



Hewlett Packard
Enterprise

SECURE GENAI ADOPTION FOR ALL!

Tom Phelan, Glyn Bowden, Saad Zaher

February 21, 2024

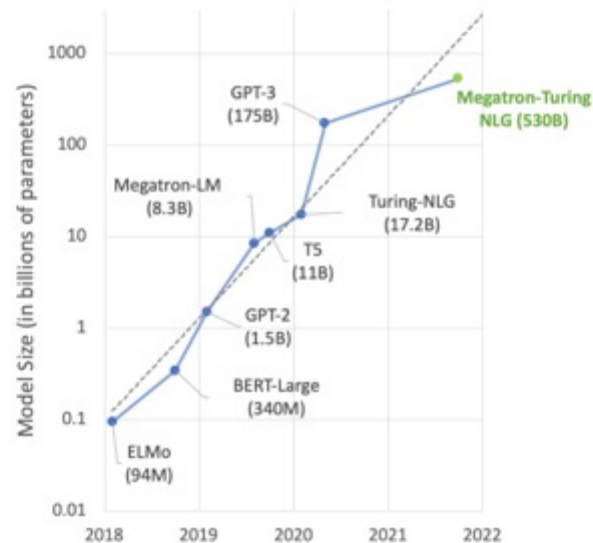
The world before November 2022

Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model

Oct 11, 2021

By Paresh Kharya and Ali Alvi

+9 Like Discuss (1)

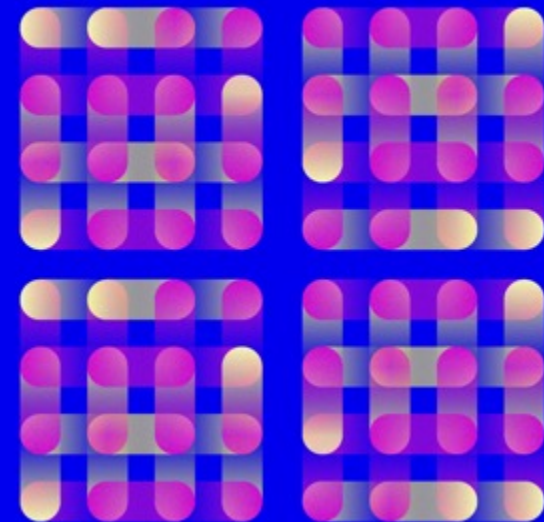


Top 9 Use Cases of Computer Vision in Manufacturing



owers the next
on of apps

re delivering GPT-3-powered
xt completion, and other advanced
API.



And then this happened ...

OpenAI's chatGPT

OpenAI's GPT-4

Google's Bard

Anthropic's Claude 2

...many more

Int

We've t
conver
for Cha
mistak
inappr

Try ChatG

The AI Arms Race



Siddharth Sharma · Follow

9 min read · Feb 9



2



"It's a new day in search. The race starts today... We're going to move fast."

— Satya Nadella, Microsoft CEO



Altman and Lecun (Forbes)

In recent years, the field of Artificial Intelligence has seen a rapid rise in the development of large language models. These models, based on deep

And things changed ...

SO

Please tell me what DH Laurence wanted to say with the following "I never saw a wild thing sorry for itself. A small bird will drop frozen dead from a bough without ever having felt sorry for itself." . Keep the analysis 10 lines long .



D.H. Lawrence's quote suggests that nature operates without self-pity. He highlights the resilience of the natural world, contrasting it with human tendencies to wallow in self-pity. Lawrence's message underscores the idea that self-pity is unique to humans, prompting reflection on our capacity to magnify suffering through it. He suggests that adopting a more resilient attitude akin to nature's might lead to a healthier perspective on life's challenges.

SO

now translate it in German



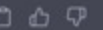
Das Zitat von D.H. Lawrence deutet darauf hin, dass die Natur ohne Selbstmitleid funktioniert. Er betont die Widerstandsfähigkeit der natürlichen Welt und stellt sie den menschlichen Neigungen gegenüber, im Selbstmitleid zu versinken. Lawrences Botschaft unterstreicht die Idee, dass Selbstmitleid nur dem Menschen eigen ist, und regt zur Reflexion über unsere Fähigkeit an, Leiden durch Selbstmitleid zu vergrößern. Er schlägt vor, eine widerstandsfähigere Einstellung, die der der Natur ähnelt, anzunehmen, um eine gesündere Perspektive auf die Herausforderungen des Lebens zu entwickeln.

SO

Please write for me an analysis about the US elections in 2020. IT should be 10 lines long



The 2020 US elections were historic and highly contentious. Incumbent President Donald Trump, representing the Republican Party, faced former Vice President Joe Biden, the Democratic candidate. Amid a global pandemic, mail-in voting played a significant role, leading to prolonged ballot counting and legal challenges in some states. Ultimately, Joe Biden secured victory with 306 electoral votes to Trump's 232. The election saw record voter turnout, reflecting intense polarization and voter engagement. Trump's refusal to concede and baseless claims of widespread voter fraud led to a tumultuous transition period. Biden's victory marked a significant shift in US politics, emphasizing the importance of unity and addressing the nation's pressing challenges.



SO

What is the best Harry Potter movie ?



The question of which Harry Potter movie is the best is subjective and can vary from person to person depending on their personal preferences. The Harry Potter film series consists of eight movies, each adapting one of the seven books in J.K. Rowling's Harry Potter series.



Some fans might prefer the earlier films, such as "Harry Potter and the Sorcerer's Stone" or "Harry Potter and the Chamber of Secrets," for their faithful adaptation of the source material and the innocence of the younger characters. Others might favor the later films, like "Harry Potter and the Deathly Hallows - Part 2," for their darker and more action-packed tone as the series reaches its climax.

SPEED OF ADOPTION

- GenAI adoption is being driven by democratization
- Public are getting exposure and awareness to GenAI faster than enterprises can keep up
- Need to move fast is driven by a real sense of market urgency. Don't get left behind!



MEETS REALITY OF ENTERPRISE

- GenAI famous for leaks and misuse
- Internal security and data governance not keeping pace with demand
- Confusion around IT requirements to create or use GenAI



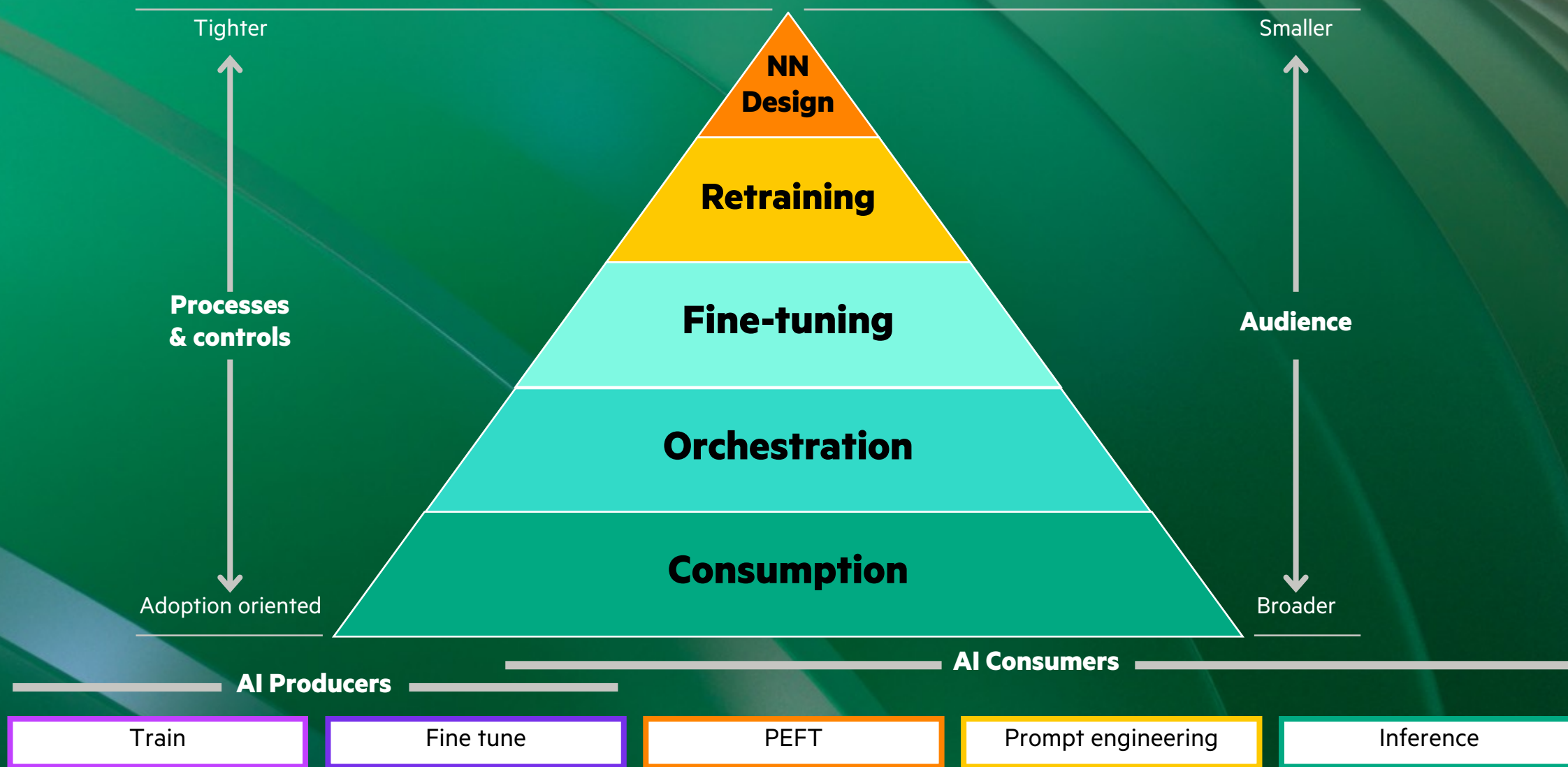
WHAT COULD GO WRONG?



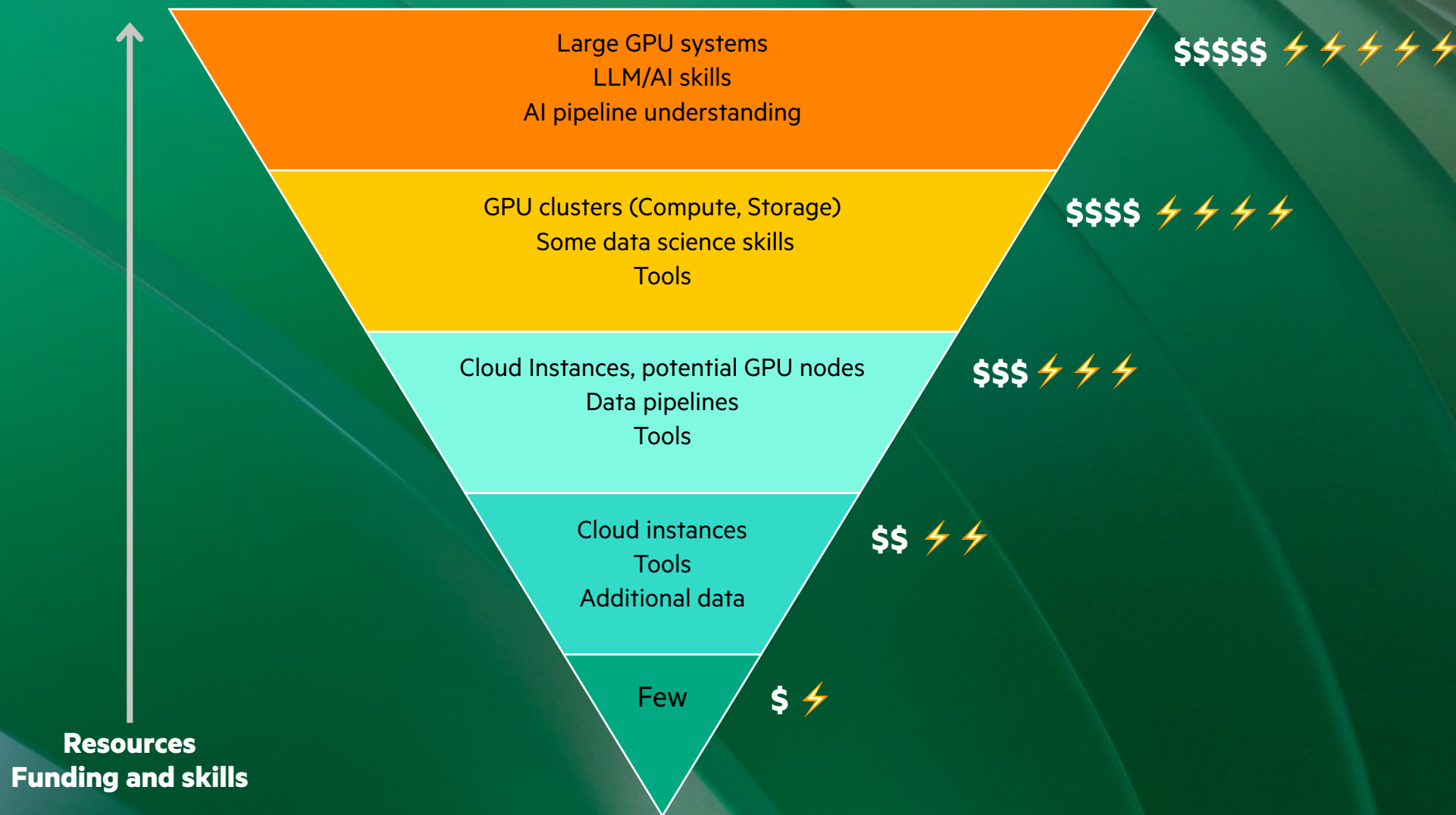
- Velocity
- Security
- Privacy
- Ethics
- Hallucination

HOW DO ENTERPRISES USE GENAI?

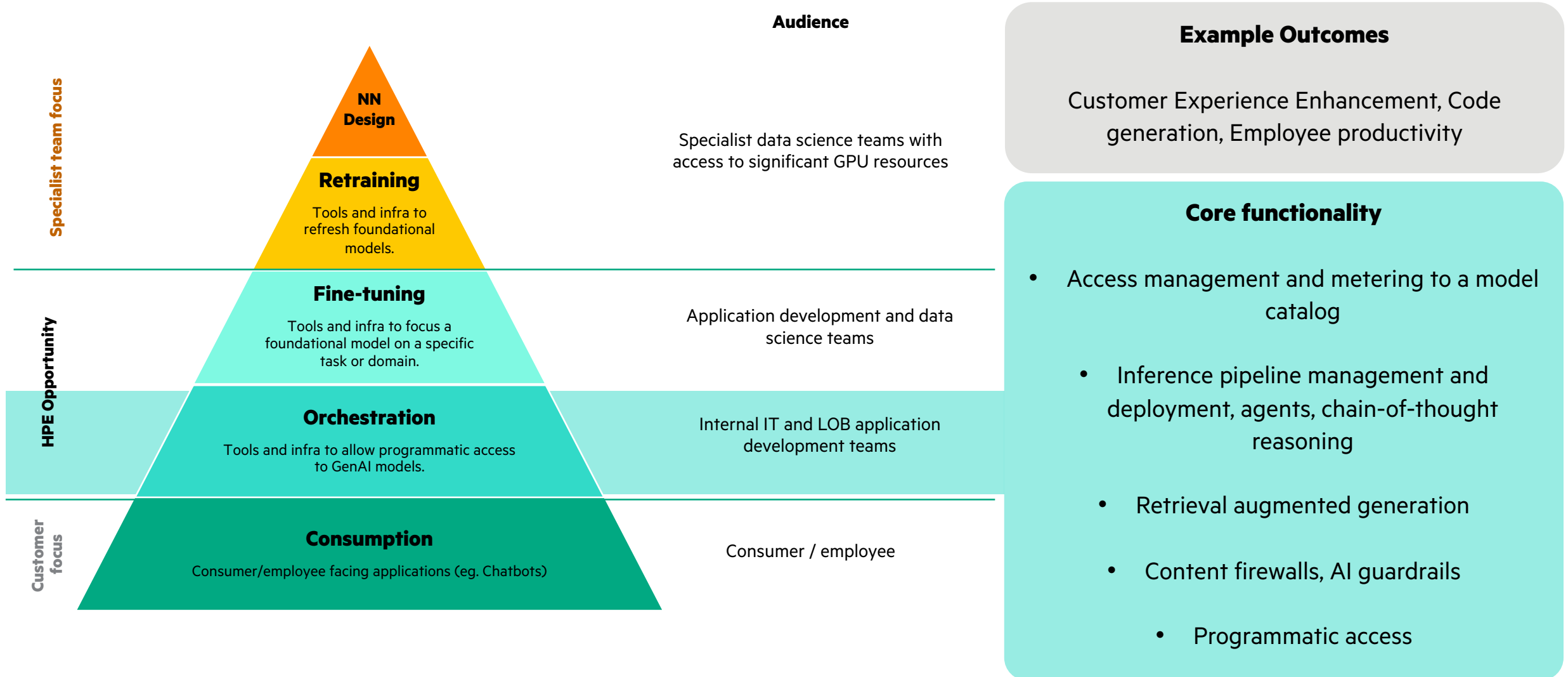
AI WORKLOAD TYPES



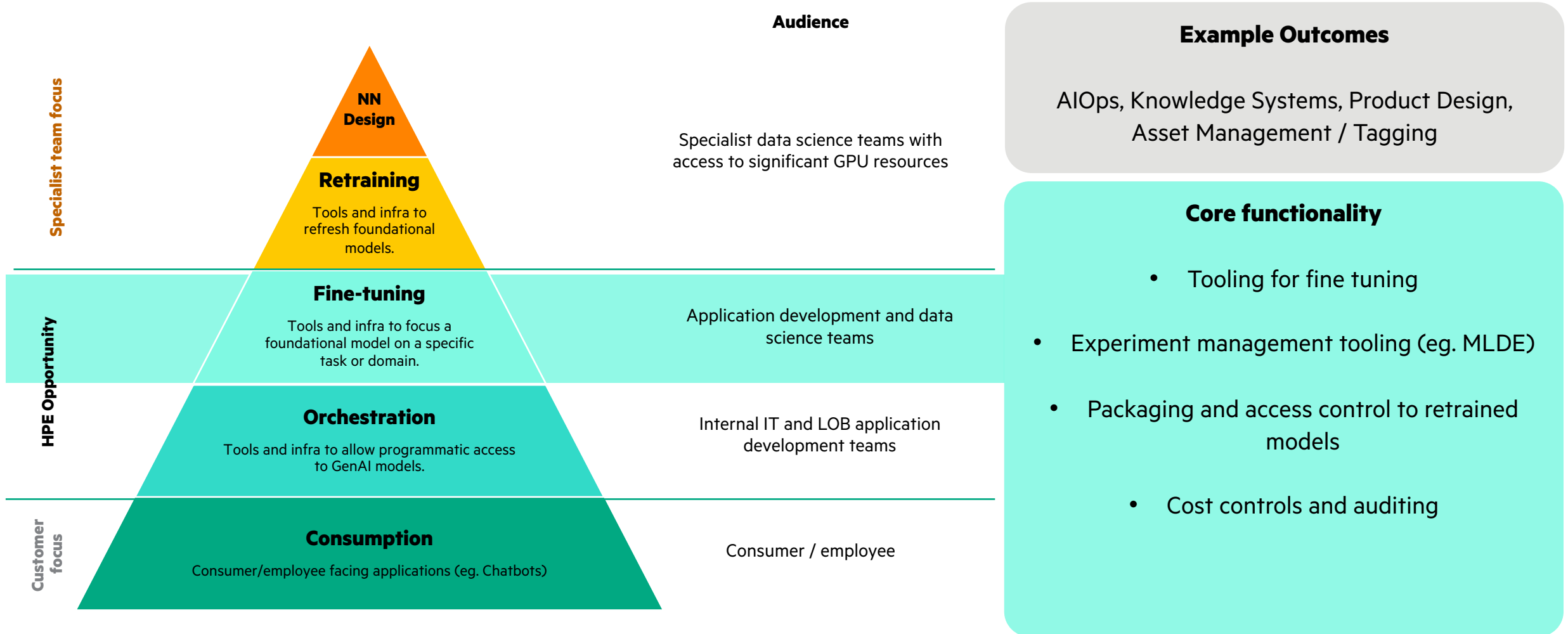
WHAT IS REQUIRED? | DEPENDS WHERE YOU WANT TO PLAY



GENAI FOR CREATIVE OUTCOMES



GENAI FOR MORE CONTEXT SPECIFIC OUTCOMES



WHAT ARE SOME OF THE QUESTIONS BEFORE ADOPTING?



1. Which One (company, model, open source)?
2. Is it production ready?
3. What about governance?
4. Which usecases are ready?
5. What do I need to get started?



SECURITY, PRIVACY, GOVERNANCE

RISKS/VULNERABILITIES FOR LLMS: OWASP

OWASP Top 10 for LLM Applications

LLM01

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins can add vulnerabilities.

LLM06

Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

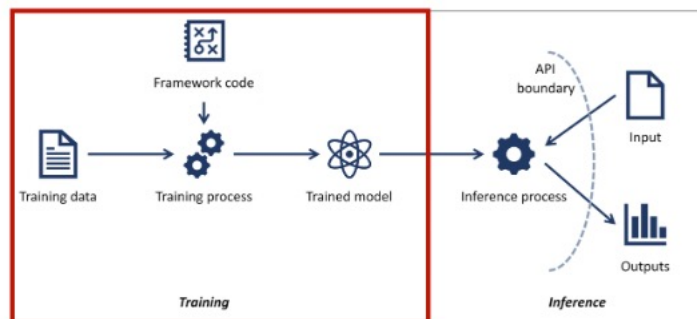
Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

AI ATTACK SCENARIOS

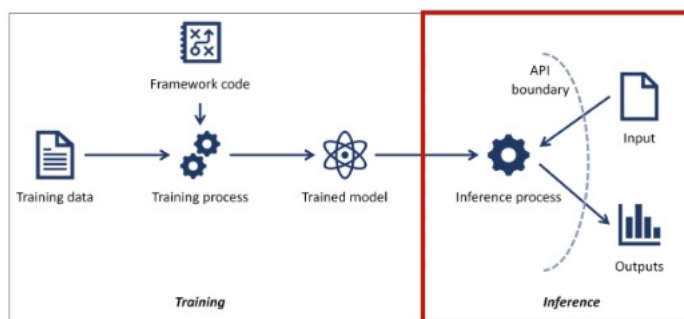
Training Time Attack

An attacker can compromise the training data via the feedback mechanism (e.g. Microsoft Tay chatbot).



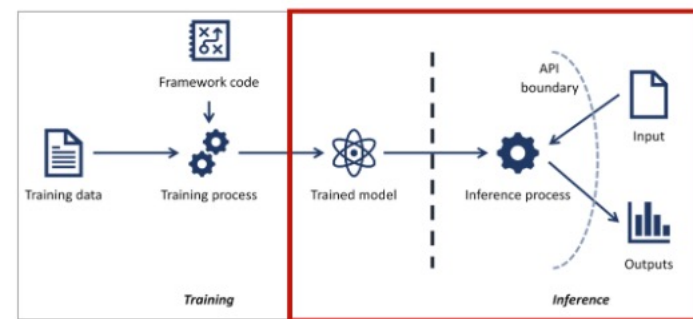
Inference Time Attack

Consider the most common deployment scenario where a model is deployed as an API endpoint. In this blackbox setting an attacker can only query the model and observe the response. The attacker controls the input to the model, but the attacker does not know how it is processed.

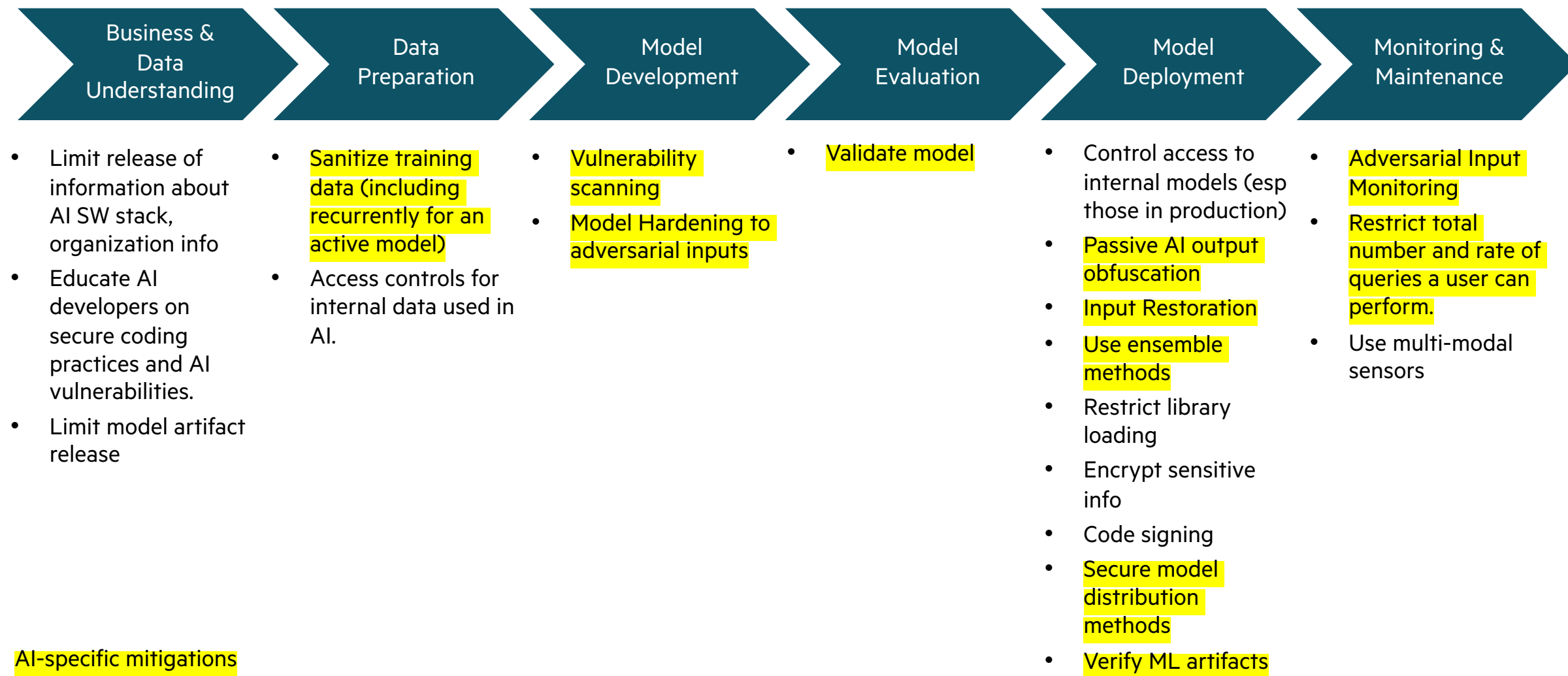


Attack on Edge/Client

A model exists on a client (e.g., phone) or on the edge (e.g., IoT). An attacker might have access to model code through reversing the service on the client.



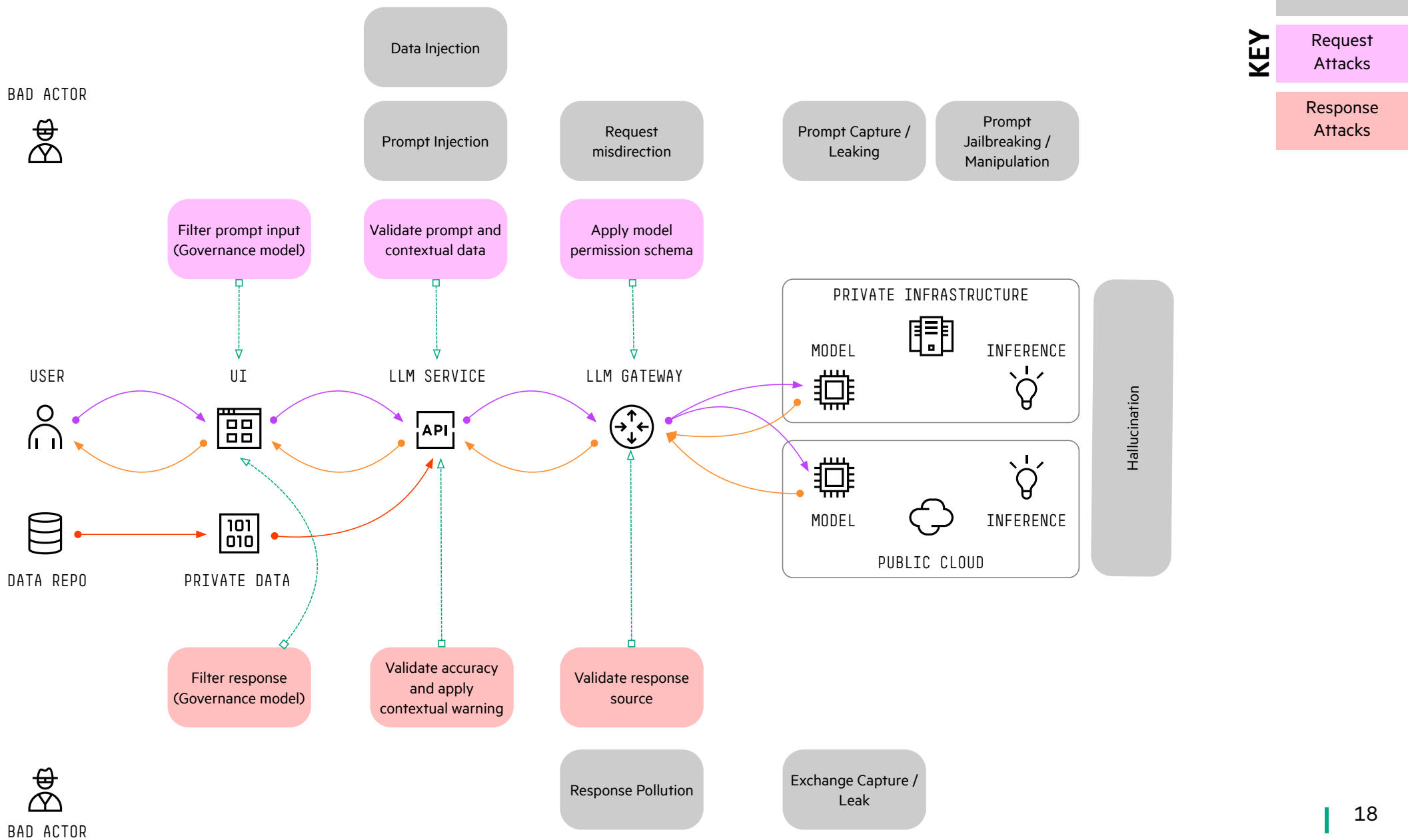
SECURITY MITIGATIONS OVER AI LIFECYCLE

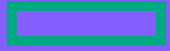


AI-specific mitigations

Adapted from: <https://atlas.mitre.org/mitigations/>

SECURITY THREAT LANDSCAPE FOR GENAI





Hewlett Packard
Enterprise

ENTER “GUARDRAILS”



VERY DYNAMIC SPACE

- Open Source ++
 - Startups ++++
 - Funding ++++
 - Success ?
-
- API's changing
 - Models needed (who checks these?)
 - Traditional “Heuristics” approach?



CONFIDENTIAL | AUTHORIZED

NEMO GUARDRAILS

- Open-source toolkit for easily adding programmable guardrails to LLM-based conversational applications.
- Guardrails (or "rails" for short) are specific ways of controlling the output of a large language model, such as ruling out topics like politics.
- Responding in a particular way to specific user requests
- Following a predefined dialog path
- Using a particular language style, extracting structured data, and more.



USE CASES

Question Answering

RAG or Retrieval Chains

QnA over a set of documents, enforcing fact checking and output moderation

Domain Specific Assistants

Chatbots

Ensures the assistant stays on topic and follows the designed conversational flows

LLM Endpoints

Custom LLMs

Add guardrails to custom LLMs for safer customer interactions

Langchain

Chains, Agents ...

Add guardrails layer around your chains and LLM-based agents.



GUARDRAILS TYPES

Input

Input Rails

applied to user input.
can reject or stop additional processing steps, or alter the input (mask sensitive data, ...etc.)

Execution

Execution Rails

applied to input/output of the custom actions (a.k.a. tools), that need to be called by the LLM.

Dialog

Dialog Rails

operate on canonical forms and determine the next steps; execute an action, generate LLM response, return predefined response, ...etc.)

Output

Output Rails

applied to the LLM generated output; it can reject the output, not returning it to the user, or alter it (e.g., removing sensitive data).

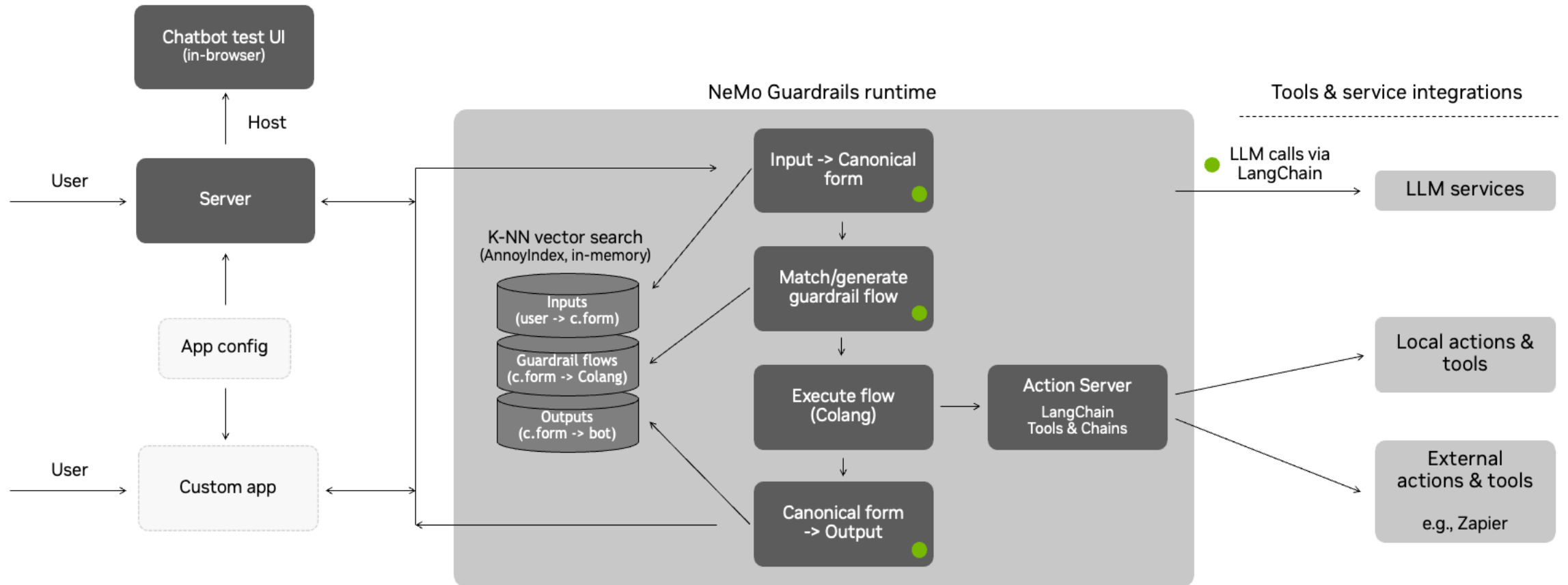
Retrieval

Retrieval Rails

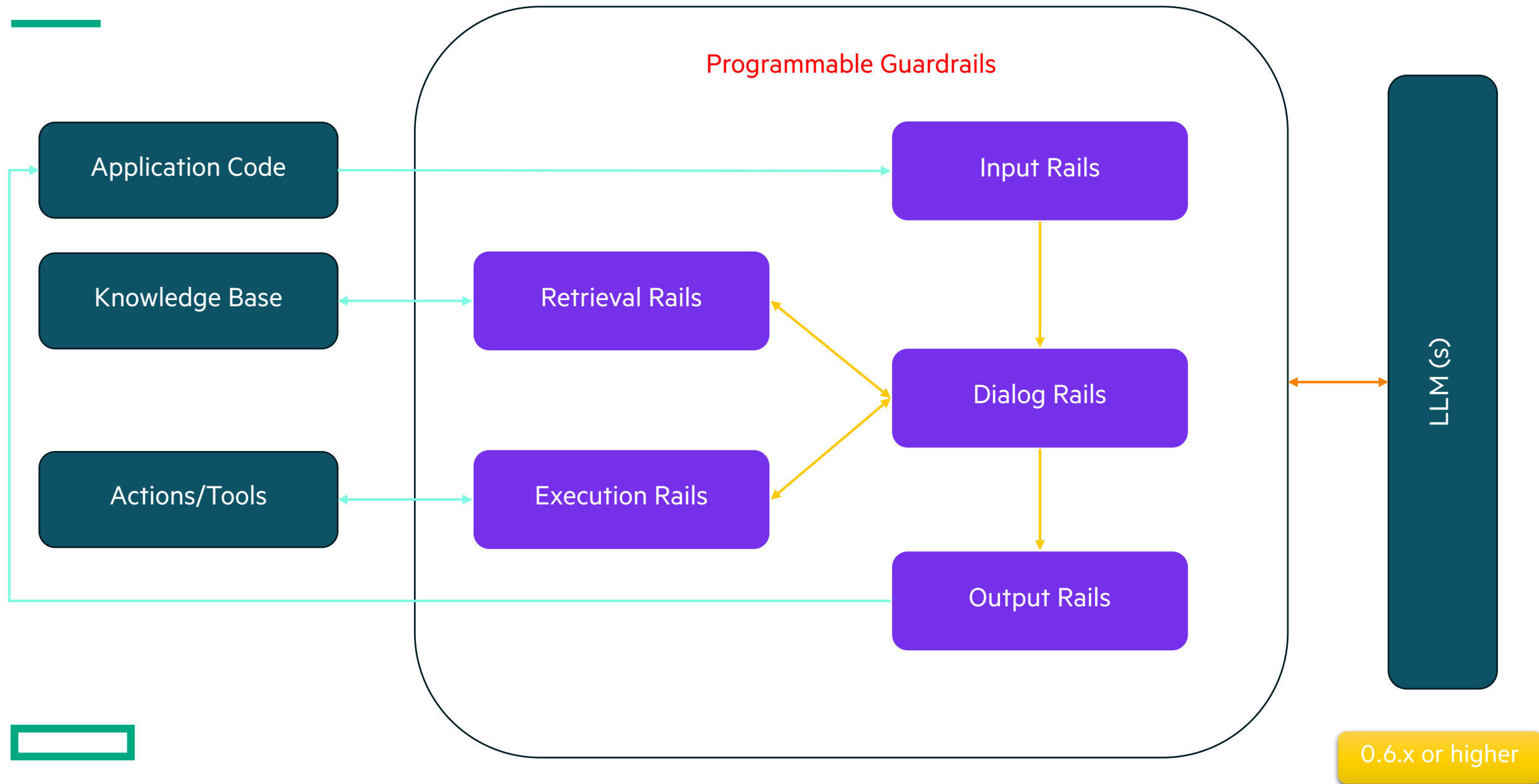
Applied to retrieved chunks in RAG use-case, it can reject a chunk, preventing it from being used to prompt the LLM, or alter relevant chunks to mask sensitive data.

0.6.x or higher

ARCHITECTURE (OLD)



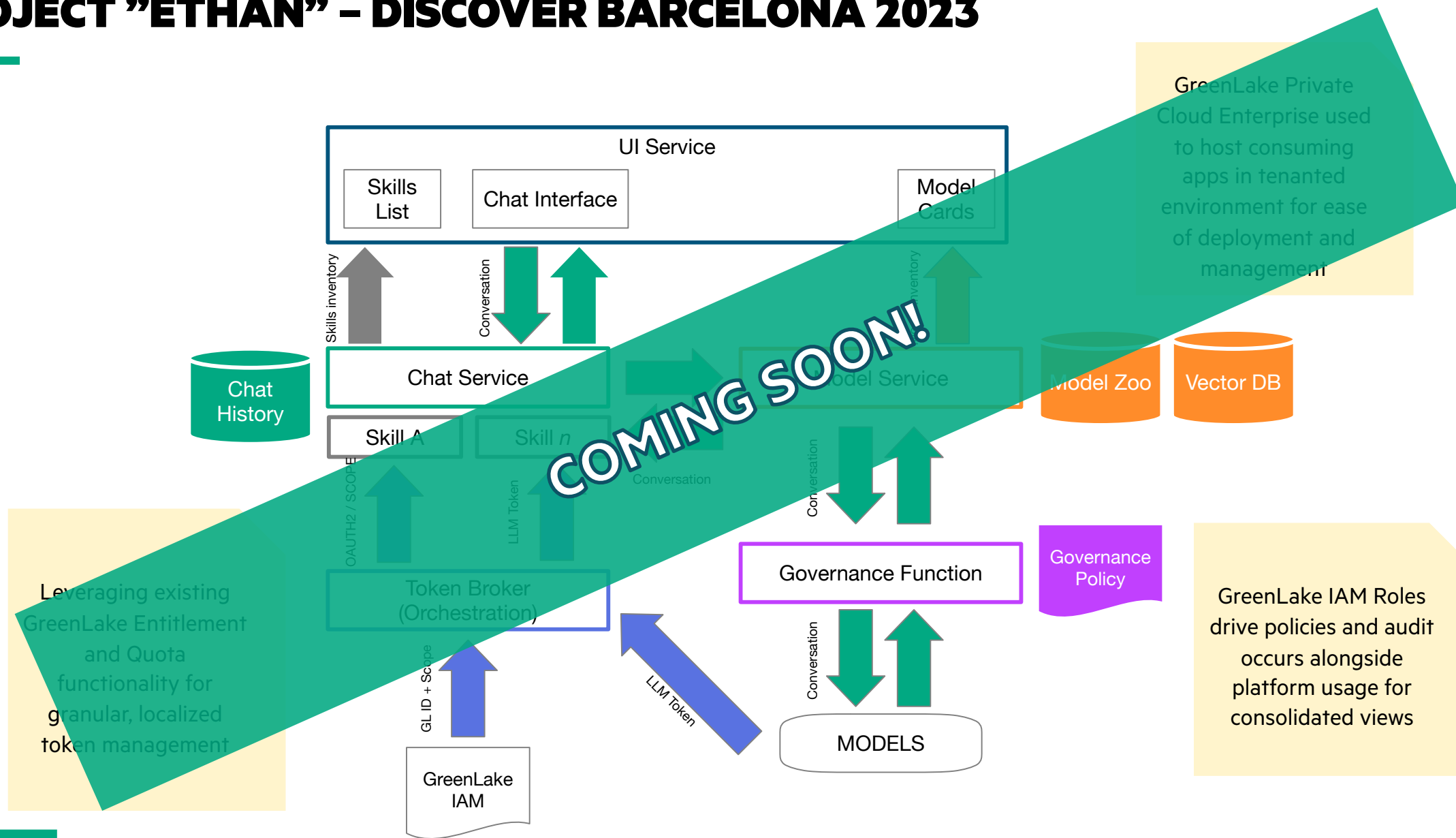
ARCHITECTURE (NEW!)



HPE

GREENLAKE

PROJECT "ETHAN" – DISCOVER BARCELONA 2023



Enabling Artificial Intelligence

HPE 
GreenLake

Infrastructure

From bare metal to containers

From training to inference

From the supercomputing to edge

+

AI Platform

Data Management

Model training

Model inference

+

Models and Services

Open-source models

Commercial partnerships

Experts that can get you started

One platform | Vendor neutral | Cloud Neutral | AI accessible for all

THANK YOU

tom.phelan@hpe.com
glyn.bowden@hpe.com
saad.zaher@hpe.com

