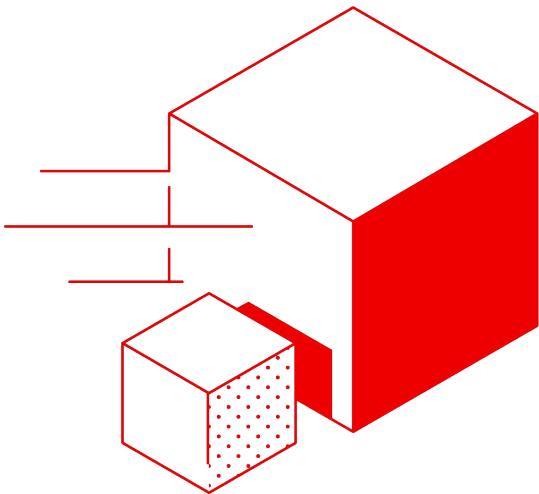
Tokens: 8559

**20x More Tokens
150x More Compute**



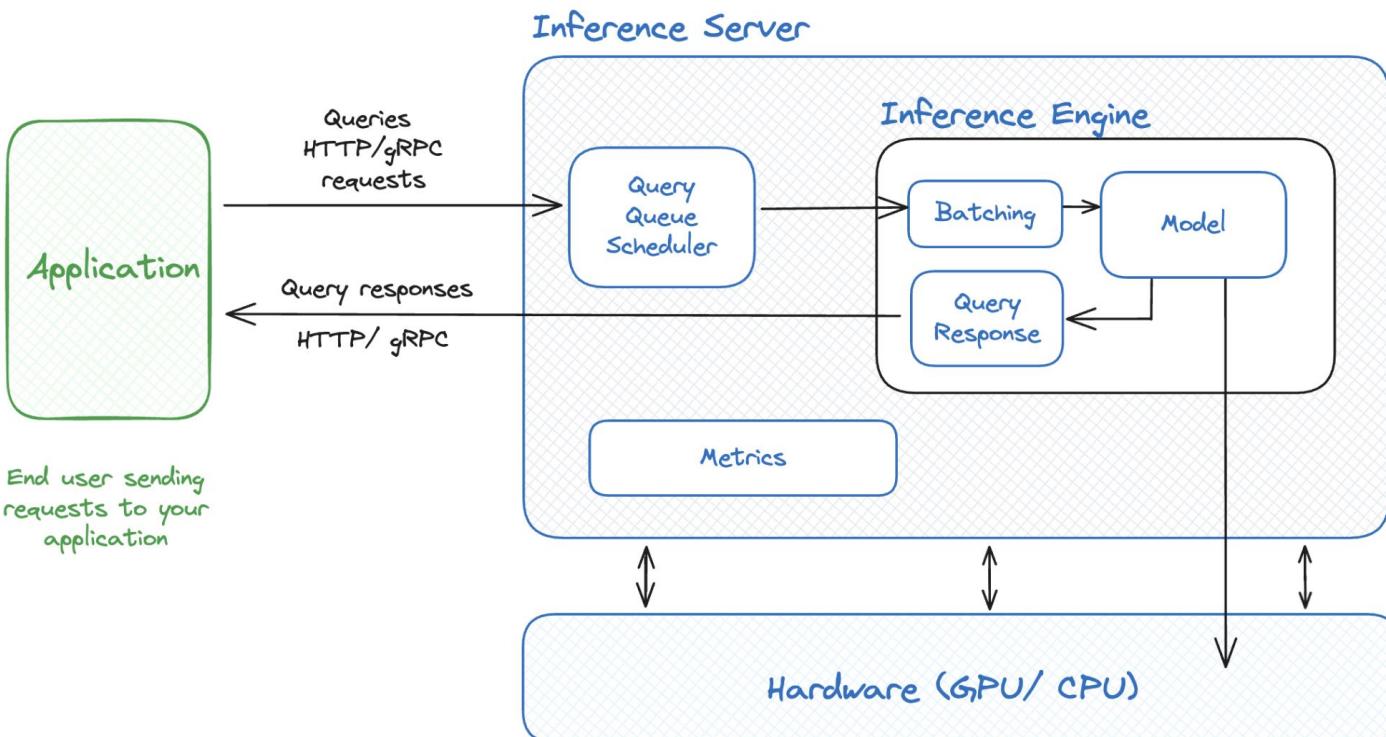
The value of inference

Inference is where the real world value happens



- ▶ **Powers the AI experience** where users interact with models
- ▶ **Can happen anywhere** across hardware, models, and the hybrid cloud
- ▶ **Creates value for AI initiatives** by delivering on desired business outcomes

LLMs need Inference Servers



Important metrics for LLM serving

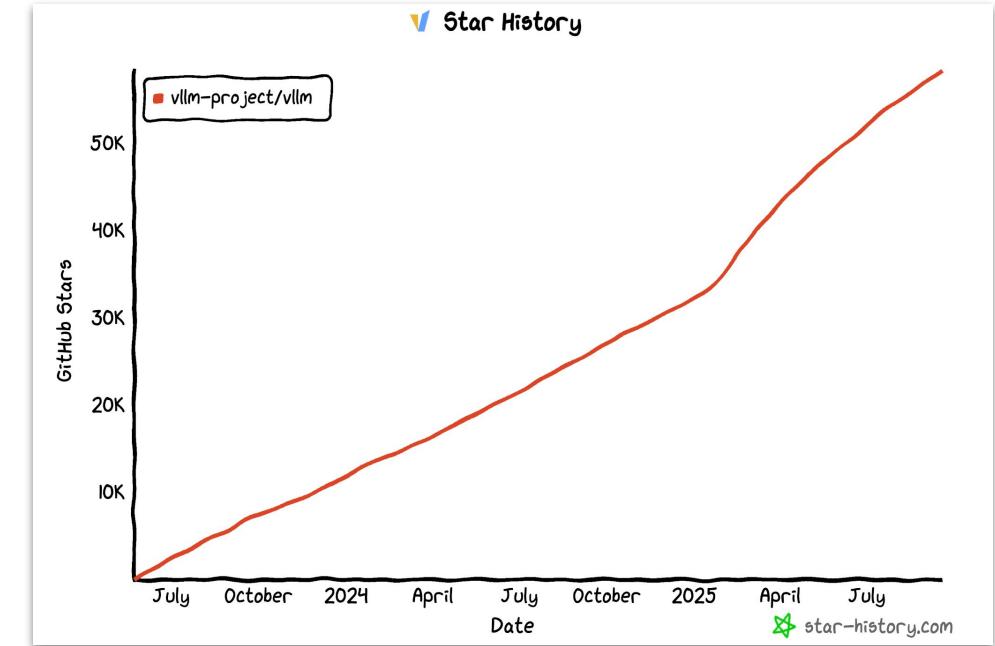
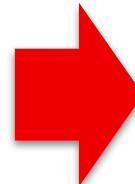
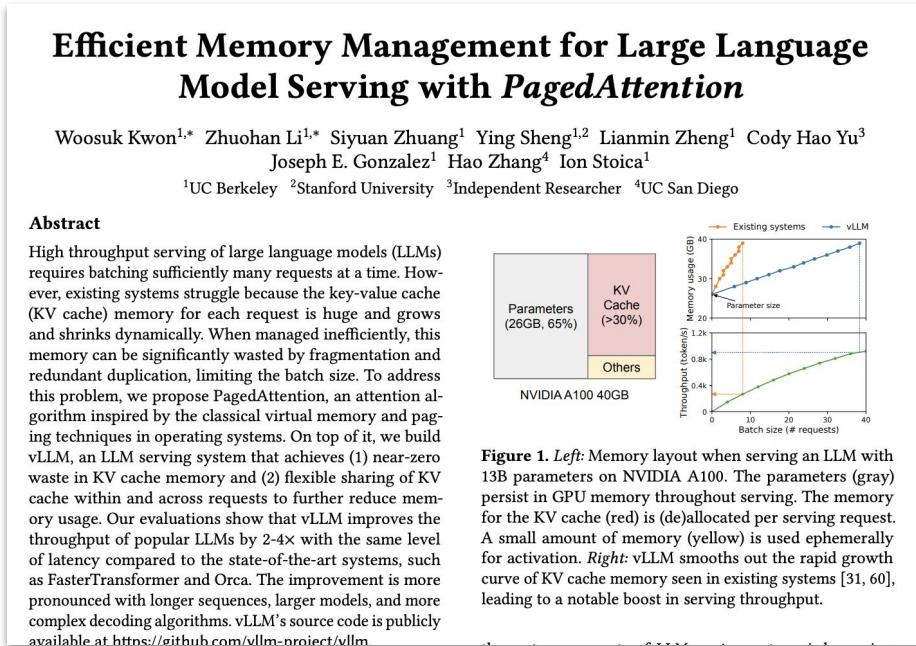
- **Throughput**: The number of requests from end users processed per second. Also measurable by the number of tokens generated by the model you are using in your application.
- **Latency**: The time taken to process a request from the time it is received to the time the response is sent. TTFT (Time to First Token) is a key metric for measuring the latency of LLMs.

vLLM's Goal

Build the **fastest** and **easiest-to-use** open-source LLM
inference & serving engine

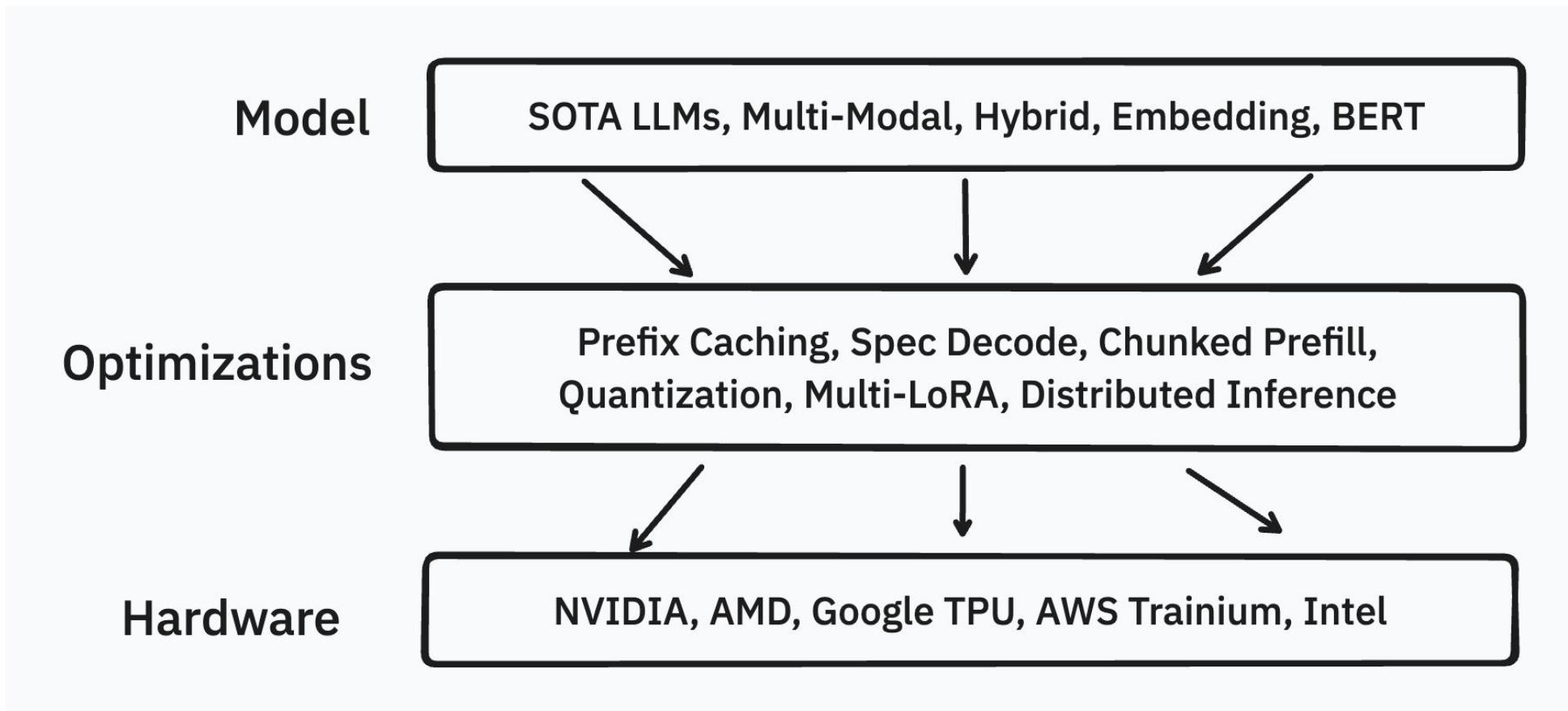
2 Year Journey Of vLLM

vLLM has rapidly evolved from a research project to the open source default



- **Pervasive** → 1+ million weekly installs; 58k GitHub stars
- **Explosive Growth** → Millions of deployed GPU hours per day
- **Vibrant Community** → 1500+ contributors

From this base, we have built...



Under the hood of vLLM

Bryon Baker



What Problem is vLLM Solving?

Production Inference Serving

- ▶ **Batch Size > 1 & Data Center Hardwares**
 - Not the same workload as on-device inference for a single user
- ▶ **How do you?**
 - Efficiently schedule requests into the next forward pass?
 - Manage KV cache context and runtime memory footprint?

Challenge 1: Batching

Naive Batching



GPU core ↑

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1				
S_2	S_2	S_2					
S_3	S_3	S_3	S_3				
S_4	S_4	S_4	S_4	S_4			

time →

GPU core ↑

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1	S_1	END		
S_2	END						
S_3	S_3	S_3	S_3	S_3	END		
S_4	END						

time →

Continuous Batching

GPU core ↑

T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1				
S_2	S_2	S_2					
S_3	S_3	S_3	S_3				
S_4	S_4	S_4	S_4	S_4			

time →

GPU core ↑

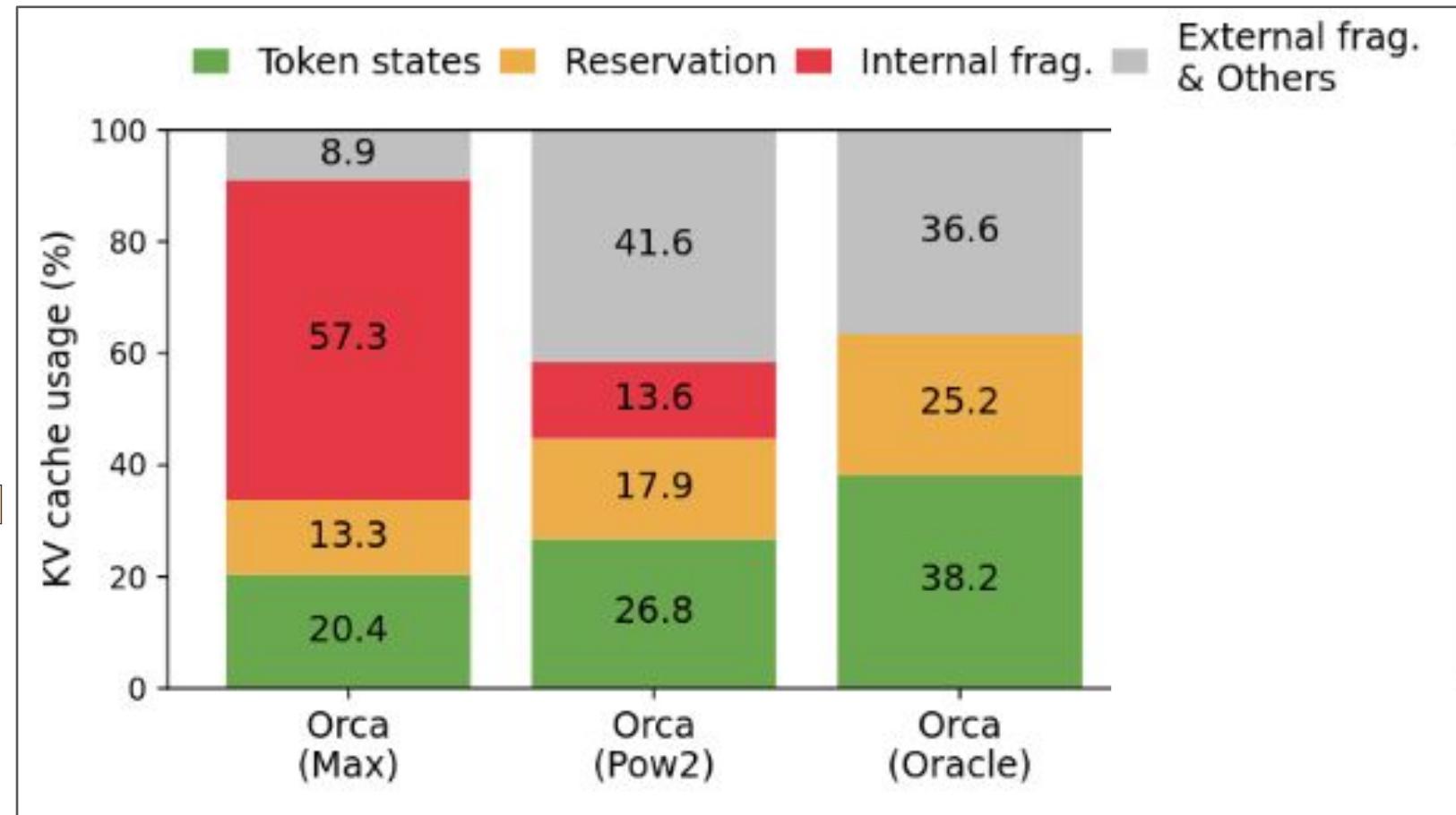
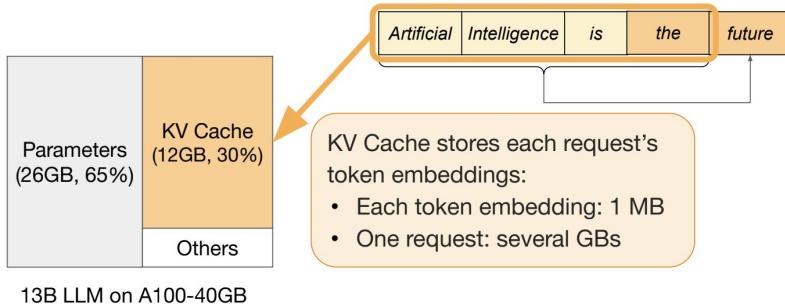
T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
S_1	S_1	S_1	S_1	S_1	END	S_6	S_6
S_2	END						
S_3	S_3	S_3	S_3	S_3	END	S_5	S_5
S_4	S_7						

time →



Challenge 2: KV Caching

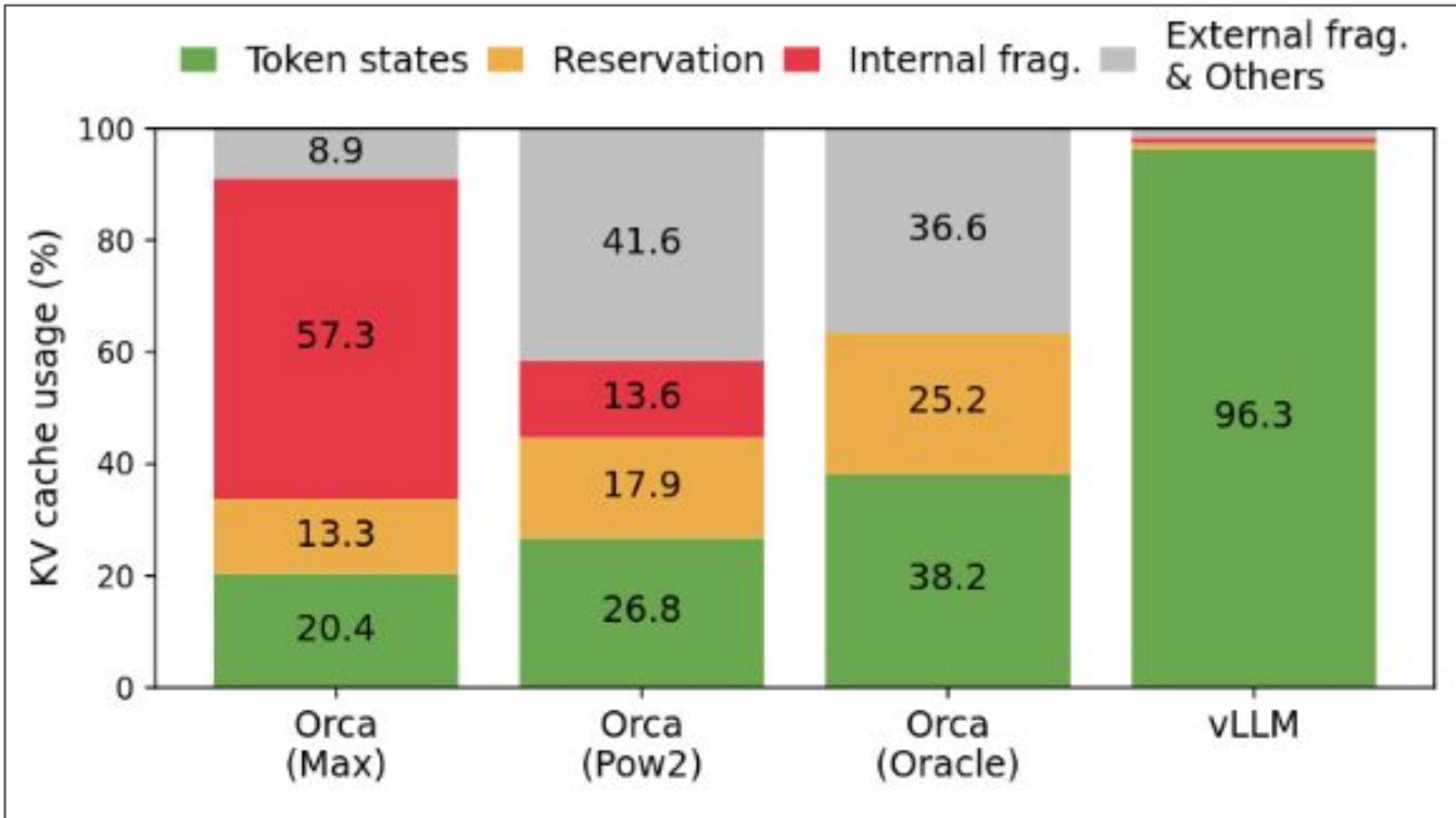
KV Cache: Caching Key and Value vectors in self-attention saves redundant computation and accelerates decoding - ***but takes up memory!***



Paged Attention



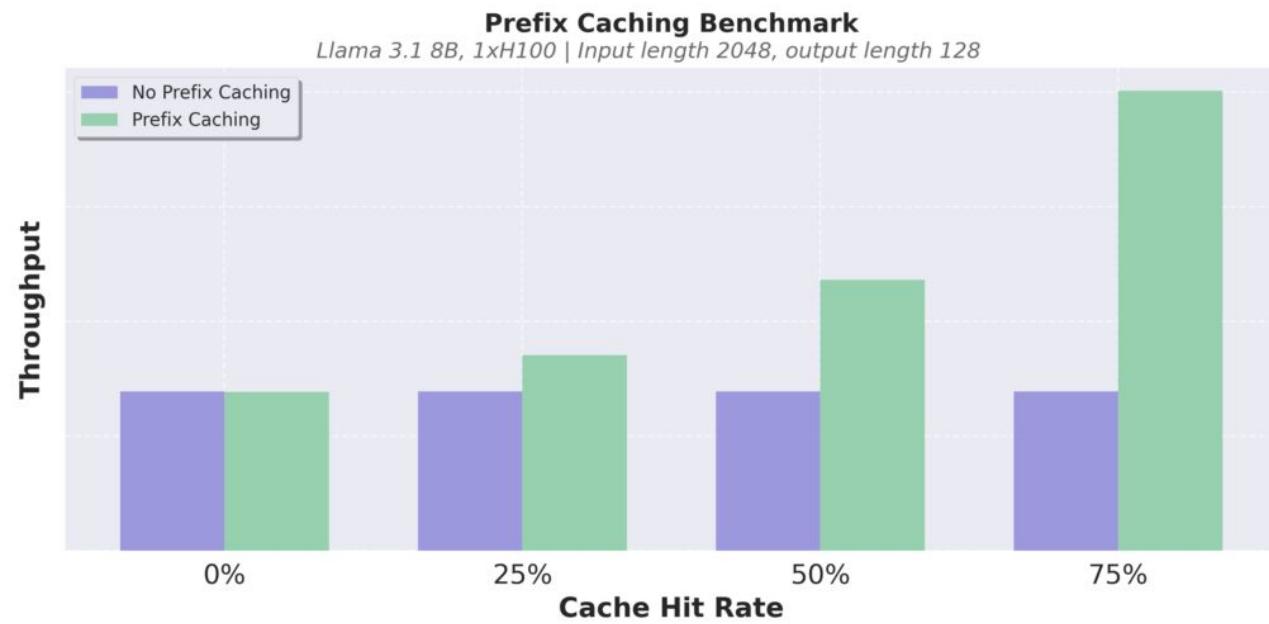
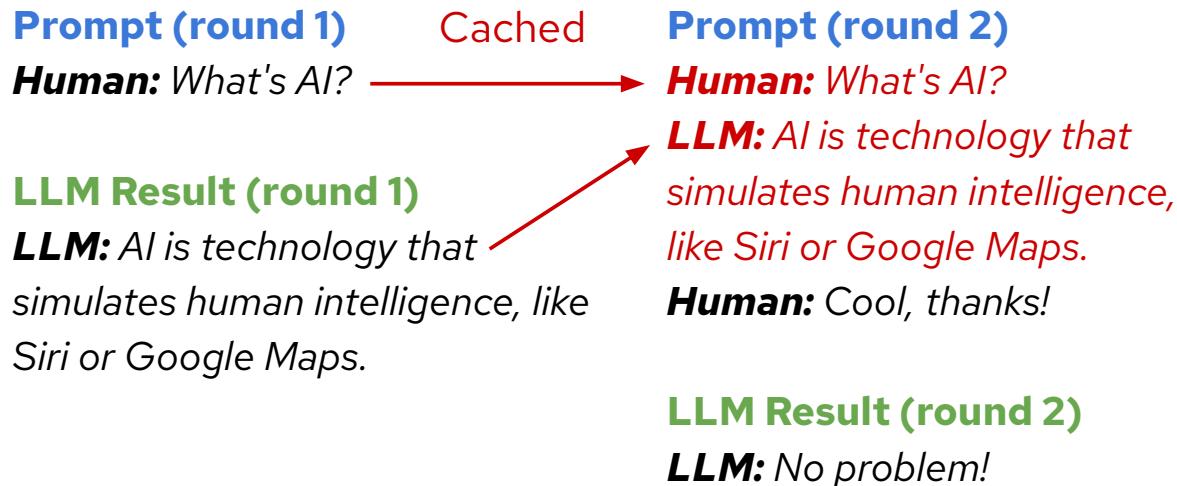
- *Virtual memory approach for KV Cache*
 - Analysis found that existing systems waste **60%-80%** of GPU memory on fragmentation and over-reservation.



Automatic Prefix Caching

Re-use KV cache blocks across requests! Improves time-to-first-token by skipping prefill

Example: Multi-turn conversation



Quantization in vLLM

Use low bit precisions (e.g., FP8, INT8, FP4) to store and compute faster

1. Weight Quantization

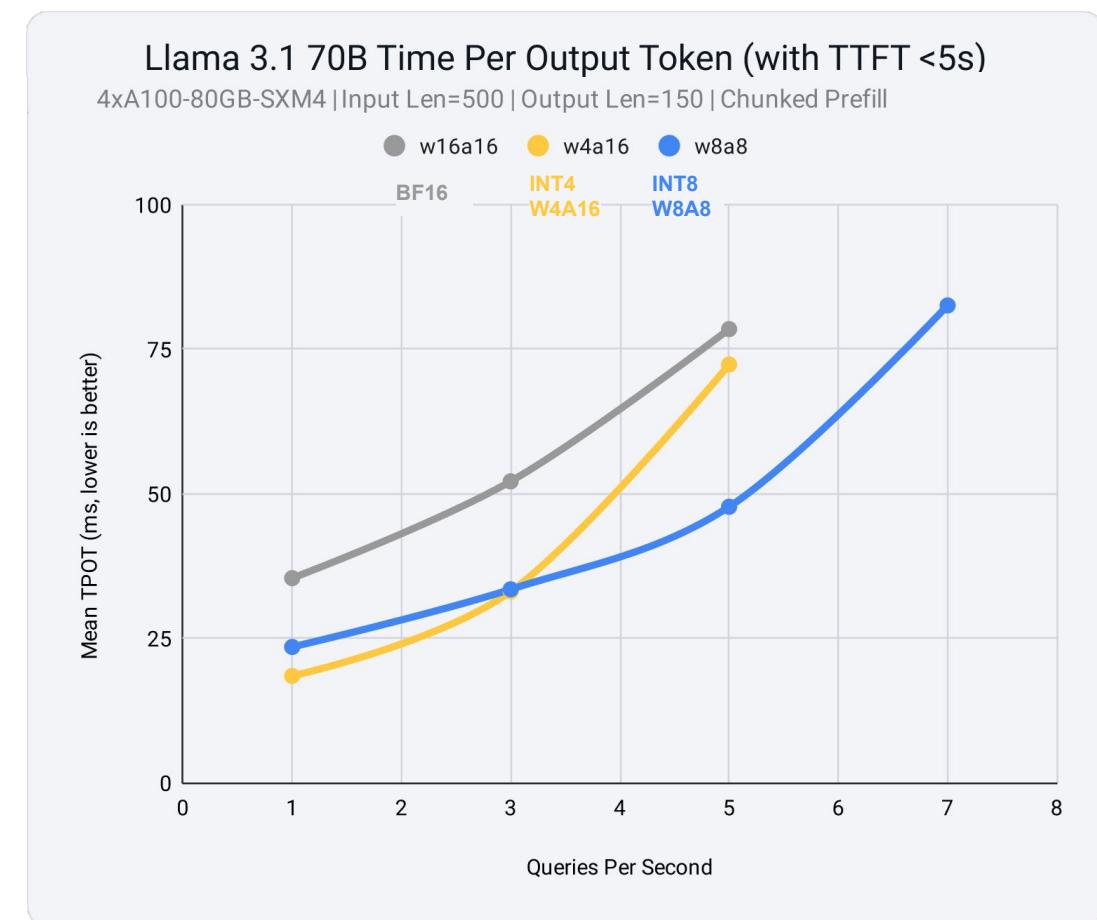
- Reduced storage & memory footprint
- E.g.) 100B parameter model takes up 200GB@BF16 but only 50GB@INT4

2. Activation Quantization

- Quantized weights and activations
- Faster linear layers by utilizing low precision tensor cores

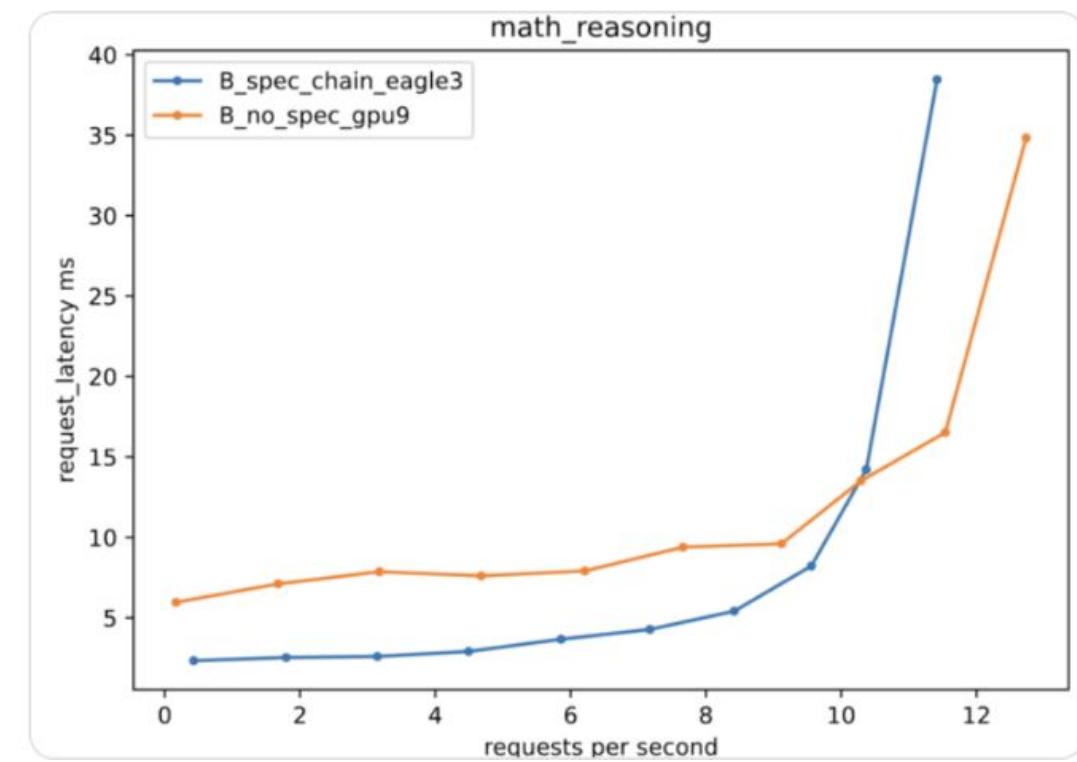
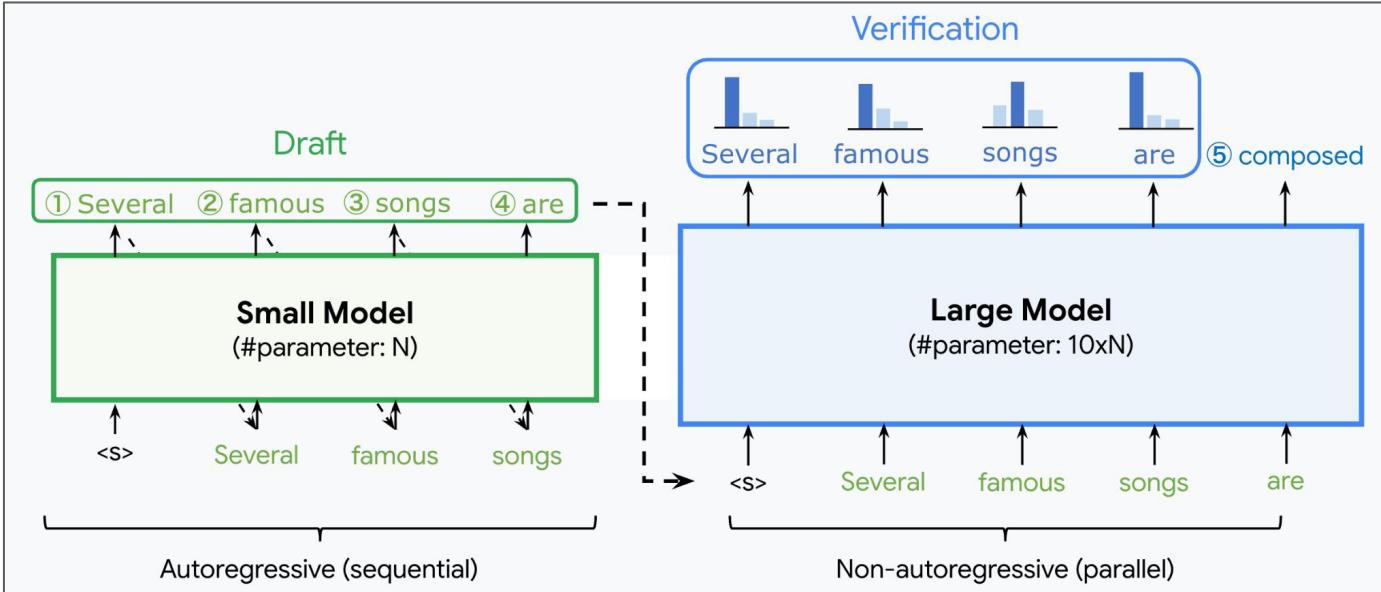
3. KV Cache Quantization

- Reduced KV cache footprint & faster attention
- Crucial for long context workloads



Speculative Decoding

Accelerate decoding phase with speculation - variety of methods: ngram, draft model, EAGLE, etc



vLLM Combines All Optimizations Together

Without Optimizations

Prompt

```
<SYSTEM> You are a helpful assistant. ...
Keep your answers precise and concise.
<USER> Generate a description for this item: ...
```

Output



Prompt

```
<SYSTEM> You are a helpful assistant. ...
Keep your answers precise and concise.
<USER> Generate a description for this item: ...
```

Output



vLLM Goes Distributed

Single Device

Single Host

Multi-Device

Multi-Host

Multi-Device

vLLM: The De Facto Open GenAI Inference Platform

vLLM has emerged as the Linux of GenAI Inference



Llama



Qwen



DeepSeek



Gemma



Mistral



Ai2



Microsoft



NVIDIA



IBM

vLLM



GPU



Instinct



TPU



Neuron



Gaudi



Spyre



Physical



Virtual



Private
Cloud



Public
Cloud



Edge

vLLM API - OpenAI-compatible server

A FastAPI-based server for online serving

Server

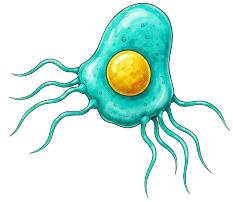
```
$ vllm serve meta-llama/Meta-Llama-3.1-8B
```

Client

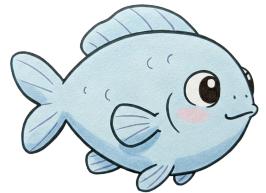
```
$ curl http://localhost:8000/v1/completions \
-H "Content-Type: application/json" \
-d '{
    "model": "meta-llama/Meta-Llama-3.1-8B",
    "prompt": "San Francisco is a",
    "max_tokens": 7,
    "temperature": 0
}'
```

Available routes are:

- /tokenize
- /detokenize
- /v1/models
- /v1/chat/completions
- /v1/completions
- /v1/embeddings
- /v1/audio/transcriptions
- /pooling
- /classify
- /score
- /rerank
- /metrics
- ...and more!



Prompt



Structured Output



RAG



Tools
MCP



Patterns

(chaining,
LLM-as-a-judge,
maker-checker,
agent-coder,
ReAct)

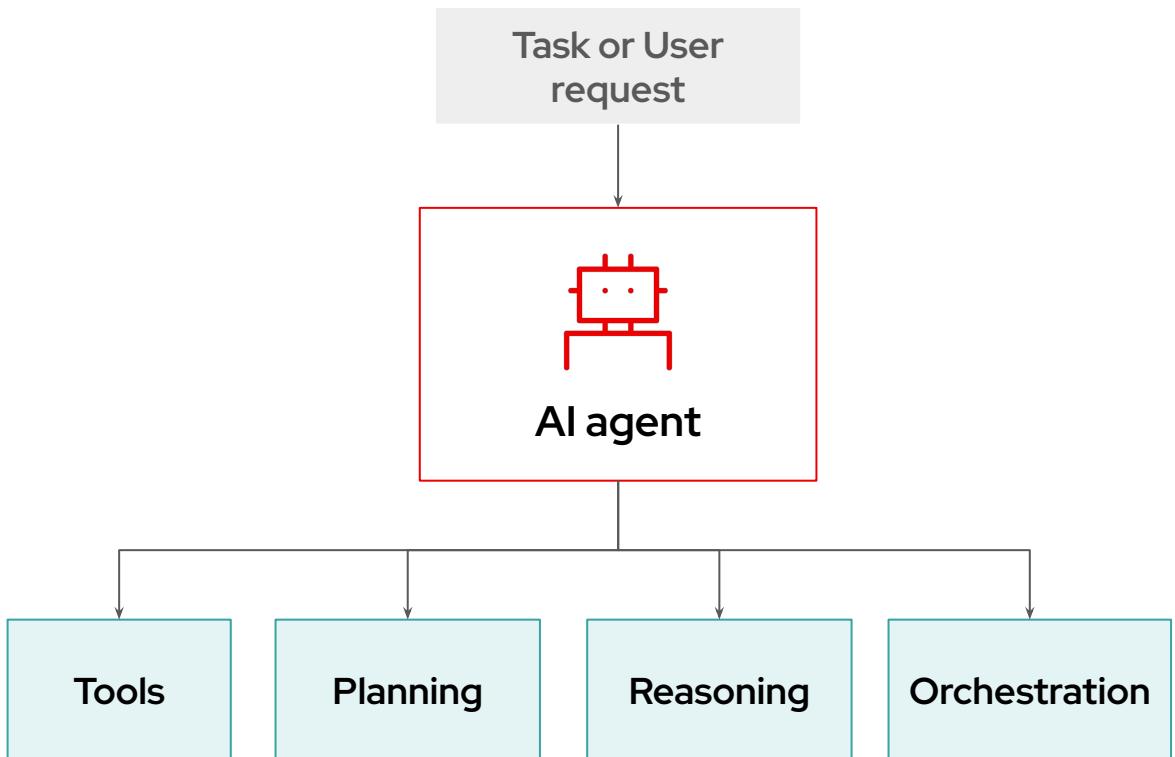


Agents

(planning,
orchestration,
retries, error-
correction)



The components of an AI Agent system



- ▶ **Tool Utilization:** Leverages external tools to gather data and perform tasks.
- ▶ **Planning and Execution:** Develops and executes multistep plans to achieve goals autonomously.
- ▶ **Reasoning:** Applies logic and contextual understanding to make informed decisions.
- ▶ **Orchestration:** Coordinates actions, tools, and agents to dynamically adjust and complete tasks.
- ▶ **Communication protocols:** enables the connections between the components.

Agentic Design Patterns



Think Step by Step & Reflect

Chain of Thought (CoT) Foundation

Chain of Thought prompting enables language models to break down complex problems into logical steps by encouraging step-by-step reasoning. This is typically achieved by adding the prompt "**Let's think step by step**" which forces the model to:

- Break down problems into manageable components
- Show intermediate reasoning steps
- Arrive at more accurate conclusions

"It was [discovered in July 2020](#) by early OA API & [AI Dungeon 2](#) users who found that GPT-3/'Dragon' would fail to solve most simple arithmetic problems like multiplication (as found by the [GPT-3 paper](#)), but could be coaxed into solving them by setting up a fictional dialogue between the player and a 'character' into solving it step by step. This discovery was widely discussed among GPT-3 enthusiasts, and highlighted on my GPT-3 page as a remarkable emergent capability of GPT-3 unlike GPT-2 or earlier models. It has been 'rediscovered' repeatedly since (by EleutherAI, and then multiple academic groups eg. as "[scratchpad](#)" or "[chain-of-thought](#)")." -Gwern

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

An example of chain of thought prompting being used to improve a models arithmetic skills. These types of problems are known to be problematic for even sophisticated language models, as they have a tendency to hallucinate responses to these forms of questions. [Source](#)



Think Step by Step & Reflect

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Answer: iPod

Standard



Reason only

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch



Act only

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced...

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ...

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Act 4: Finish[yes]



ReAct

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

Thought 3: Front Row is not found. I need to search Front Row (software) .

Act 3: Search[Front Row (software)]

Obs 3: Front Row is a discontinued media center software ...

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

Act 4: Finish[keyboard function keys]



ReAct (Reasoning and Action) builds upon CoT by combining reasoning with action-taking capabilities:

- Large Language Model (LLM) as the reasoning engine
- External tools for gathering information and performing actions
- Chain-of-Thought prompting for reasoning traces
- Action execution capabilities



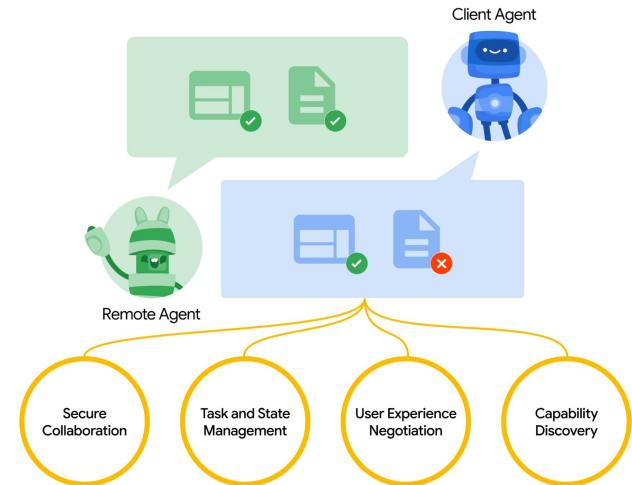
Agent2Agent Protocol (A2A)

How agents (will) communicate

With A2A, agents can:

- Discover each other's capabilities.
- Negotiate interaction modalities (text, forms, media).
- Securely collaborate on long running tasks.
- Operate without exposing their internal state, memory, or tools.

This will be used if you talk to an Agent aaS, e.g. current SaaS business will all provide an A2A endpoint for agents to use their service.



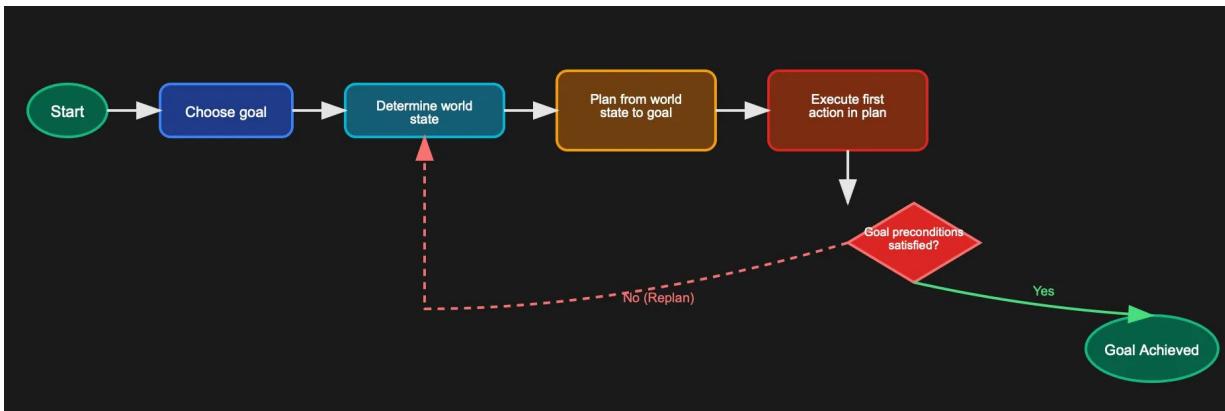
Google Cloud

Partners contributing to the Agent2Agent protocol



Example Agentic Frameworks

Framework / Toolkit	Main Idea
Llama Stack	Meta's ecosystem stack around Llama models
LangChain	Chain LLMs, tools, retrievers
LangGraph	Graph-based orchestration for LLMs
CrewAI	Multi-agent "crews" with roles
AutoGen (Microsoft)	Multi-agent conversation framework
Agno	Lightweight framework for building multi-agent systems
Pydantic	Data validation & typed models in Python



Embabel Agent Framework

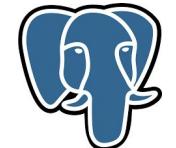
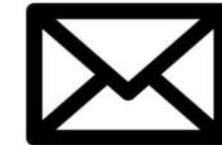
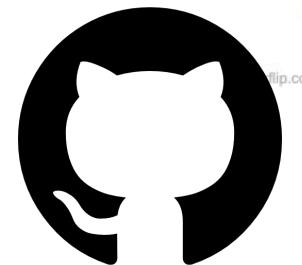
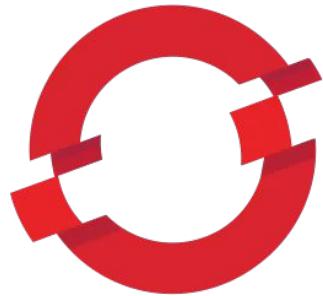
<https://github.com/embabel>



What are AI Tools?

Tools and Tool Calling are a capability that allows an LLM to interact with external systems to retrieve information, or take action.

Tool calling is essential to transform LLMs from being able to simply respond to a prompt, to taking action as an agent.



CALL ALL OF THE TOOLS!



Tools use Model Context Protocol (MCP)

The USB-C of AI applications

- ▶ A protocol, just like http
- ▶ introduced by Anthropic in November 2024
- ▶ an open standard, open-source framework
- ▶ Connect users, LLM, and “real-world”
- ▶ MCP provides a standardized way to connect AI models to different data sources and tools.
- ▶ Someone has built an MCP Server for that:
[punkpeye/awesome-mcp-servers](https://github.com/punkpeye/awesome-mcp-servers)

<https://modelcontextprotocol.io/introduction>

<https://github.com/modelcontextprotocol>

<https://docs.anthropic.com/en/docs/agents-and-tools/mcp>

<https://github.com/modelcontextprotocol/servers>

MCP is an open protocol that standardizes how applications provide context to LLMs. Think of MCP like a USB-C port for AI applications. Just as USB-C provides a standardized way to connect your devices to various peripherals and accessories, MCP provides **a standardized way to connect AI models to different data sources and tools.**

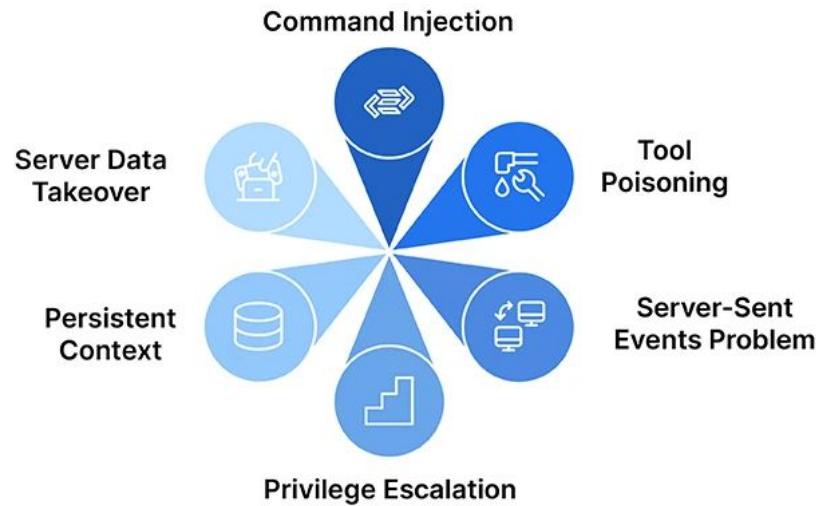


Version number here V00000



Is It Secure?

Vulnerabilities in MCP



MCP protocol has its own peculiar vulnerabilities, and the number is growing

Tool Poisoning Attacks, Prompt Injections, Rug Pull Attacks and much more

Attacks can exfiltrate sensitive data also hijack the agent's behavior even with a trusted infrastructure

Would you allow unauthorized users to access your backend data and services?

Strong Advice: Review the Security Report First



Why ToolHive?

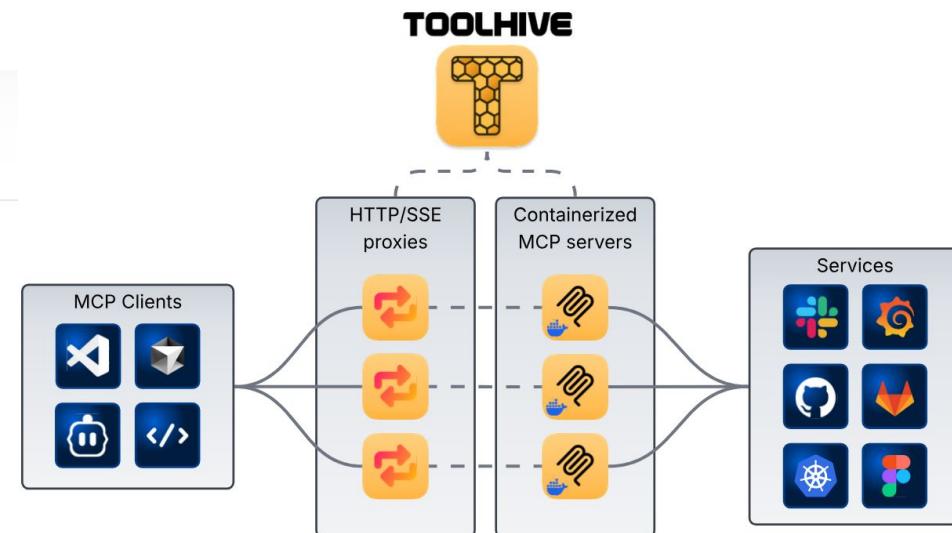
MCP Servers Made Simple and Secure

One-click deployment with enterprise-grade security built in. No complex setup. No risky config. Just safe and easy server management.

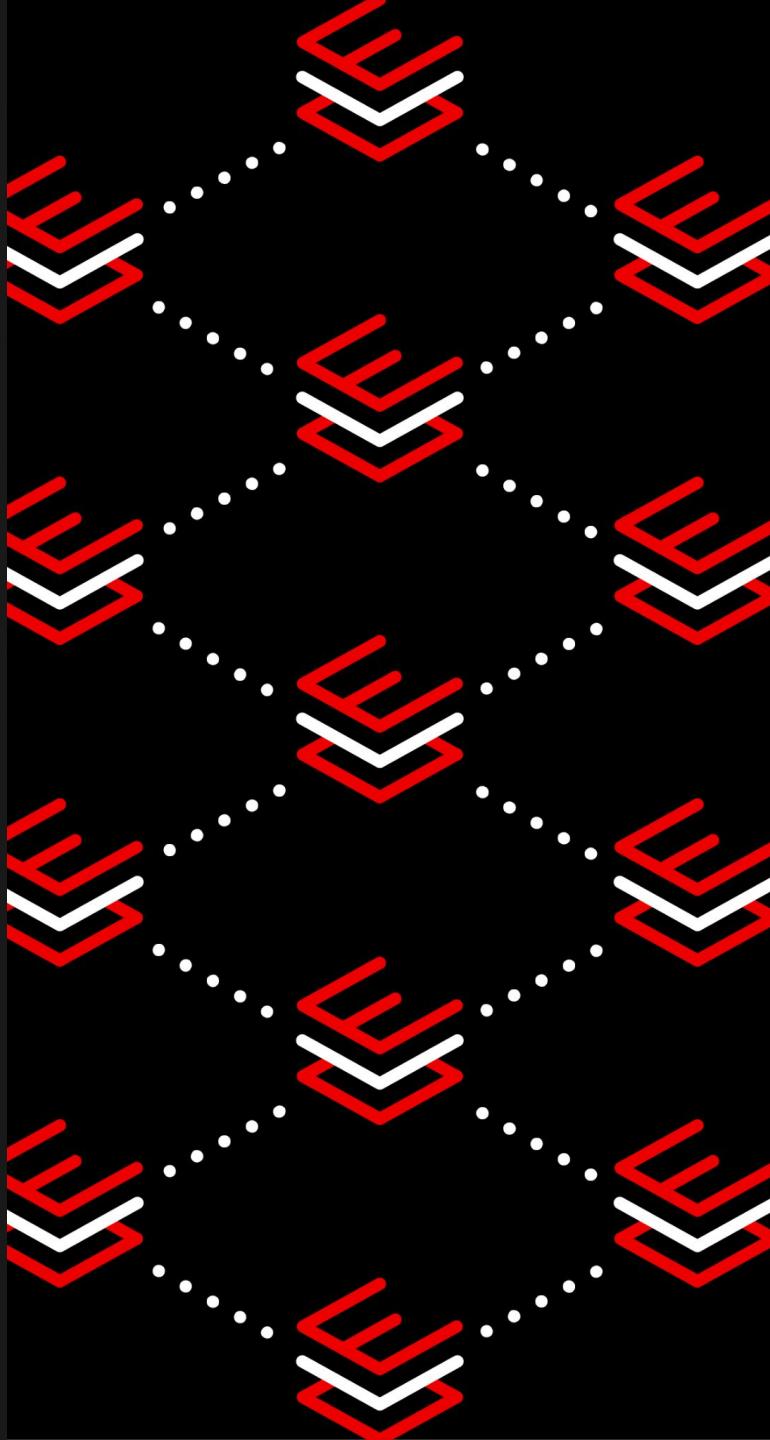
ToolHive's key features

Client Integration	Server Registry	Built-in Security
Connect your MCP Servers to your desktop clients automatically, no code configuration required.	Choose from a selection of verified MCP servers to reduce the security risks of running untrusted third-party code.	MCP servers run in isolated containers with software provenance verification to prevent malicious code execution.

<https://dev.to/stacklok/toolhive-making-mcp-servers-easy-secure-and-fun-7hi>



Red Hat
Power Hour Program



AI Ops

Incident Investigator Agent

Incident Investigator Agent

Demonstrates an autonomous AI Agent implemented with the Prompt Chaining and **Plan and Execute** pattern.

1. Provide a Goal
2. Devise a plan to achieve the Goal
3. Execute & iterate until Goal achieved
 - a. Call tools as needed
 - b. Check execution results
 - c. Determine if need for replanning or completion

Components used

- ▶ **Data Sources accessed via MCP**
 - Incident Details (Git)
 - Router Config
 - Traffic Monitoring System
 - Network Config
 - Knowledge Base
- ▶ **Automation access via MCP**
 - Ansible Automation Platform
- ▶ **AI**
 - OpenShift AI
 - Red Hat AI Inference Server
 - Llama Stack Server
 - Qwen Model
 - ToolHive for MCP



Incident Investigator Agent

Investigation Goal

Please investigate and analyze the cause of the latest network incident and report the results.

[Start Investigation](#)

🏁 Investigation Completed

📄 Final Report

The root cause of the StreamFlix video streaming service slowdowns during peak hours has been identified as misclassification of UDP-based streaming traffic in the QoS policies. Specifically, on `edge-router-01`, the configuration lacks a class-map for UDP-based streaming traffic, even though it defines a class-map `APP_STREAMFLIX_TCP` for TCP-based traffic. The StreamFlix service uses UDP protocol on port 443 from the IP range 198.51.100.0/24, but this traffic is not being properly classified and is instead being treated as unclassified traffic. This misclassification leads to the traffic being subject to the unclassified traffic policer that limits it to 1Gbps, while the actual traffic volume reaches 19,000 Mbps, resulting in significant packet drops and service degradation. The lack of a corresponding UDP-based class-map in the QoS policy configuration is the direct cause of this issue.

TPL001 Network Equipment Configuration Change Work Procedure Template

1. Work Overview

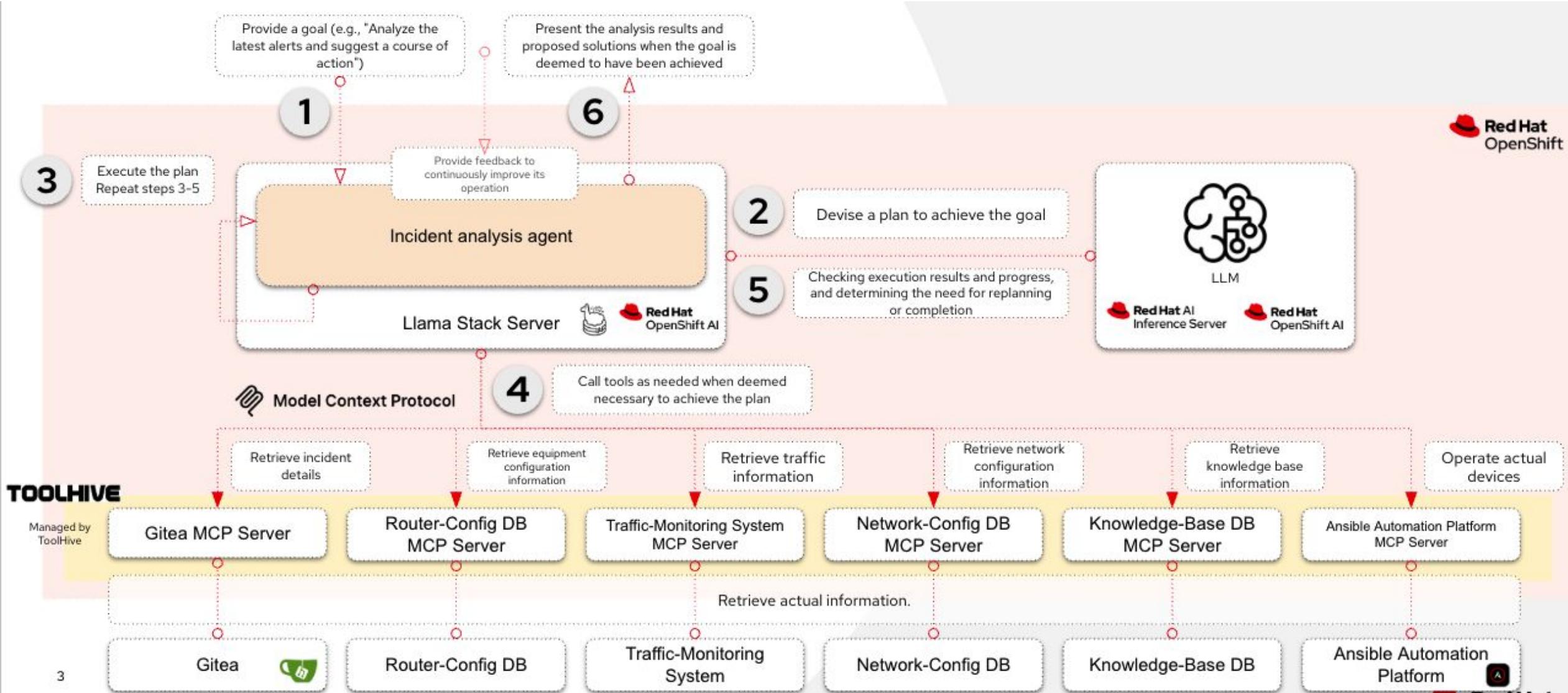
- **Purpose of Work:** To correct the QoS policy misclassification of UDP-based StreamFlix traffic on `edge-router-01` to prevent service degradation.
- **Description of Work:** Create a new class-map and policy-map entry to classify and prioritize UDP-based StreamFlix traffic, similar to the existing TCP-based classification.
- **Target Equipment:** `edge-router-01`
- **Date and Time of Work:** YYYY-MM-DD HH:MM - HH:MM (JST)

2. Related Information

- **Related Incidents:** Incident Report: 20250809_video_service_slowdown
- **Related Knowledge:** Misclassification of real-time streaming traffic as unclassified in QoS policies causes severe network latency and service degradation.
- **Related Operation Manuals:** MAN001 Cisco IOS-XE QoS Configuration Manual

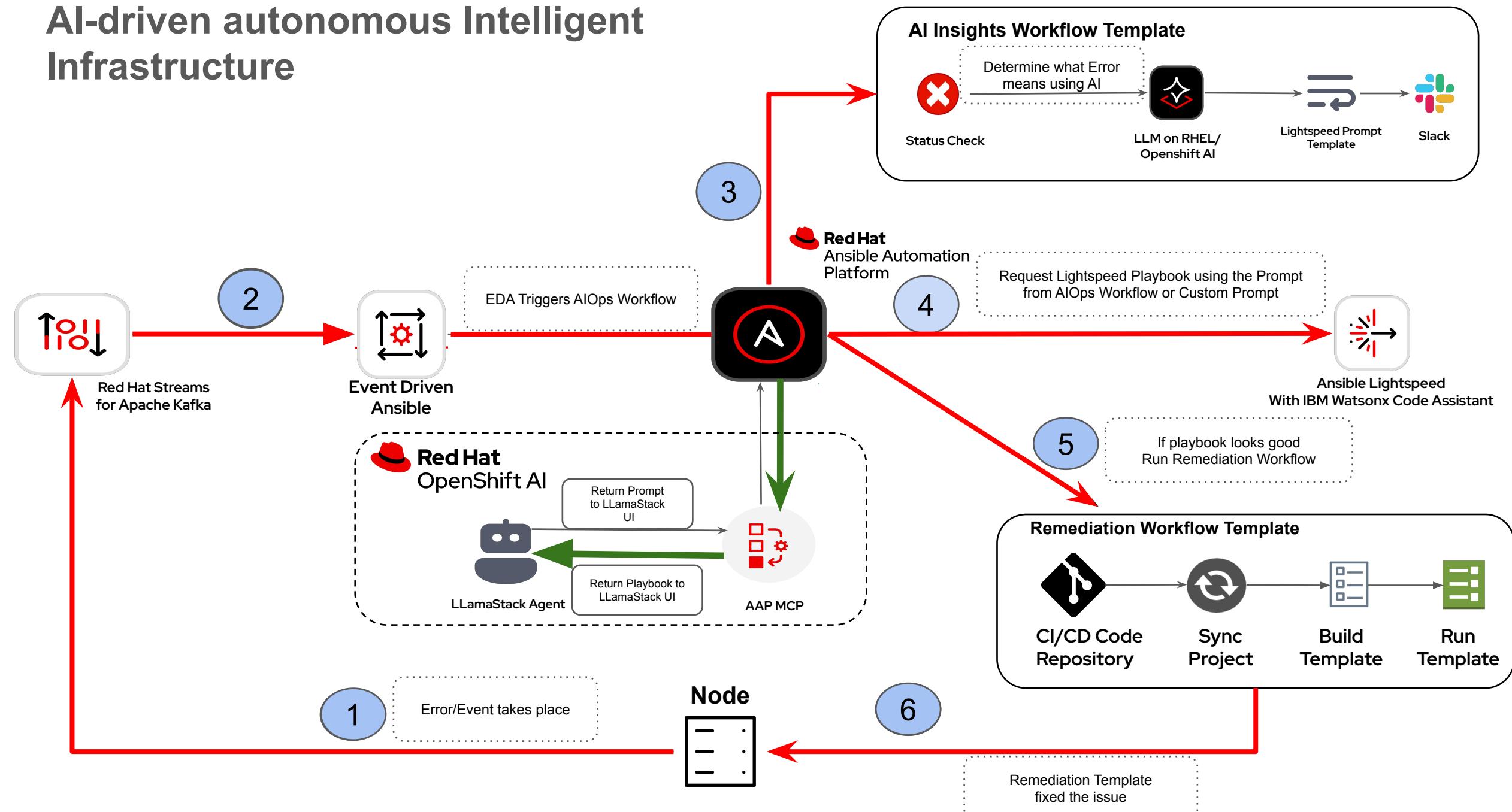
3. Pre-check

- Have you obtained the current configuration of the target equipment?
- Is the current status of the target equipment (CPU, memory, traffic) normal?
- Has the scope of impact from the change been identified?
- Have you notified the relevant parties in advance?



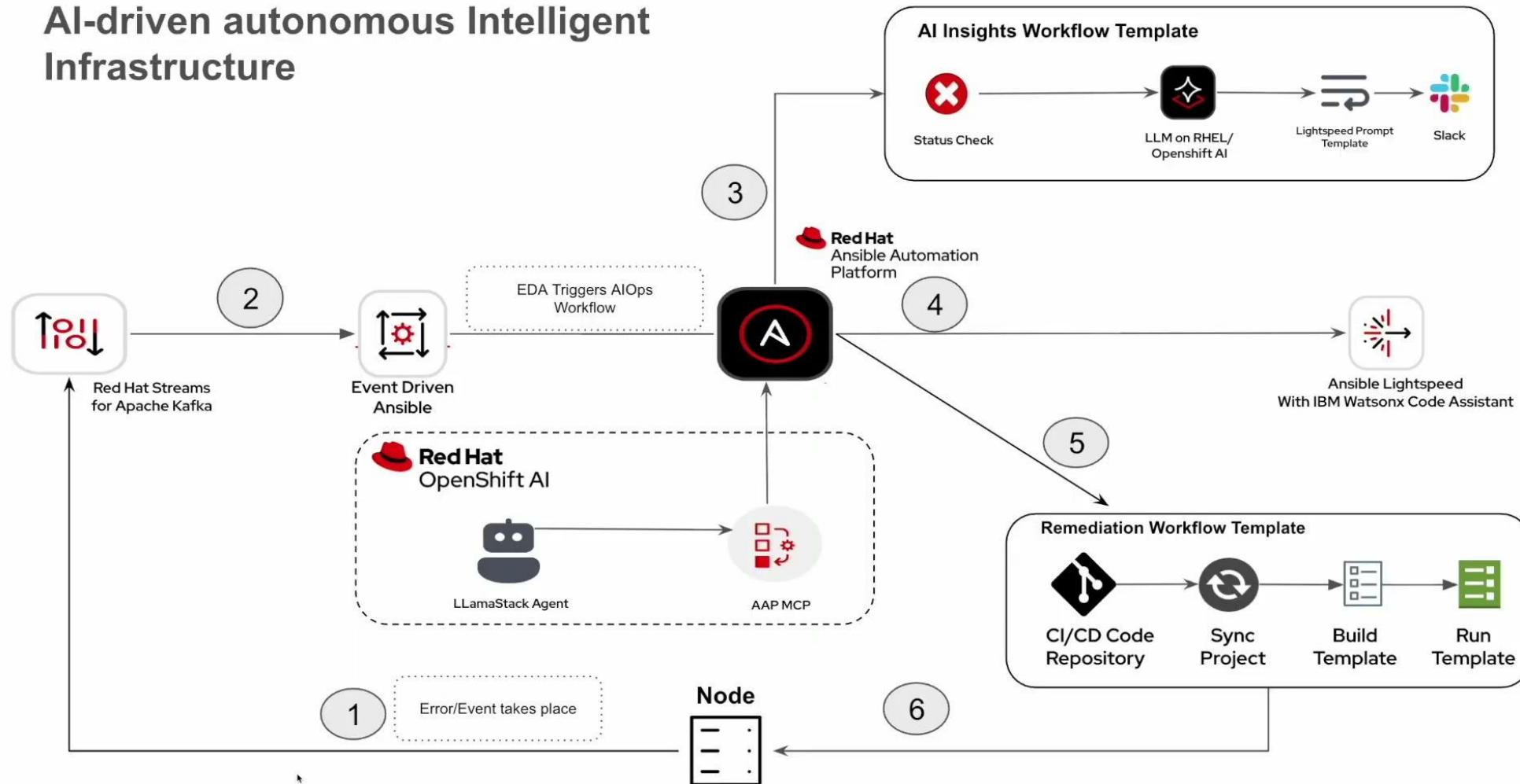
AI Insights with Actional Automation

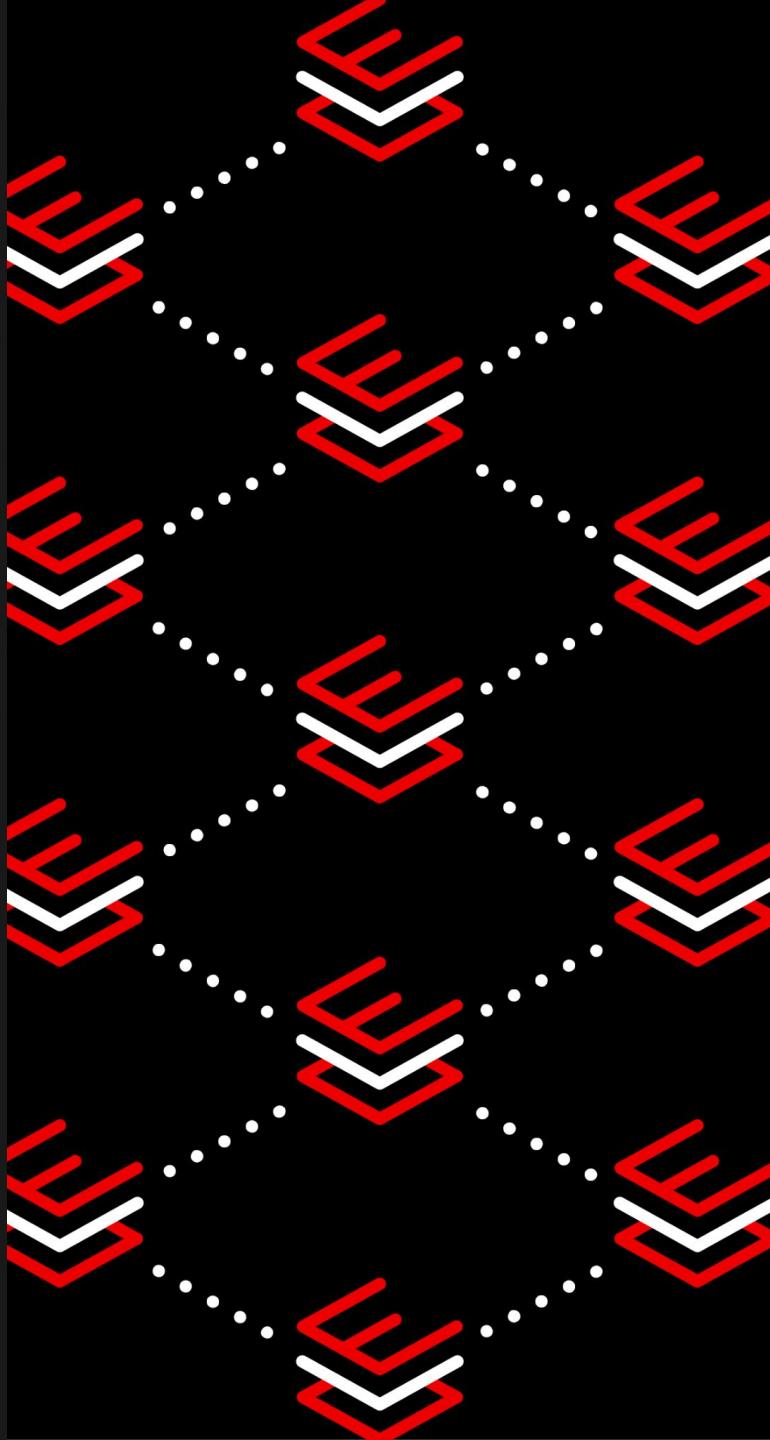
AI-driven autonomous Intelligent Infrastructure



Solution Demo

AI-driven autonomous Intelligent Infrastructure





How can you
get started

Podman AI - Get Hands on with these Recipes

The screenshot shows the Podman AI web interface. On the left is a sidebar with icons for Dashboard, Containers, Pods, Images, Volumes, Kubernetes, Extensions, AI Lab (which is selected), and Bootable Containers. The main area displays two recipe cards:

RAG Chatbot
Natural Language Processing
Summary Running

RAG (Retrieval Augmented Generation) Chat Application

This demo provides a simple recipe to help developers start to build out their own custom RAG (Retrieval Augmented Generation) applications. It consists of three main components; the Model Service, the Vector Database and the AI Application.

There are a few options today for local Model Serving, but this recipe will use '[llama-cpp-python](#)' and their OpenAI compatible Model Service. There is a Containerfile provided that can be used to build this Model Service within the repo, '[model_servers/llamacpp_python/base/Containerfile](#)'.

In order for the LLM to interact with our documents, we need them stored and available in such a manner that we can retrieve a small subset of them that are relevant to our query. To do this we employ a Vector Database alongside an embedding model. The embedding model converts our documents into numerical representations, vectors, such that similarity searches can be easily performed. The Vector Database stores these vectors for us and makes them available to the LLM. In this recipe we can use [chromaDB](#) or [Milvus](#) as our Vector Database.

Our AI Application will connect to our Model Service via its OpenAI compatible API. In this example we rely on [Langchain's](#) python package to simplify communication with our Model Service and we use [St reamlit](#) for our UI layer. Below please see an example of the RAG application.

Node.js Function calling
Natural Language Processing
Summary Running

Function Calling Application

This recipe helps developers start building their own AI applications with function calling capabilities. It consists of two main components: the Model Service and the AI Application.

There are a few options today for local Model Serving, but this recipe will use '[llama-cpp-python](#)' and their OpenAI compatible Model Service. There is a Containerfile provided that can be used to build this Model Service within the repo, '[model_servers/llamacpp_python/base/Containerfile](#)'.

The AI Application will connect to the Model Service via its OpenAI compatible API. The recipe relies on [Langchain's](#) Typescript package to simplify communication with the Model Service and [langgraph.js](#) to enable the LLM to call functions. It uses [fastify](#) as the backend-server and chart.js to plot the weather data returned. You can find an example of the chat application below.

Hands on Try out MCP

1. Download an MCP client examples

Claude - <https://claude.ai/download>

Goose - <https://block.github.io/goose/docs/quickstart/>

MCP-CLI - <https://github.com/chrishayuk/mcp-cli>

2. Use a local or remote model

Ollama - <https://ollama.com/>

vLLM - https://docs.vllm.ai/en/latest/getting_started/quickstart.html

3. Setup MCP with tool

mcp-cli --model gpt-oss:20b --server sqlite

<https://github.com/rlopez133/mcp>

4. Let's Look at how an MCP Server is built

- https://github.com/pavanjava/kafka_mcp_server/tree/main
- <https://github.com/modelcontextprotocol/servers>

Red Hat AI platforms



Red Hat Enterprise Linux AI

Foundation model platform for developing, testing, and running Granite family LLMs

- ▶ Provides a simplified approach to get started with generative AI that includes open source models
- ▶ Makes AI accessible to developers and domain experts with little data science expertise
- ▶ Provides the ability to do training & inference on individual production server deployments



Red Hat OpenShift AI

Integrated MLOps platform for model lifecycle management at scale anywhere

- ▶ Provides support for both generative and predictive AI models with a BYOM approach
 - ▶ Includes distributed compute, collaborative workflows, model serving and monitoring
- ▶ Offers enterprise MLOps capabilities and the ability to scale across hybrid-clouds
- ▶ Includes Red Hat Enterprise Linux AI, including the Granite family models



Model development

Bring your own models or customize Granite models to your use case with your data. Supports integration of multiple AI/ML libraries, frameworks, and runtimes.



Model serving and monitoring

Deploy models across any OpenShift footprint and centrally monitor their performance.



Lifecycle management

Expand DevOps practices to MLOps to manage the entire AI/ML lifecycle.



Resource optimization and management

Scale to meet workload demands of gen AI and predictive models. Share resources, projects, and models across environments.

Integrated AI platform

Create and deliver gen AI and predictive models at scale across hybrid cloud environments.



Llama Stack

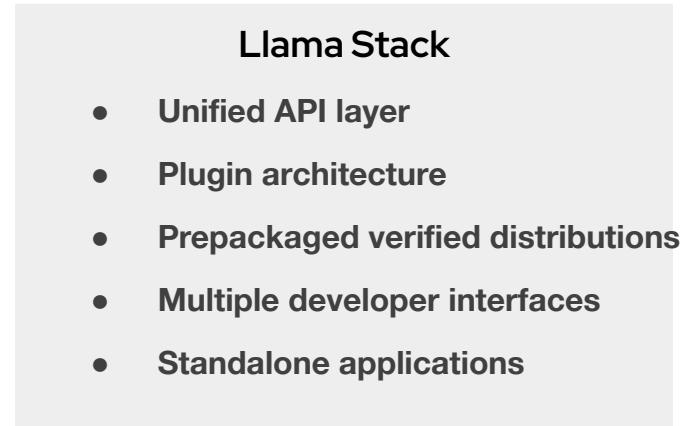
<https://github.com/llamastack/llama-stack>



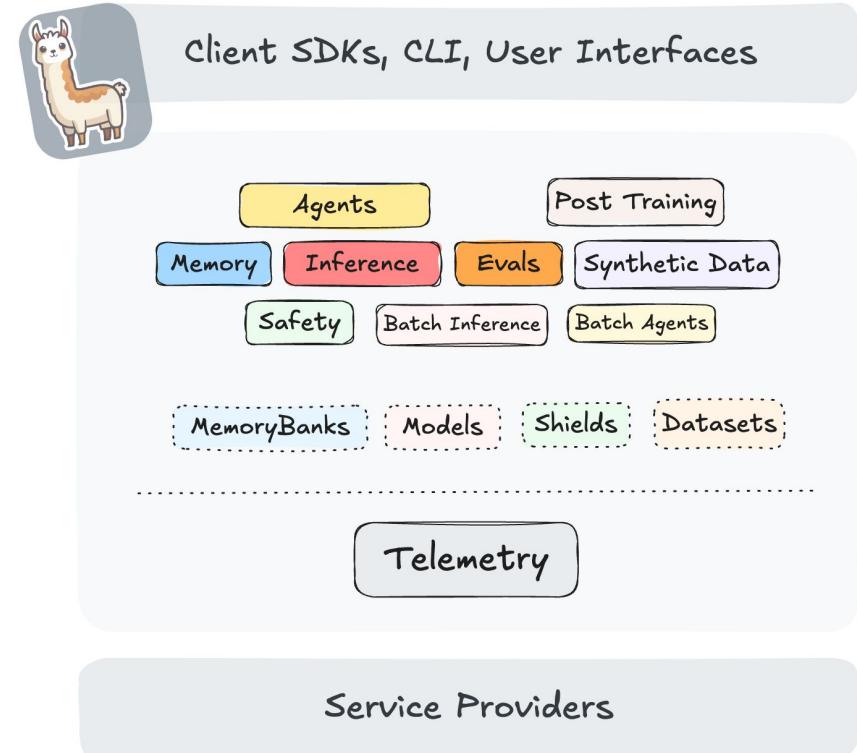
<https://youtu.be/MrAx-rV36I8>



<https://youtu.be/nevwVVSSzGw>

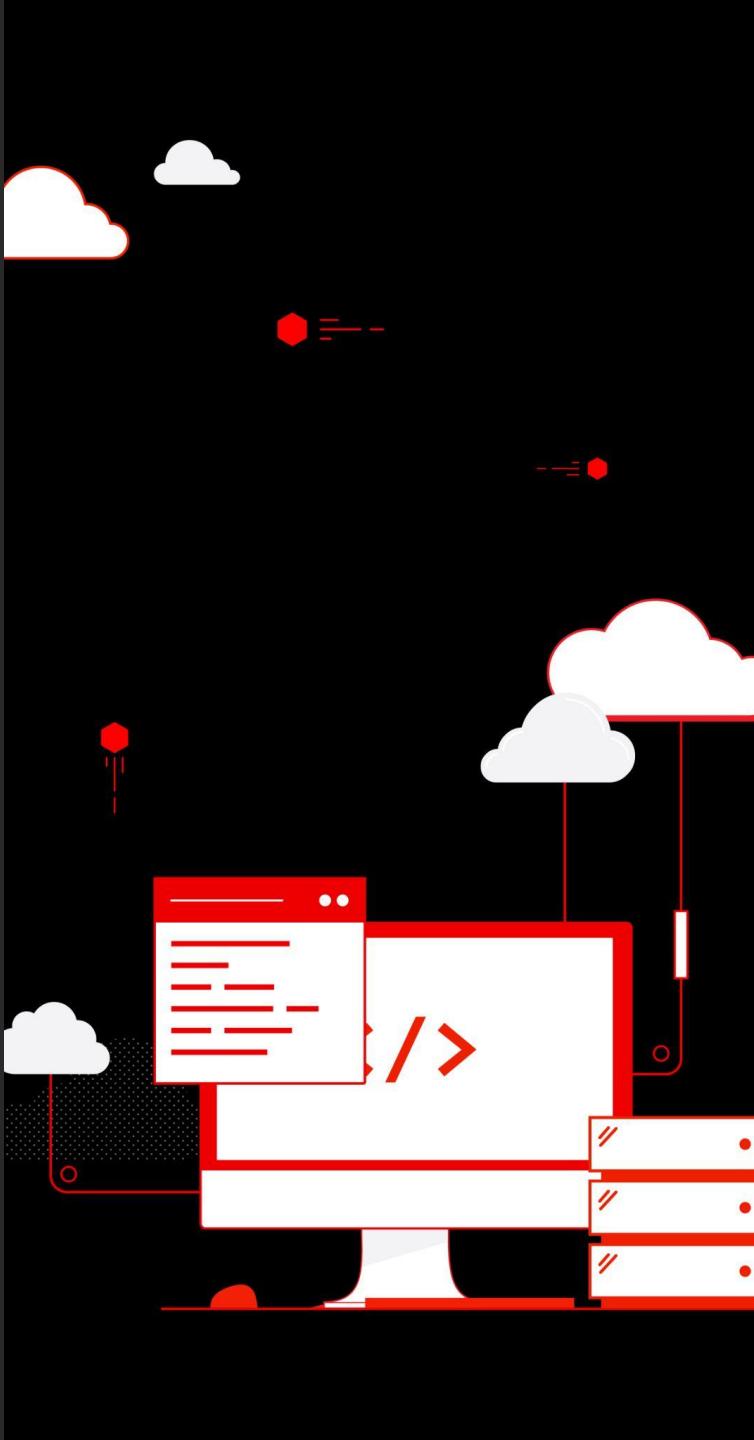


<https://youtu.be/K9JW3DzURh0>



Areas of focus for 2025 and Beyond

 Models	Multi-lingual; Multimodal Reasoning models	Granite Models 3rd party open models	Model Validation Model Optimizations
 Aligning to Enterprise AI Data ingestion and chunking Synthetic Data Generation Training Evaluation RAG	 Model Serving Optimizations Service Level Objectives Massive scaling Model Hub and Registry	 AI Agents and Apps API: Agents, Inference, datasets, safety, vector_io, telemetry Tool/Function Calling	
 LLMOps MLOps	Observability: TFT, ITL, TPS, Metrics, Logs, Traces Costs: \$/million i/o tokens	Security: Provenance, Signing, Encryption Governance: guardrails	
 Hybrid Cloud  Open Source  Partners	Hardware Accelerators, OEMs, Clouds Models, Tools, Frameworks Model Providers, OEMs, CSPs, ISVs		



Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.



[linkedin.com/company/red-hat](https://www.linkedin.com/company/red-hat)



[facebook.com/redhatinc](https://www.facebook.com/redhatinc)



[youtube.com/user/RedHatVideos](https://www.youtube.com/user/RedHatVideos)



twitter.com/RedHat