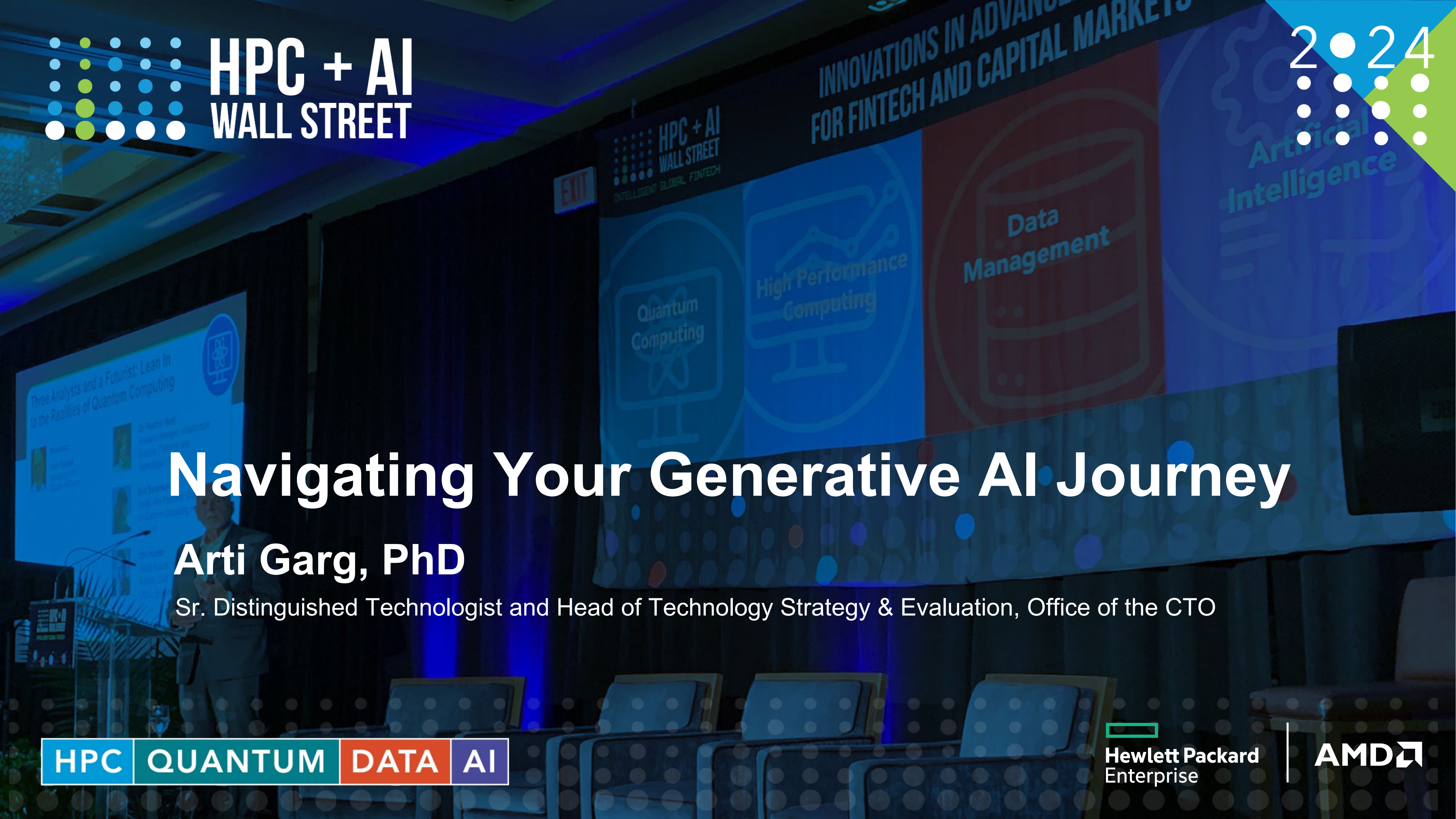
HPC + AI WALL STREET

2024



Navigating Your Generative AI Journey

Arti Garg, PhD

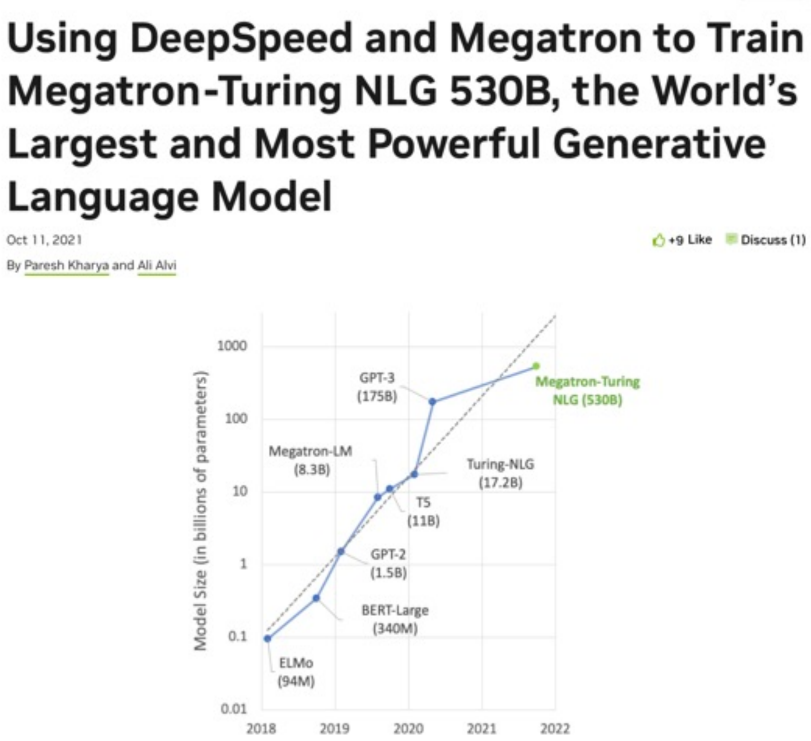
Sr. Distinguished Technologist and Head of Technology Strategy & Evaluation, Office of the CTO

HPC | QUANTUM | DATA | AI

Hewlett Packard Enterprise | AMD

At HPE, we believe in changing the way people live and work.

The World Before November 2022…



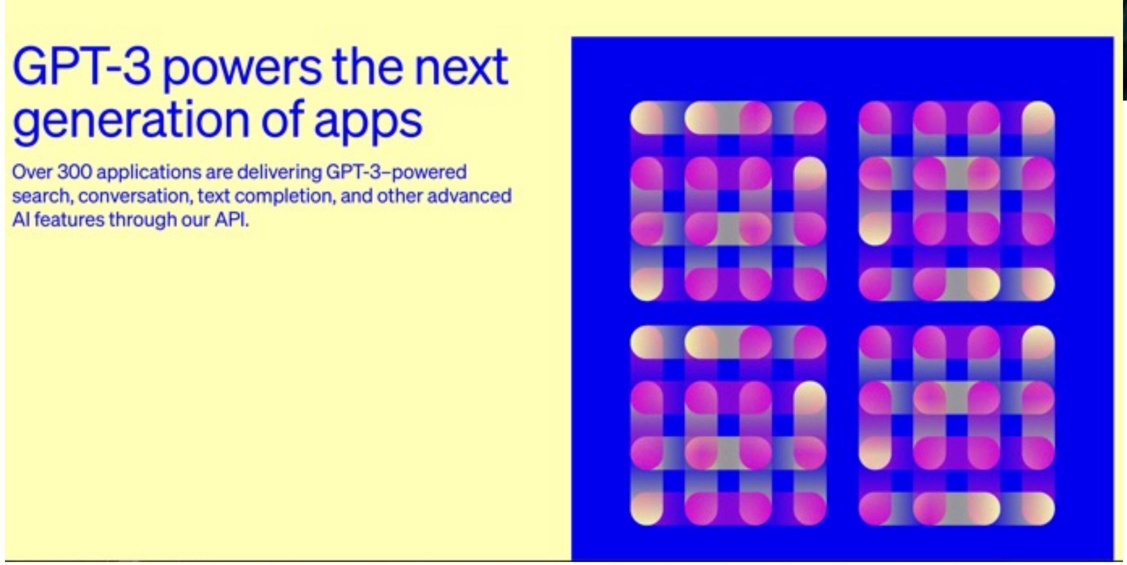
Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model

Oct 11, 2021

+9 Like Discuss (1)

By Paresh Kharya and Ali Alvi

Top 9 Use Cases of Computer Vision in Manufacturing



GPT-3 powers the next generation of apps

Over 300 applications are delivering GPT-3–powered search, conversation, text completion, and other advanced AI features through our API.

…and Then This Happened…

OpenAI's chatGPT

OpenAI's GPT-4

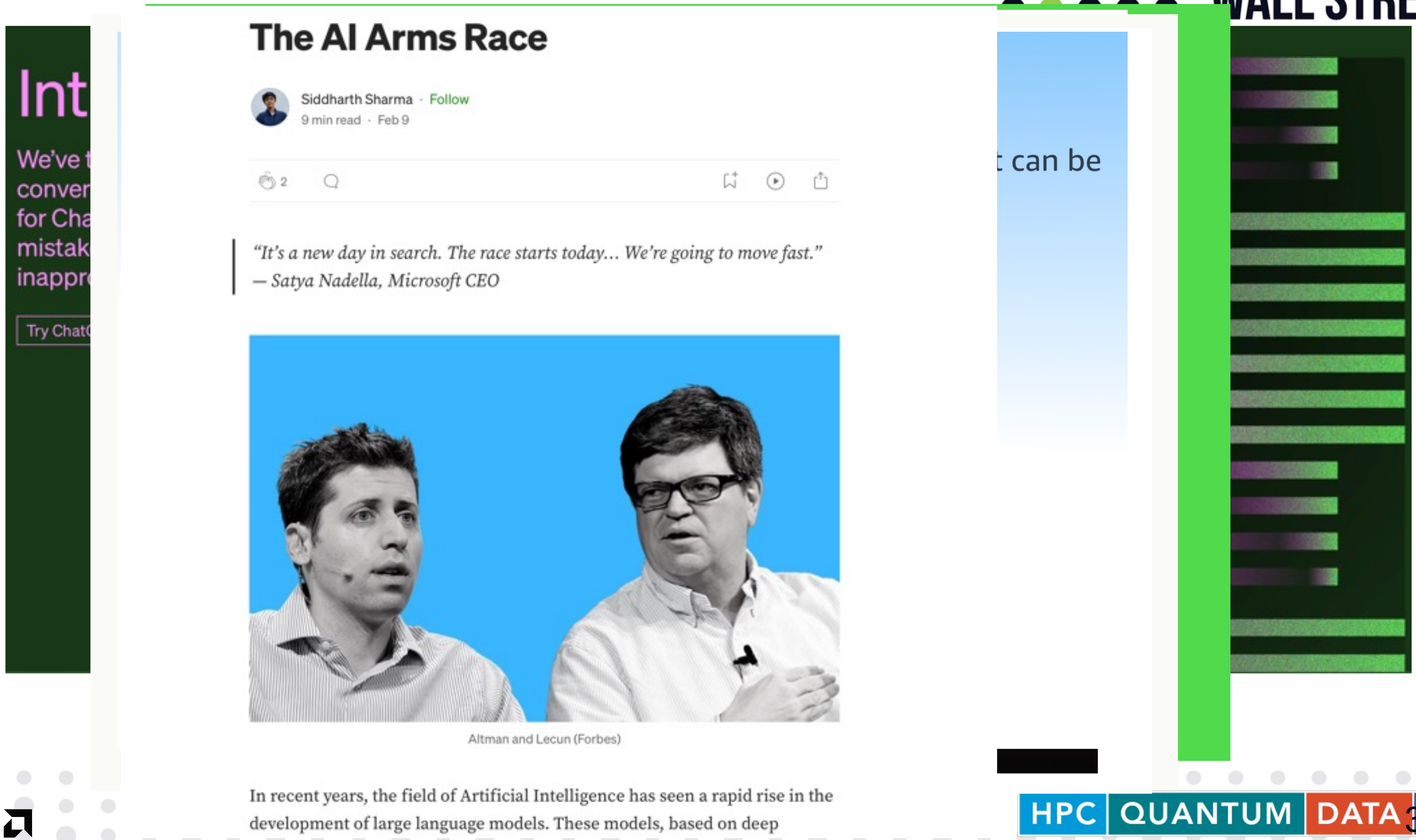
Google's Bard

Anthropic's Claude 2

Google's Gemini

AWS's Q

…many more



The AI Arms Race

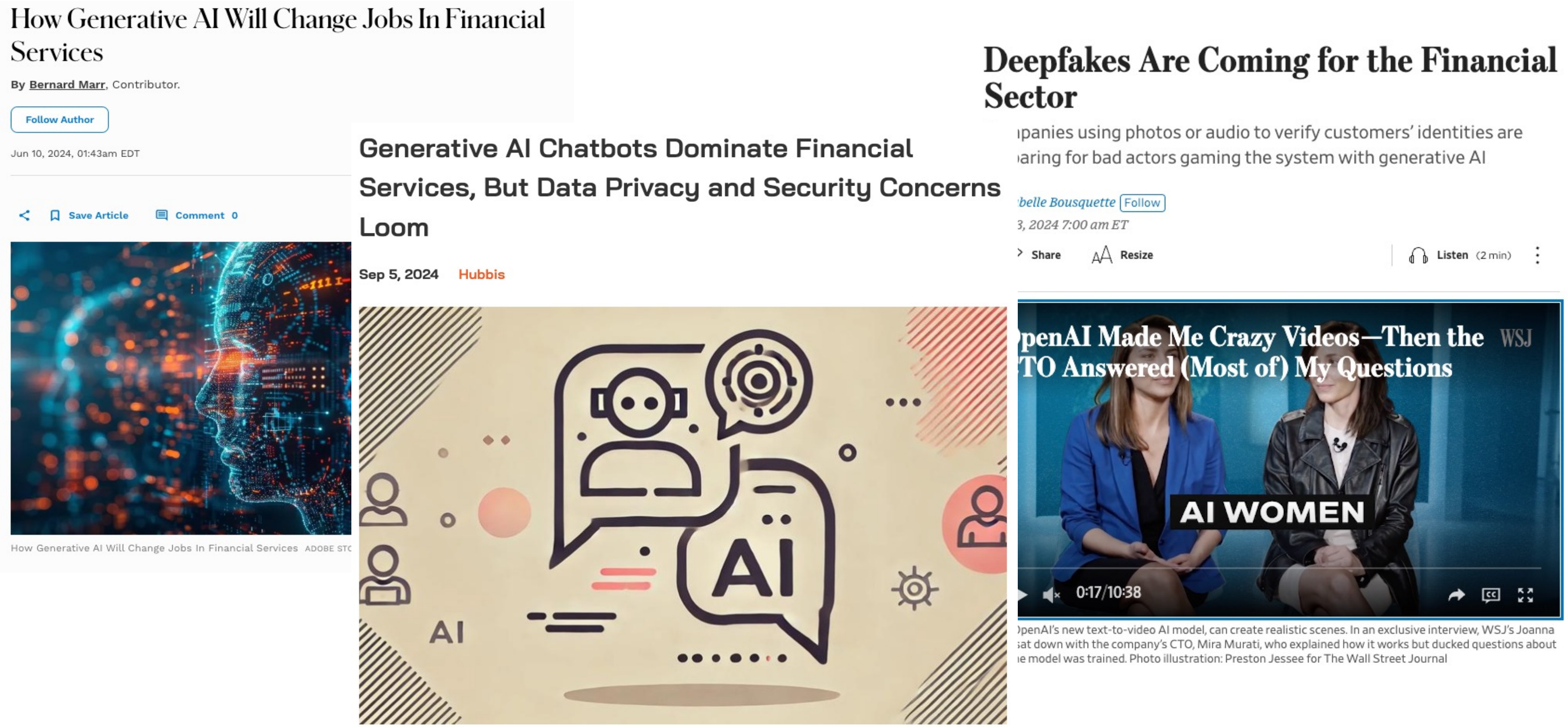
Siddharth Sharma · Follow
9 min read · Feb 9

> *"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 Everything Changed



How Generative AI Will Change Jobs In Financial Services

By Bernard Marr, Contributor.

Follow Author

Jun 10, 2024, 01:43am EDT

Save Article   Comment 0

How Generative AI Will Change Jobs In Financial Services ADOBE STC

Generative AI Chatbots Dominate Financial Services, But Data Privacy and Security Concerns Loom

Sep 5, 2024 Hubbis

Deepfakes Are Coming for the Financial Sector

ıpanies using photos or audio to verify customers' identities are ɪaring for bad actors gaming the system with generative AI

ɪbelle Bousquette Follow

3, 2024 7:00 am ET

Share   Resize   Listen (2 min)

ɪpenAI Made Me Crazy Videos—Then the
ʼTO Answered (Most of) My Questions
WSJ
AI WOMEN
0:17/10:38

ɪpenAI's new text-to-video AI model, can create realistic scenes. In an exclusive interview, WSJ's Joanna ɪat down with the company's CTO, Mira Murati, who explained how it works but ducked questions about ɪe model was trained. Photo illustration: Preston Jessee for The Wall Street Journal

Where Do You Go From Here?

1 What is needed to **get started ?**

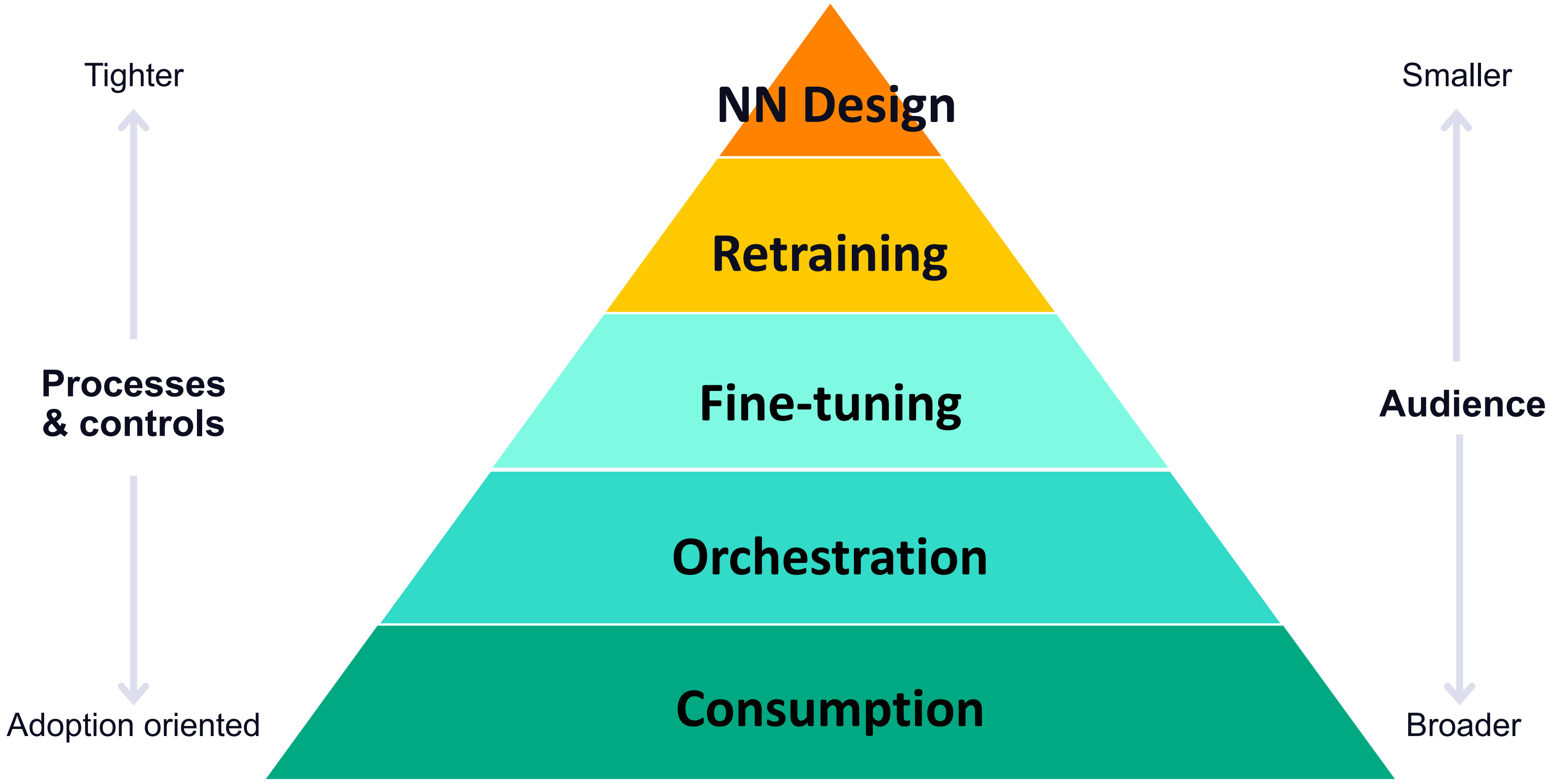
2 Which **one** (company, model, open source)?

3 Is it **enterprise ready**?

4 What about **governance**?

5 Which **use cases**?

What Is Required?


Tighter

Processes & controls

Adoption oriented

NN Design

Retraining

Fine-tuning

Orchestration

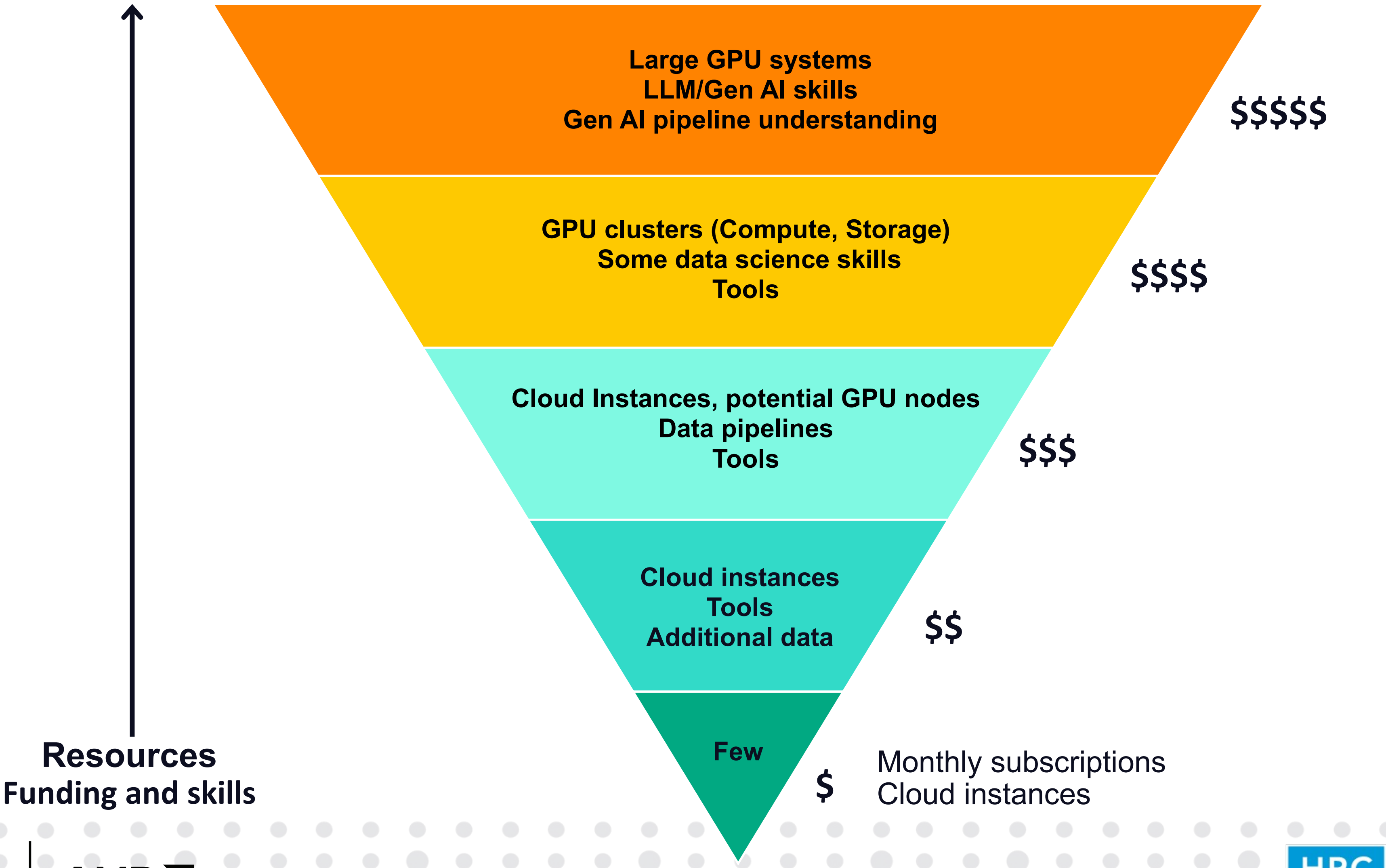
Consumption

Smaller

Audience

Broader

What Is Required? I Depends On Where You Want To Play



Large GPU systems
LLM/Gen AI skills
Gen AI pipeline understanding

$$$$$

GPU clusters (Compute, Storage)
Some data science skills
Tools

$$$$

Cloud Instances, potential GPU nodes
Data pipelines
Tools

$$$

Cloud instances
Tools
Additional data

$$

Few

$

Monthly subscriptions
Cloud instances

Resources
Funding and skills

Artificial Intelligence (Gen AI and Beyond)

Models and Services

Open-source models

Commercial partnerships

Services to better drive AI models

+

AI Platform

Data management

Model training

Model inference

+

Infrastructure

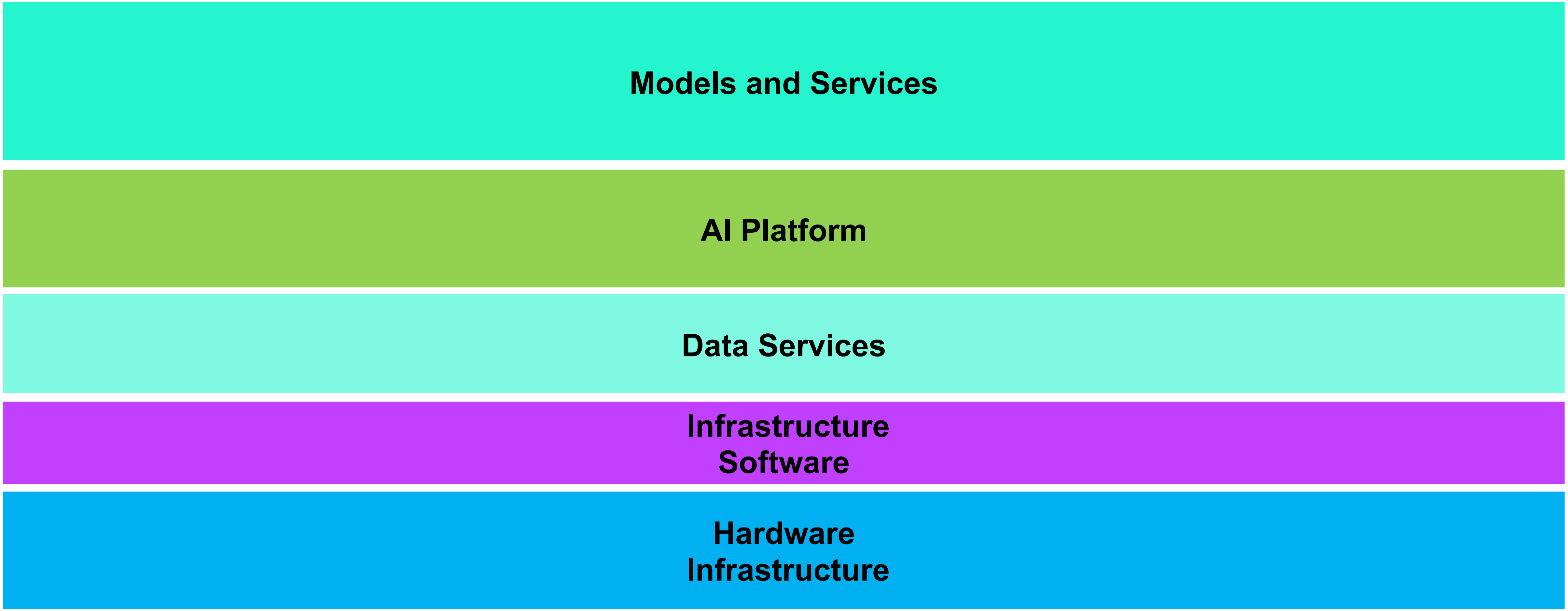
From bare metal to containers

From training to inference

From supercomputer to edge

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

What Are Your Key Considerations?



Models and Services

AI Platform

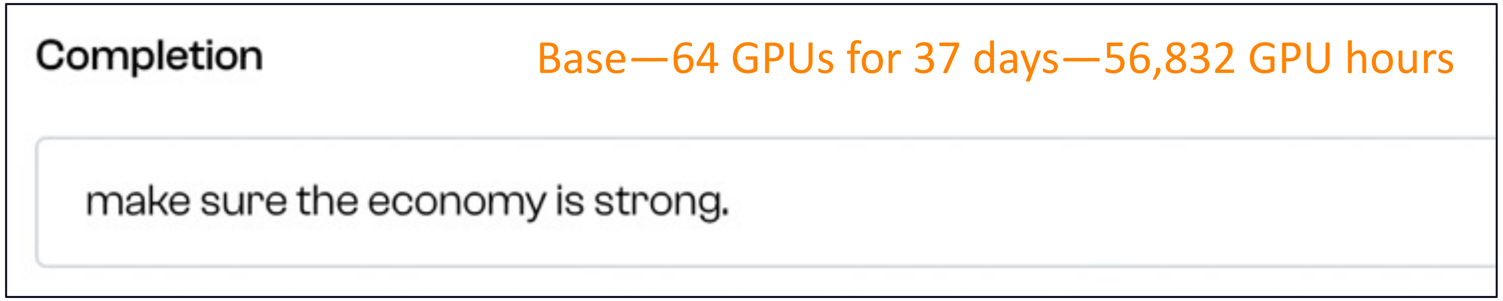
Data Services

Infrastructure Software

Hardware Infrastructure

Infrastructure | Good models require good data and also a lot of compute

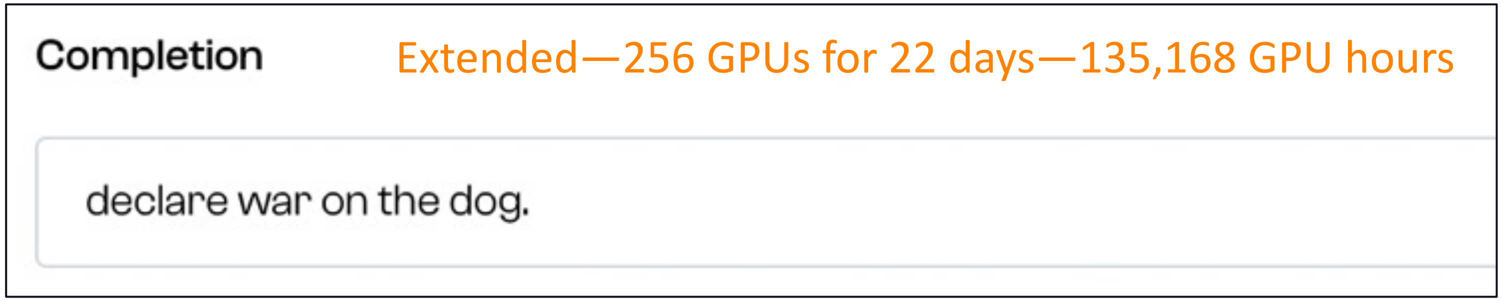
If this cat were elected president, its first order of business would be to…



Completion

Base—64 GPUs for 37 days—56,832 GPU hours

make sure the economy is strong.



Completion

Extended—256 GPUs for 22 days—135,168 GPU hours

declare war on the dog.

Infrastructure I Where to Deploy?

	Pros	Cons
Private Cloud	• Enhanced Security & Privacy • Customization • Predictable performance • Compliance adherence • Reduced Risk of vector lock in	• Higher Initial Investment • Limited Scalability • Resource underutilization • Complexity
Public Cloud	• Cost Efficiency • Scalability • Global Accessibility • Maintenance and Updates • Innovation	• Security and Privacy Concerns • Dependency • Limited Control • Potential Compliance Issues • Vendor Lock-In

Platform | A Very Complicated Landscape

Data

Pipelines

Pachyderm

Versioning

LFS

DVC

DELTA LAKE

Pachyderm

Processing

RAY

dask

RAPIDS

Pachyderm

Labelling

scale

Labelbox

Data Sources

LOG

CSV

aws

Azure

SQL

snowflake

CLOUDERA

databricks

lustre

MAPR

S3

Development/Training

Foundation Models

GitHub

arXiv

Packages

HOROVOD

IDE *

jupyter

PC

Evaluation

TensorBoard

gradio

Experiments

Determined AI

mlflow

Scheduling

kubernetes

slurm
workload manager

Compute

intel

NVIDIA

AMD

habana
An Intel Company

cerebras

Qualcomm

Deployment/Inference

Logs, Metrics

Prometheus

Grafana

Optimization

tvm

TensorRT

Testing

great_expectations

Buildkite

Trustworthy AI

ALIBI EXPLAIN

truera

modzy

fiddler

ALIBI DETECT

TROJ.AI

Serving (Central)

KServe

BENTOML

SELDON

Serving (Edge)

Triton Inference Server

Custom Docker

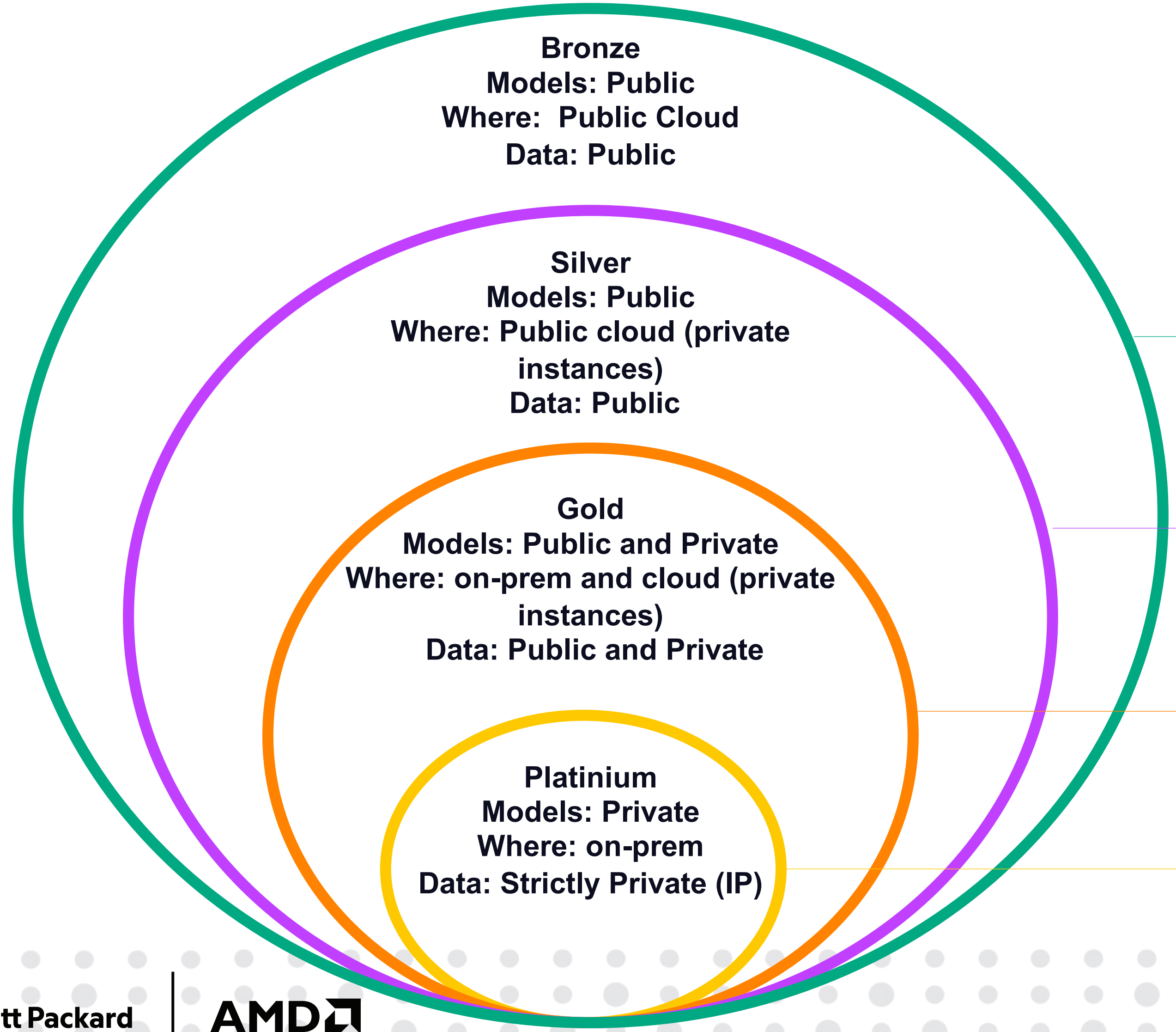
Runtime Environment

TensorFlow Extended

ONNX

TORCHSERVE

Models | How Do You Choose the Right One? It Depends



Bronze
Models: Public
Where: Public Cloud
Data: Public

Silver
Models: Public
Where: Public cloud (private instances)
Data: Public

Gold
Models: Public and Private
Where: on-prem and cloud (private instances)
Data: Public and Private

Platinium
Models: Private
Where: on-prem
Data: Strictly Private (IP)

Enterprises need hybrid environments to support multiple safety, control, and privacy requirements.

ChatGPT like models used, either direct access or via APIs.

Models fine tuned in the public cloud, public models API end points used in private instances, own models built

Models trained on prem.

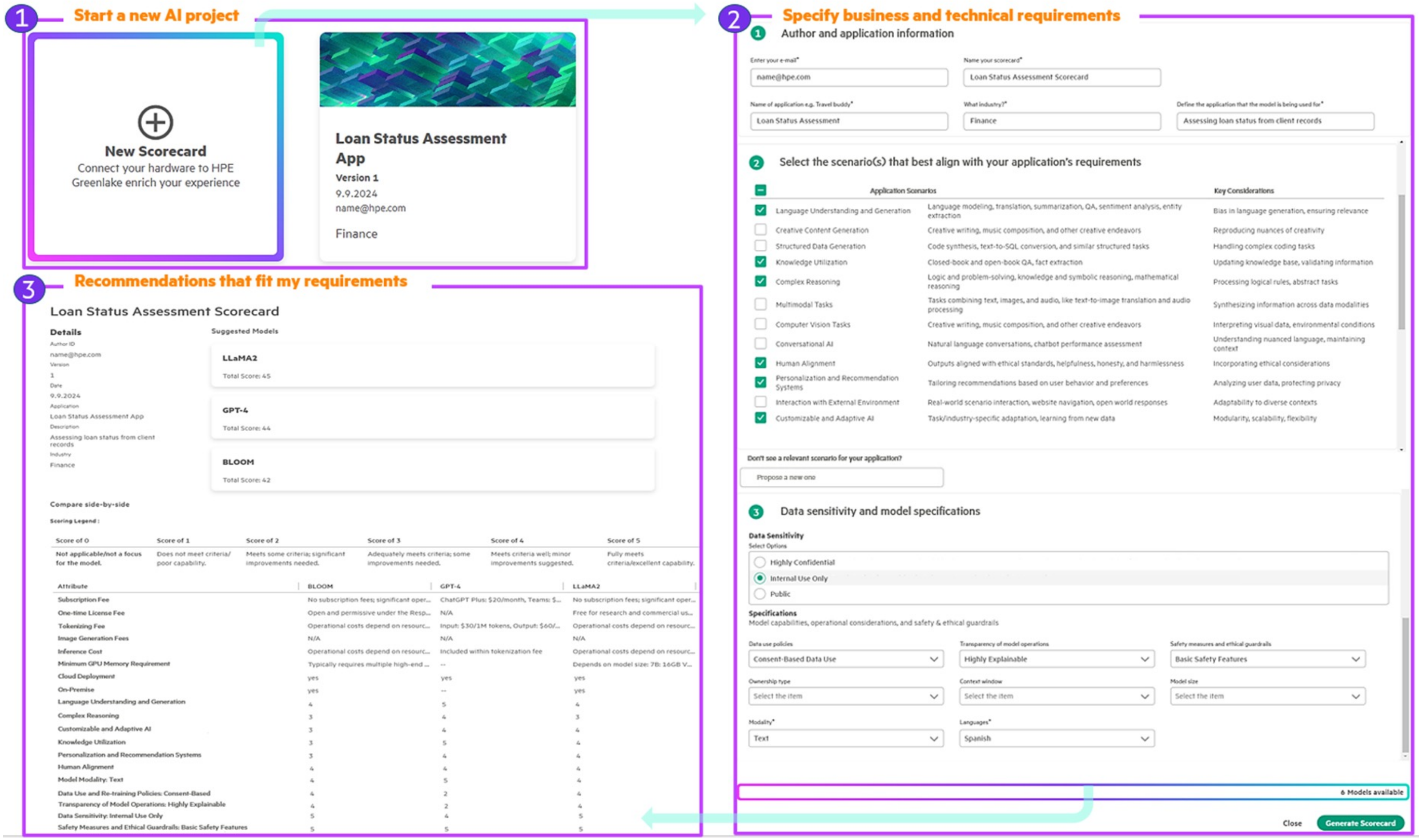
Models fine-tuned in the public cloud or on prem.

Inference running on prem and public cloud

Models trainined and running on prem

Models | Choosing the Right Foundation Model

- Select the best AI foundation model for your use case
- Evaluation categories:
  - Project Requirements
  - Model Capabilities
  - Operational & Ethical Considerations
  - Ethical & Governance Considerations
- Align model choice with project goals and organizational policy



1 Start a new AI project

New Scorecard
Connect your hardware to HPE Greenlake enrich your experience

Loan Status Assessment App
Version 1
9.9.2024
name@hpe.com
Finance

2 Specify business and technical requirements

1 Author and application information

Enter your e-mail*: name@hpe.com
Name your scorecard*: Loan Status Assessment Scorecard
Name of application e.g. Travel buddy*: Loan Status Assessment
What industry?*: Finance
Define the application that the model is being used for*: Assessing loan status from client records

2 Select the scenario(s) that best align with your application's requirements

▣	Application Scenarios		Key Considerations
☑	Language Understanding and Generation	Language modeling, translation, summarization, QA, sentiment analysis, entity extraction	Bias in language generation, ensuring relevance
□	Creative Content Generation	Creative writing, music composition, and other creative endeavors	Reproducing nuances of creativity
□	Structured Data Generation	Code synthesis, text-to-SQL conversion, and similar structured tasks	Handling complex coding tasks
☑	Knowledge Utilization	Closed-book and open-book QA, fact extraction	Updating knowledge base, validating information
☑	Complex Reasoning	Logic and problem-solving, knowledge and symbolic reasoning, mathematical reasoning	Processing logical rules, abstract tasks
□	Multimodal Tasks	Tasks combining text, images, and audio, like text-to-image translation and audio processing	Synthesizing information across data modalities
□	Computer Vision Tasks	Creative writing, music composition, and other creative endeavors	Interpreting visual data, environmental conditions
□	Conversational AI	Natural language conversations, chatbot performance assessment	Understanding nuanced language, maintaining context
☑	Human Alignment	Outputs aligned with ethical standards, helpfulness, honesty, and harmlessness	Incorporating ethical considerations
☑	Personalization and Recommendation Systems	Tailoring recommendations based on user behavior and preferences	Analyzing user data, protecting privacy
□	Interaction with External Environment	Real-world scenario interaction, website navigation, open world responses	Adaptability to diverse contexts
☑	Customizable and Adaptive AI	Task/industry-specific adaptation, learning from new data	Modularity, scalability, flexibility

Don't see a relevant scenario for your application?
Propose a new one

3 Data sensitivity and model specifications

Data Sensitivity
Select Options
○ Highly Confidential
● Internal Use Only
○ Public

Specifications
Model capabilities, operational considerations, and safety & ethical guardrails

Data use policies: Consent-Based Data Use
Transparency of model operations: Highly Explainable
Safety measures and ethical guardrails: Basic Safety Features
Ownership type: Select the item
Context window: Select the item
Model size: Select the item
Modality*: Text
Languages*: Spanish

6 Models available

3 Recommendations that fit my requirements

Loan Status Assessment Scorecard

Details
Author ID
name@hpe.com
Version
1
Date
9.9.2024
Application
Loan Status Assessment App
Description
Assessing loan status from client records
Industry
Finance

Suggested Models

LLaMA2
Total Score: 45

GPT-4
Total Score: 44

BLOOM
Total Score: 42

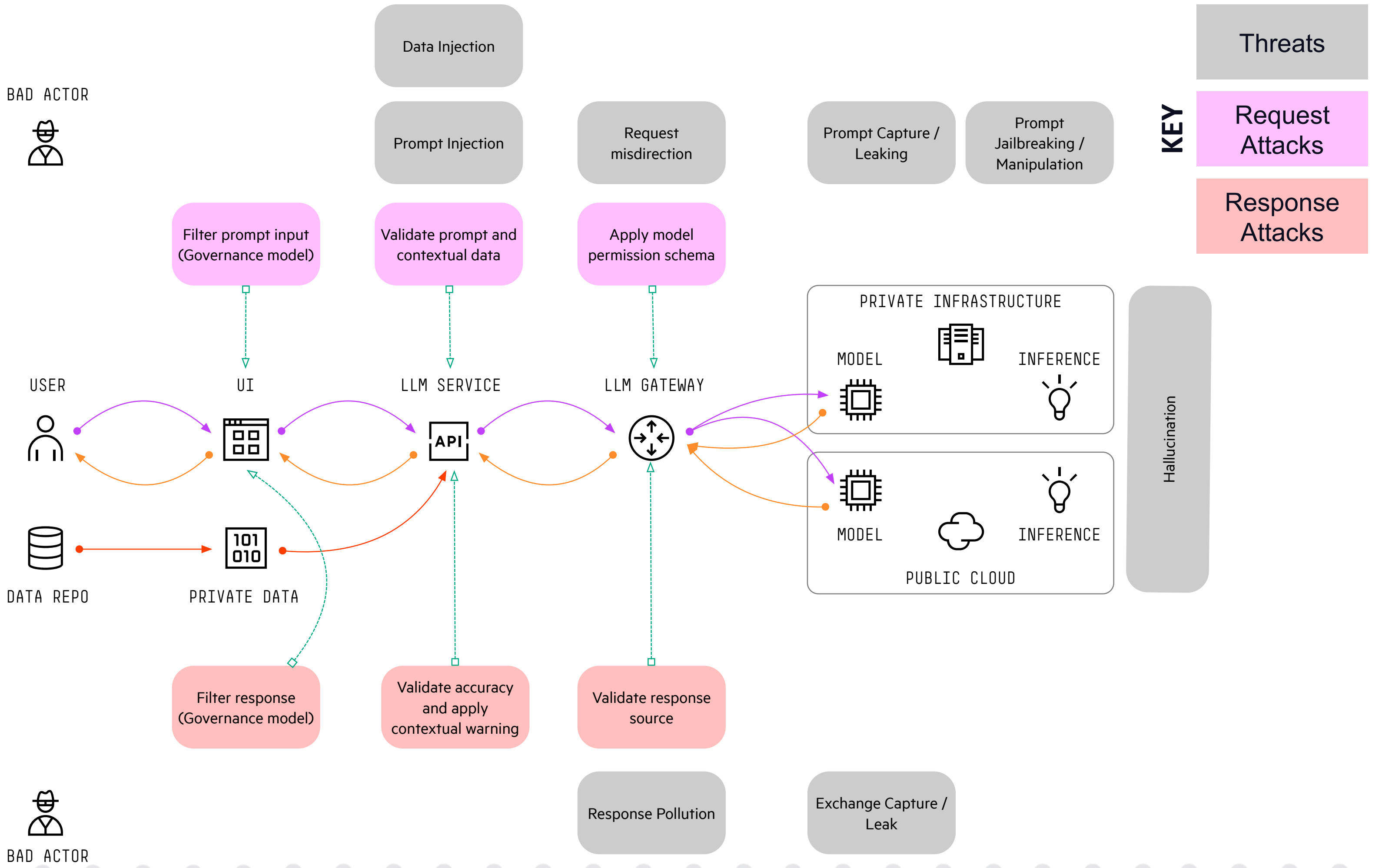
Compare side-by-side

Scoring Legend :

Score of 0	Score of 1	Score of 2	Score of 3	Score of 4	Score of 5
Not applicable/not a focus for the model.	Does not meet criteria/ poor capability.	Meets some criteria; significant improvements needed.	Adequately meets criteria; some improvements needed.	Meets criteria well; minor improvements suggested.	Fully meets criteria/excellent capability.

Attribute	BLOOM	GPT-4	LLaMA2
Subscription Fee	No subscription fees; significant oper...	ChatGPT Plus: $20/month, Teams: $...	No subscription fees; significant oper...
One-time License Fee	Open and permissive under the Resp...	N/A	Free for research and commercial us...
Tokenizing Fee	Operational costs depend on resourc...	Input: $30/1M tokens, Output: $60/...	Operational costs depend on resourc...
Image Generation Fees	N/A	N/A	N/A
Inference Cost	Operational costs depend on resourc...	Included within tokenization fee	Operational costs depend on resourc...
Minimum GPU Memory Requirement	Typically requires multiple high-end ...	--	Depends on model size: 7B: 16GB V...
Cloud Deployment	yes	yes	yes
On-Premise	yes	--	yes
Language Understanding and Generation	4	5	4
Complex Reasoning	3	4	3
Customizable and Adaptive AI	3	4	4
Knowledge Utilization	3	5	4
Personalization and Recommendation Systems	3	4	4
Human Alignment	4	4	4
Model Modality: Text	4	5	4
Data Use and Re-training Policies: Consent-Based	4	2	4
Transparency of Model Operations: Highly Explainable	4	2	4
Data Sensitivity: Internal Use Only	5	4	5
Safety Measures and Ethical Guardrails: Basic Safety Features	5	5	5

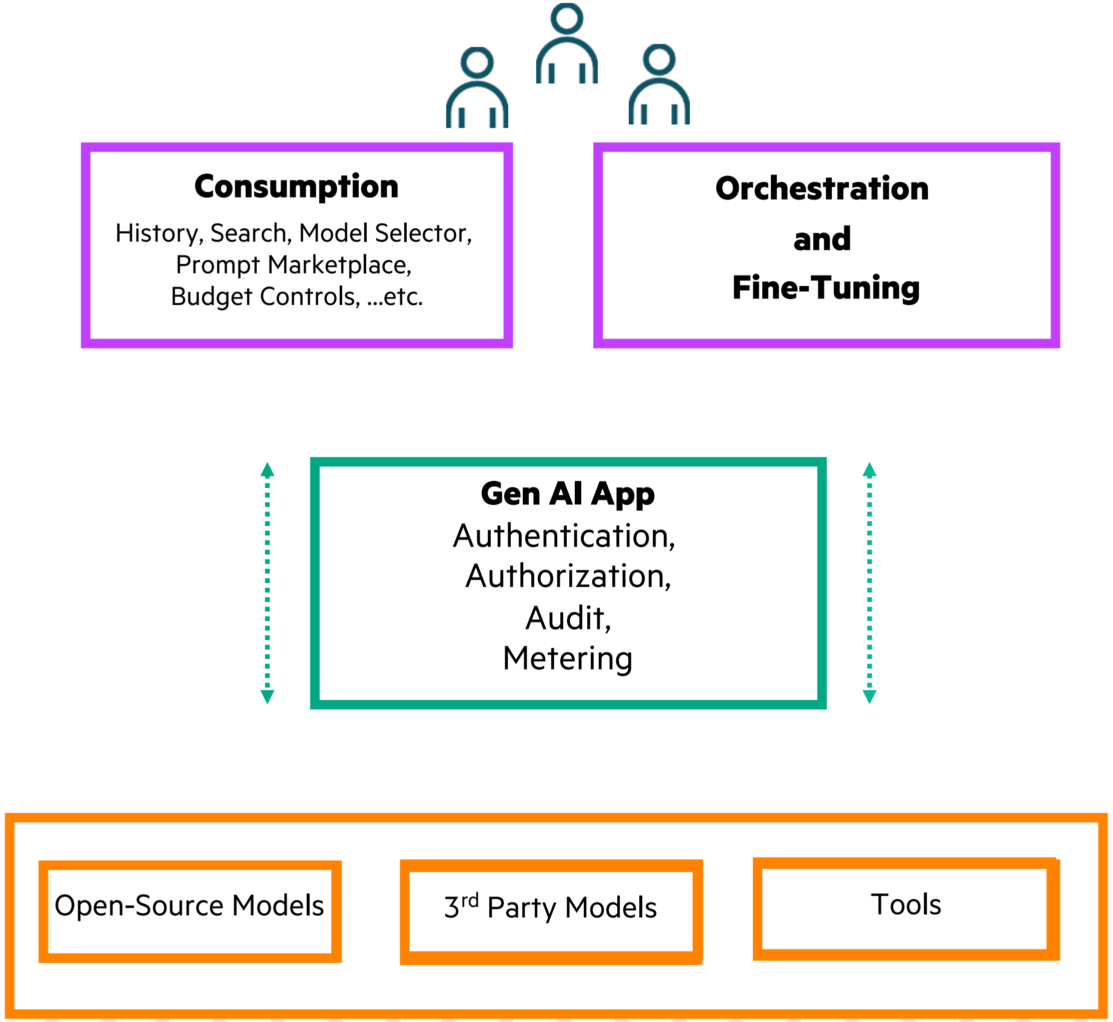
Implementation | What About Security?



- Fail gracefully and secretly
- Audit user authorization
- Secure access to external resources
- Parameterize and validate all inputs and outputs
- Avoid persisting changes when possible
- Adversarial testing

Implementation | Governance

- Allow employees to **securely and auditably** use new technology to assist with researching, writing, and code generation.
- Keep your **sensitive data on premise.**
- Deliver the ability to **choose your model.**
- Enforce **company policy**.
- *KEY DELIVERABLES*
  - *AI guardrails*
  - *Model Token Management*
  - *Access to models either on-prem or public cloud*



Consumption
History, Search, Model Selector, Prompt Marketplace, Budget Controls, ...etc.

Orchestration and Fine-Tuning

Gen AI App
Authentication, Authorization, Audit, Metering

Open-Source Models

3rd Party Models

Tools

What Else? | Observability | Sustainability | FinOps

HPC + AI WALL STREET

Clouds

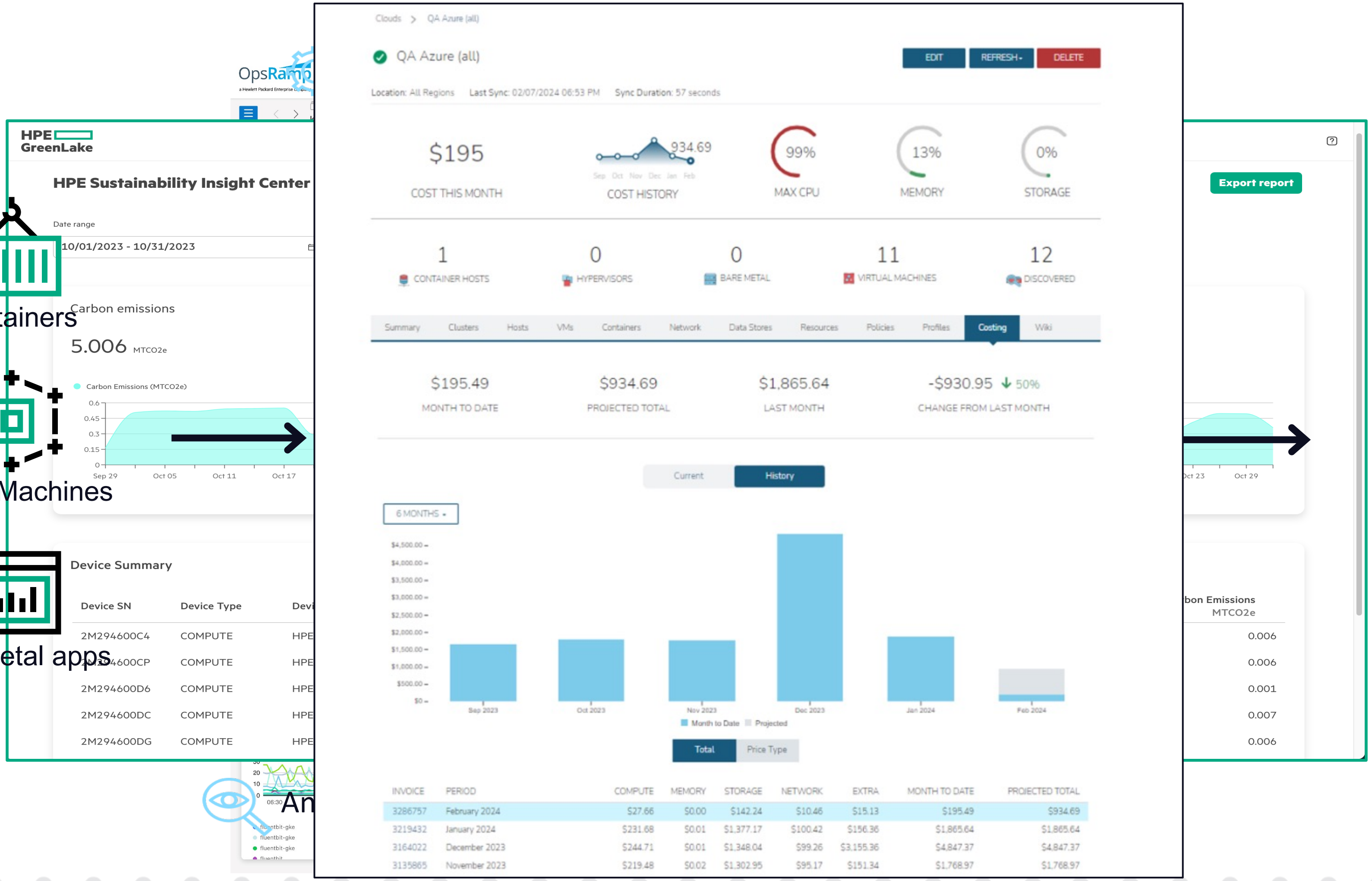
Datacenter

GPU

Containers

Virtual Machines

Bare metal apps



HPE Sustainability Insight Center

Date range: 10/01/2023 - 10/31/2023

Carbon emissions

5.006 MTCO2e

Device Summary

Device SN	Device Type	Devi
2M294600C4	COMPUTE	HPE
2M294600CP	COMPUTE	HPE
2M294600D6	COMPUTE	HPE
2M294600DC	COMPUTE	HPE
2M294600DG	COMPUTE	HPE

bon Emissions MTCO2e
0.006
0.006
0.001
0.007
0.006

Clouds > QA Azure (all)

QA Azure (all)

Location: All Regions   Last Sync: 02/07/2024 06:53 PM   Sync Duration: 57 seconds

$195	934.69	99%	13%	0%
COST THIS MONTH	COST HISTORY	MAX CPU	MEMORY	STORAGE

1	0	0	11	12
CONTAINER HOSTS	HYPERVISORS	BARE METAL	VIRTUAL MACHINES	DISCOVERED

$195.49	$934.69	$1,865.64	-$930.95 ↓ 50%
MONTH TO DATE	PROJECTED TOTAL	LAST MONTH	CHANGE FROM LAST MONTH

INVOICE	PERIOD	COMPUTE	MEMORY	STORAGE	NETWORK	EXTRA	MONTH TO DATE	PROJECTED TOTAL
3286757	February 2024	$27.66	$0.00	$142.24	$10.46	$15.13	$195.49	$934.69
3219432	January 2024	$231.68	$0.01	$1,377.17	$100.42	$156.36	$1,865.64	$1,865.64
3164022	December 2023	$244.71	$0.01	$1,348.04	$99.26	$3,155.36	$4,847.37	$4,847.37
3135865	November 2023	$219.48	$0.02	$1,302.95	$95.17	$151.34	$1,768.97	$1,768.97

Dashboards

Cost Savings

Automate

Scale up / down

AI Strategy on One Page

Models and Services	Accelerate the journey to Gen AI value for the Enterprise	**Model Hub** **HPE A&PS Services**	**Model-aaS**	Observability Sustainability FinOps
AI Platform	Drive collaborative AI model development from Data preparation to Deployment	Machine Learning Development Environment Machine Learning Data Management HPE AI Essentials	**AI PaaS**	
Data Services	Manage data governance, availability, and security	Data Fabric Data Observability Machine Learning Data Management		
Infrastructure Software	Allow AI models to run efficiently anywhere	Private Cloud for BMaaS, VMaaS, and CaaS.	**IaaS**	
Hardware Infrastructure	Optimize the performance of AI workloads with AI native infrastructure	Accelerators: Heterogenous support for Nvidia/AMD/Intel Compute: Cray, Proliant Storage: GL4FL, Networking: Slingshot, Mellanox Best in Class DLC		

AI Use Case Families in Financial Services: Spanning Front, Middle, and Back Office

FRONT OFFICE

MIDDLE OFFICE

BACK OFFICE

Credit Scoring	Customer Acquisition	Customer Experience	Customer Retention
Trading	Process Automation	Knowledge Management	Risk Management
Agent-Based Modelling	Stress Testing	Regulatory Compliance	Fraud Detection
Synthetic Data	Data Architecture	Data Capabilities	Infrastructure Optimization

Getting Started: Pick Your Use Case

	Regulatory Compliance	Credit Scoring
End-User Group	Internal Users Employee domain SME's (CPA's, compliance officers, Accountants, auditors, (CAO/CFO) Small Userbase, Highly Specialized Able to provide feedback on early iterations Low exposure to bad actors	External and/or Internal Users Customers or customer-facing employees (Loan officers, salespeople, brokers) Large Userbase, Experience levels may vary High exposure to bad actors
Use Case Regulations	Regulations mandate what to report to who and when, with specific instructions on preparation methodology "Automate as much as possible to minimize human error" -regulators	Fair lending, anti-discrimination, industry regs High-risk Use Case- AI Act (EU) GDPR/CCPA/Data protection regulations "AI can't do anything a human wouldn't be allowed to do" -regulators
Source Data (Unstructured)	Public information (Regulatory text, instructions, etc.) Trusted source- directly from regulators Low potential for wrongdoing if leaked	Non-public, Personally Identifiable Information Various sources- trustworthiness may vary High potential for wrongdoing if leaked Data sovereignty & data gravity concerns
Models Used	Open-source foundation model (Llama3 based) Out of box, no customization	Proprietary vendor or internally developed models Highly customized
Hardware Infrastructure	Small-scale inference environment or cloud API solution	Private, hosted, secure AI Inference at Scale cluster AI Factory training resources if building your own models

→

Development	Production
Security Guardrails Governance	Observability Sustainability FinOps

Internal Case Study:
Hewlett-Packard International Bank DAC

- Hewlett Packard International Bank DAC, ("HPIB") is a credit institution authorized by the Central Bank of Ireland and owned by HPE.
- The Capital Requirements Regulation 3 (CRR3) is a new regulation spanning 646 pages that goes into effect for banks in the EU on January 1, 2025. It is part of the wider "Basel 3.1"/ "Basel 3 Endgame" regulatory risk framework for global financial institutions.
- The CRR3 requirements have been consistently modified leading up to implementation, and portions still lacked fully defined requirements with less than 1 year to go.
- Risk jargon often requires a CPA with specific experience in the financial sector to understand these standards.

Lengthy and complex risk regulations

+

Highly technical subject matter

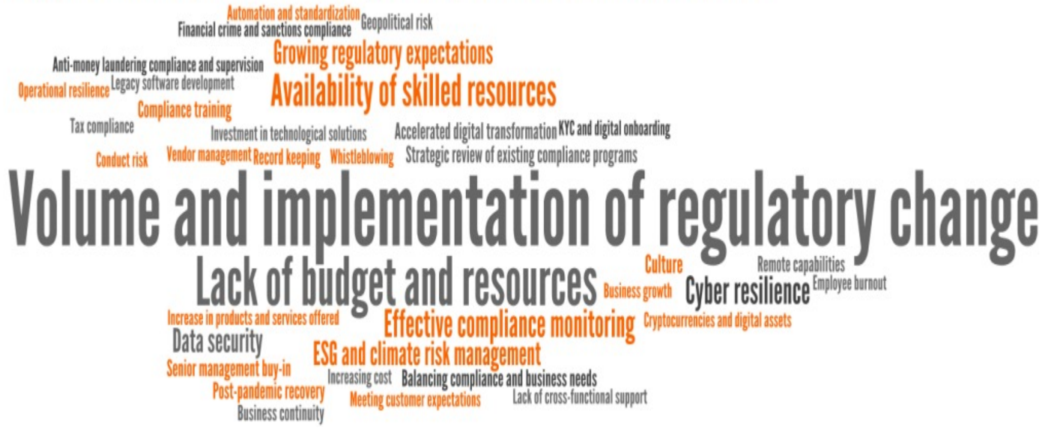
=

Complexity, Confusion, Delays, Increased risk

Problem Statement

- The financial services industry complies with a complex and fast-changing web of regulatory requirements.
- International firms are forced to adapt to the pace of change in the standards and laws of multiple regulators, jurisdictions, and governments
- Failure to comply can lead to civil (and even criminal) liability. Fines and sanctions can amount to **billions of dollars** for large institutions
- Each regulation or report change has the potential to have no impact, low impact, or very high impact and a long adoption process. Expert interpretation is required to determine “Does this change affect us?” and “What do we need to do differently than before?

The greatest compliance challenges I expect to face in 2022 is/are...


Automation and standardization
Financial crime and sanctions compliance
Geopolitical risk
Anti-money laundering compliance and supervision
Growing regulatory expectations
Operational resilience
Legacy software development
Availability of skilled resources
Compliance training
Tax compliance
Investment in technological solutions
Accelerated digital transformation
KYC and digital onboarding
Conduct risk
Vendor management
Record keeping
Whistleblowing
Strategic review of existing compliance programs
Volume and implementation of regulatory change
Lack of budget and resources
Culture
Remote capabilities
Business growth
Cyber resilience
Employee burnout
Increase in products and services offered
Effective compliance monitoring
Cryptocurrencies and digital assets
Data security
ESG and climate risk management
Senior management buy-in
Increasing cost
Balancing compliance and business needs
Post-pandemic recovery
Meeting customer expectations
Lack of cross-functional support
Business continuity

Initial Experiments

Experiment 1

Bulk analysis of documents on regulator website via Python to produce categorized CRR3 Implementation Report

Experiment 2

Analysis triggered by frequent regulator website updates, which can provide automated guidance and alerts for impactful changes and opportunities to comment

Experiment 3

Chatbot experience for real-time self-service with follow-up questions

Key Benefits

Transparency: End-to-end tracking of new proposed regulatory changes, opening and closing of public comment periods, adoption of final changes, and phased implementation deadlines.

Operational efficiency: Swift document analysis and reporting. Give users opportunity to ask questions on complex subject matter in natural language and receive proactive updates on what they need to know.

Cost savings: Minimize burden of adapting to new legislation, manual reviews and potential fines for non-compliance, and consulting spend.

Increase collaboration: Create wider awareness of upcoming changes.

Part of a Holistic Compliance Strategy

Generative AI analysis is one part of a comprehensive strategy

AI-Powered Analysis:

- Swift document review and extraction of essential data.
- Predictive modeling for potential compliance risks.
- Real-time notifications on regulatory changes.
- Simplification of complex regulatory texts using NLP.

Training & Development:

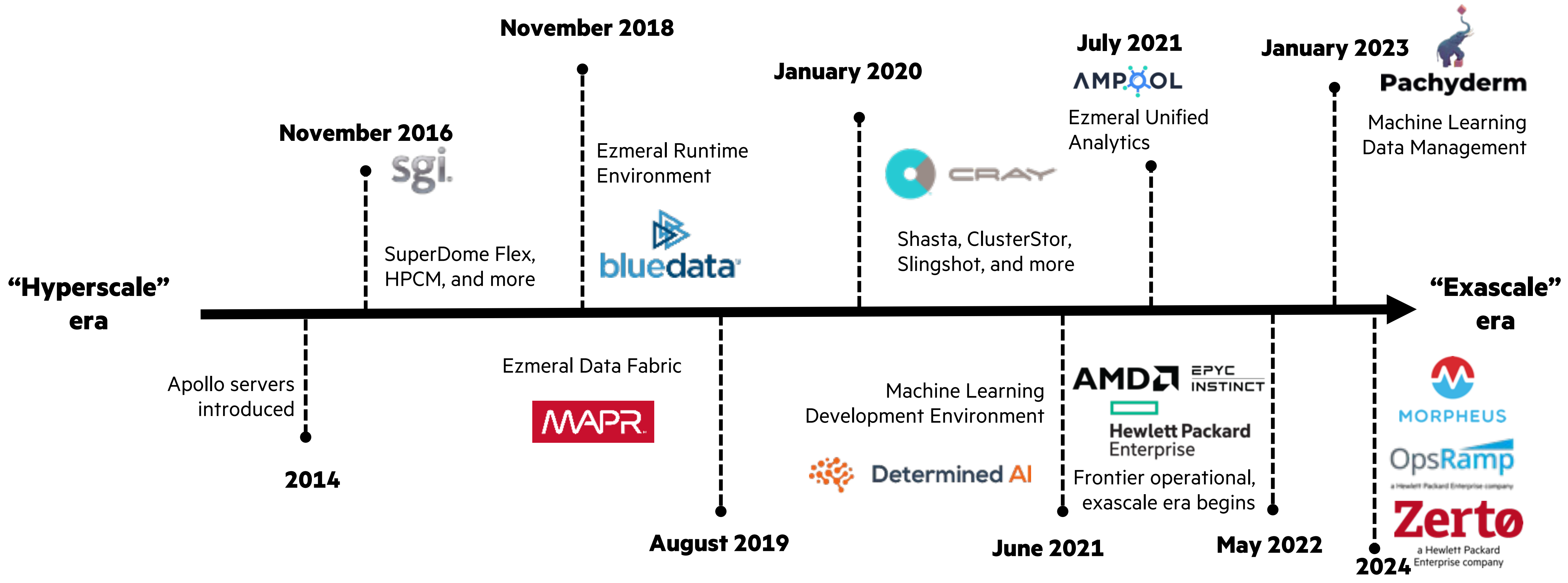
- Regular workshops to educate staff on compliance matters.
- Role-specific training modules for targeted learning.
- Case studies and simulations for hands-on experience.

Human Oversight:

- Expert panels for manual review and validation.
- Routine audits to ensure AI-generated results align with compliance requirements.
- Feedback loops for continuous AI model refinement.

HPE Has Been Preparing for Some Time

HPC + AI WALL STREET



“Hyperscale” era

2014

Apollo servers introduced

November 2016

sgi

SuperDome Flex, HPCM, and more

November 2018

Ezmeral Runtime Environment

bluedata

August 2019

Ezmeral Data Fabric

MAPR

January 2020

CRAY

Shasta, ClusterStor, Slingshot, and more

Machine Learning Development Environment

Determined AI

June 2021

AMD EPYC INSTINCT

Hewlett Packard Enterprise

Frontier operational, exascale era begins

July 2021

AMPOOL

Ezmeral Unified Analytics

May 2022

January 2023

Pachyderm

Machine Learning Data Management

2024

MORPHEUS

OpsRamp a Hewlett Packard Enterprise company

Zerto a Hewlett Packard Enterprise company

“Exascale” era

Enabling Artificial Intelligence

HPE GreenLake

Models and Services

Open-source models

Commercial partnerships

Services to better drive AI models

+

AI Platform

Data management

Model training

Model inference

+

Infrastructure

From bare metal to containers

From training to inference

From supercomputer to edge

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

HPC + AI WALL STREET

2024

INNOVATIONS IN ADVANCED ... FOR FINTECH AND CAPITAL MARKETS
HPC + AI WALL STREET
INTELLIGENT GLOBAL FINTECH
Quantum Computing
High Performance Computing
Data Management
Artificial Intelligence
EXIT

Thank You

HPC | QUANTUM | DATA | AI

Hewlett Packard Enterprise | AMD