Red Hat Summit

Connect

AI Optimization: from Intelligent Training to Distributed AI

Giovanni Spina, *CTO Accenture Cloud Innovation Center, Rome*, Accenture

Alessandro Tringali, *Senior Manager Data & AI*, Accenture

accenture | Red Hat

Who are we

Giovanni Spina
Chief Technology Officer,
Accenture Cloud
Innovation Center,
Accenture

Alessandro Tringali
Senior Manager
Data & AI,
Accenture

Our Innovation Center

Accenture Cloud Innovation Center Rome
via Ostiense 92, Rome
https://www.accenture.com/it-it/services/cloud/innovation-center-roma

www.accenture.it
www.accenture.com

Agenda

- Executive Summary
- Approaches to AI Optimization
- Red Hat Openshift AI for distributed AI
- LLM tuning results

Executive Summary

Need

- Commercial LLMs are trained on massive public datasets and lack **enterprise-specific knowledge**.
- External or cloud-based models raise **data sovereignty and confidentiality risks**.
- The **growing demand for high-performance, expensive hardware** is accelerating as AI workloads become increasingly complex

Approach

- **Sovereign AI framework** for enterprise ownership of data, models, and infrastructure.
- **Fine-Tuning & SFT** on proprietary datasets for business language and logic.
- Leveraging **distributed AI and scalable infrastructure** to optimize resources and minimize reliance on expensive hardware

Value

- Shift from Prompt Engineering to **System Design with AI Optimization**.
- Build **specialized LLMs** fully aligned with enterprise data and logic.
- Improve **accuracy, performance, and cost-efficiency** – a concrete step toward a **Sovereign AI ecosystem**.

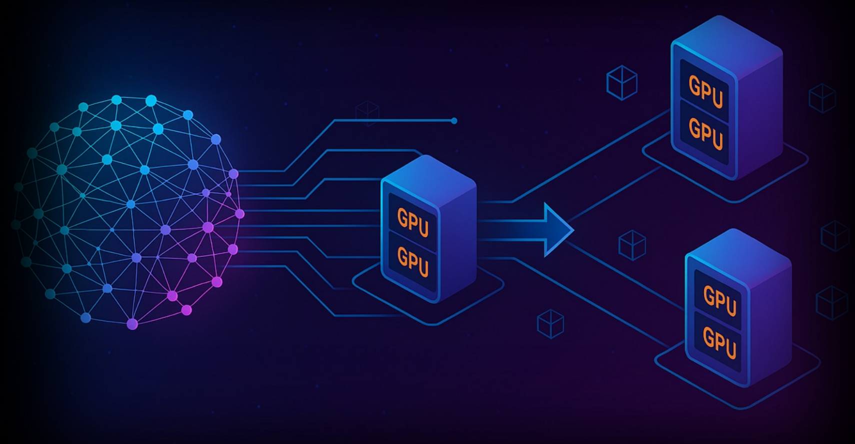
Our case study

Goal	Prove the efficiency of Fine Tuning and Reinforcement Learning on a medium sized model in a Sovereign Environment, in terms of accuracy and ability to save computational resources.
Platform	ACIC Private AI Platform (Dell + Red Hat). An on-prem infrastructure, able to accelerate time to value, simplify management, and assist in creating a security focused AI environment.
Optimize	Implementation of load distribution strategies leveraging Red Hat OpenShift, Red Hat OpenShift AI and Red Hat Distributed Training Operator, to minimize hardware requirements and maximize the efficiency of the Private AI platform.
Model & Domain	We fine-tuned Phi-3-mini-128k-instruct on CyberSecurity domain, using CyberMetric for accuracy evaluation.
Training	The process included 4 phases, each using dedicated datasets: Pre-Training on a large CyberSecurity corpus, Instruction Fine-Tuning, Reasoning Fine-Tuning and a final Reinforcement Learning phase.
Result	We increased the accuracy of the model to match the accuracy of 2X parameters Model while cutting training time by over 50% and maximizing hardware efficiency through distributed AI on Red Hat OpenShift AI.

Approaches to AI Optimization

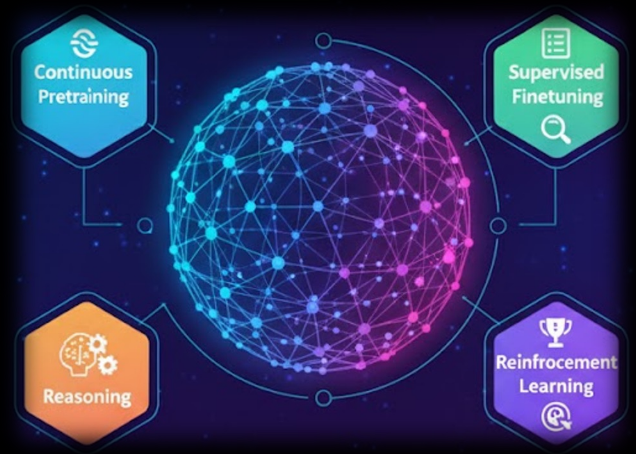
Distributed AI

This involves distributing the workload of fine tuning and inference of AI models across multiple resources/hardware , leveraging on all GPU, to improve efficiency, scalability and costs

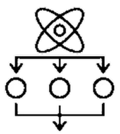


LLM Tuning Process

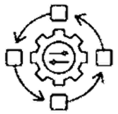
By combining Continuous Pretraining, Supervised Finetuning, Reasoning, and Reinforcement Learning, organizations can build AI systems that continuously learn, adapt, and improve over time. This holistic tuning process enhances model accuracy, contextual understanding, and responsiveness.



Distributed AI



LLM Tuning Process



Red Hat Openshift AI for distributed AI

Red Hat® OpenShift® AI is a flexible, scalable artificial intelligence (AI) and machine learning (ML) platform that enables enterprises to create and deliver AI-enabled applications at scale across hybrid cloud environments

Key Features:

- Built using **open source technologies**
- End-to-end AI model **lifecycle management**
- **Kubernetes-native**, container-based architecture
- **GPU acceleration and optimized inference performance**
- **Collaborative and self-service** AI workspace
- Enterprise-grade **governance and security**
- Hybrid-cloud and multi-environment **deployment flexibility**

One technology serving diverse use cases

Standard Cluster

Standard solution based on infrastructure designed according to Red Hat best practices. It can deliver private AI services (safe by design) and enable GPU-as-a-Service and Model-as-a-Service scenarios

Single Node

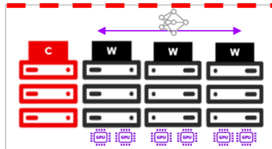
Using OpenShift AI on Single Node architectures allows the deployment of artificial intelligence in resource-constrained environments (including low-cost DEV/TEST activities), while maintaining operational consistency and enterprise-grade security. It is also suitable for EDGE scenarios

Single Node – Multi GPU

Deploying OpenShift AI on a Single Node with multi-GPU support allows you to maximize computing power even in compact or decentralized environments, while retaining all the advantages of the OpenShift platform (security, governance, automation).

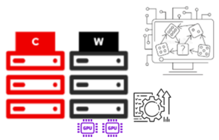
Multi Node – Multi GPU

Using OpenShift AI on a standard infrastructure with 3 worker nodes and 3 control plane nodes, enhanced with multi-GPU solutions, enables combining the horizontal scalability typical of OpenShift clusters with the vertical scalability provided by multiple GPUs per node



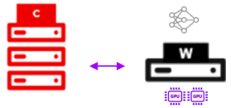
Distributed AI

Not just horizontal and vertical scalability—it's also possible to adopt solutions that parallelize resource-intensive tasks across all cluster nodes, optimizing infrastructure usage and reducing the need for additional investments



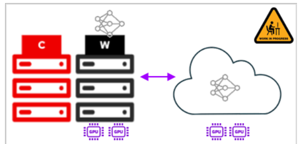
AI Inference Optimization

Red Hat AI Inference Server provides fast and cost-effective inference at scale. Powered by vLLM, the inference server maximizes GPU utilization and enables faster response times increasing efficiency without sacrificing performance.



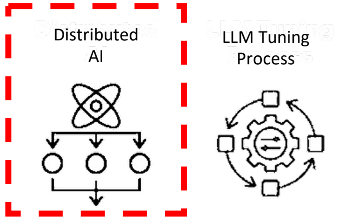
Remote Worker/Edge

Remote Worker Nodes bring AI/ML processing closer to data sources, reducing bandwidth usage and enabling real-time decisions, while remaining part of the central cluster to ensure consistent security, governance, and automated MLOps across distributed environments.

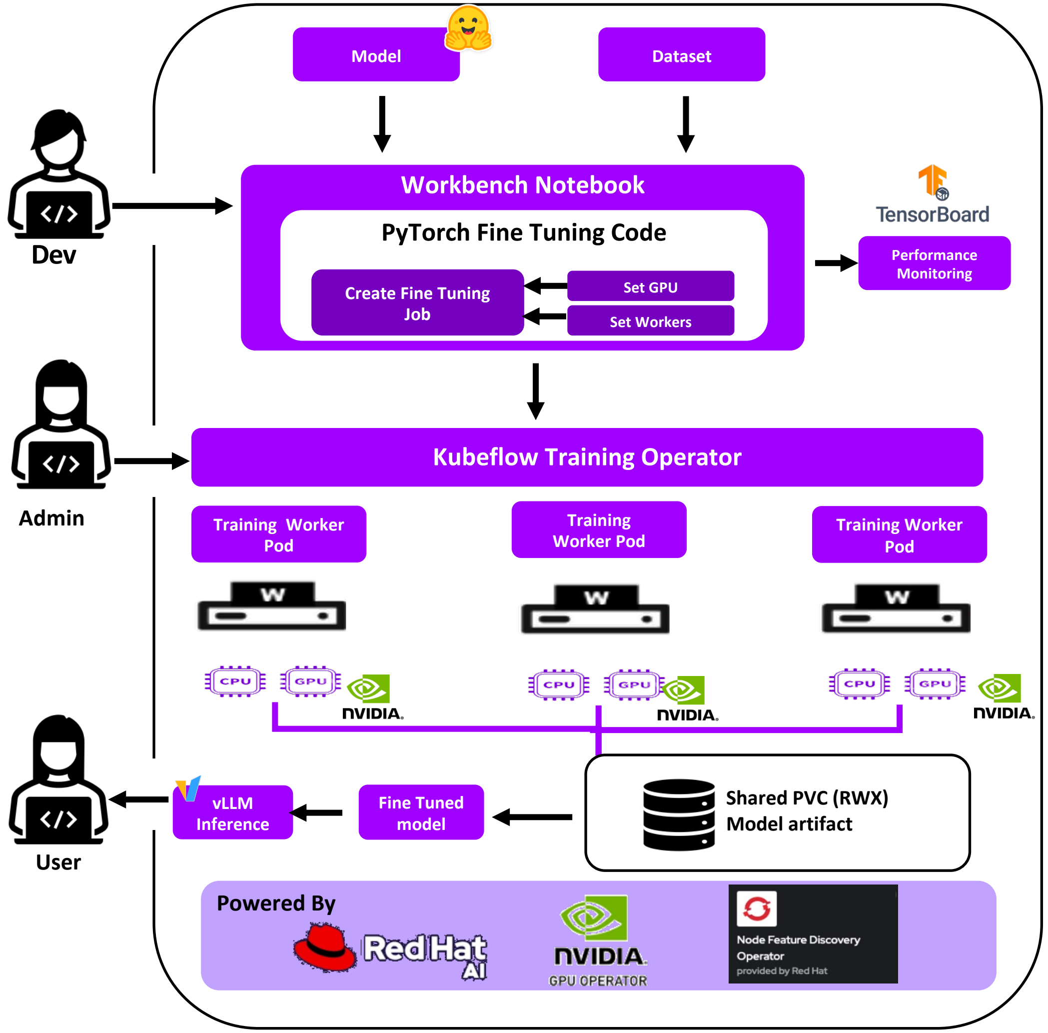


Secure Cloud Burst

Leveraging Confidential Containers, it's possible to extend the OCP environment to the cloud. This enables dynamic management of infrastructure resource needs, avoiding service disruptions and reducing the need to purchase new servers for short-term demands



Distributed Fine Tuning from Red Hat



Red Hat enables enterprises to streamline model fine-tuning and distributed training through the **Kubeflow Training Operator** on OpenShift AI. By integrating the operator into the OpenShift ecosystem, Red Hat provides a consistent, cloud-native way to orchestrate **large-scale AI/ML workloads** across Kubernetes clusters

- The Kubeflow Trainer Operator is designed to **facilitate** distributed fine tuning of ML complex models on Kubernetes clusters.
- Is a Kubernetes-native project for fine-tuning and scalable distributed training created with different ML frameworks (PyTorch, TensorFlow)
- Lowers infra costs by **offloading data** loading and model initialization to CPUs and streamlining asset distribution across training nodes, keeping GPUs focused on computation.

* By leveraging GPU Direct RDMA with high-speed interconnects like **InfiniBand or RoCEv2** in OpenShift AI, organizations can, significantly reduce tuning time, improve resource efficiency, and accelerate AI time-to-market.

AI Optimization: from intelligent Training to Distributed AI

Accenture's approach to LLModels domain optimization

Distributed AI

LLM Tuning Process

LLM Tuning Process

01

Continuous Pretraining

Continue the pretraining phase well beyond the official release

02

Supervised Fine Tuning

Teach the model what to know and how to respond

03

Reasoning

Teach the model what knowledge to specialize in and how to apply it

04

Reinforcement Learning

Let the model learn from feedback and improve its response

How

Unsupervised learning
Expose the model to large amounts of domain-specific data to learn the language and context of that field

Supervised learning
Provide questions and answers to train instruction-following and fine-tune on domain-specific content

Supervised reasoning training
Use structured questions and answers emphasizing logical steps and analytical thinking

Reinforcement learning with human feedback (RLHF)
Reward better answers to refine style, clarity, and compliance

Purpose

Enhance domain knowledge and make the model “speak” the domain language

Ensure the model **follows instructions** and **becomes** a reliable domain **expert**

Enable the model to use the “think” process effectively, ensuring that its reasoning is **useful**, **analytical**, and **logical**

Allow the model to **learn from its mistakes**, improving choices of wording, conciseness, and ethical alignment

LLM Tuning Results

Methodological

	Accuracy
Base (Phi-3)	81,8%
Our Fine-Tuned Model	84,2%

+2.4 points accuracy gain matching a model with **2× parameters**

Infrastructural

Pretraining	Cybersecurity corpus *(Unsupervised)*	**22h -> 8h**
Reasoning Fine-Tuning	Math/Logical dataset	**8h -> 4h**
Instruction Fine-Tuning	Domain-specific data	**14h -> 6h**

Over 50% reduction in total training time with consistent efficiency gains across pretraining, reasoning, and instruction fine-tuning phases.

Pretraining and **Supervised Fine-Tuning** significantly **enhance** the performance of **smaller models**, enabling them to match or even surpass larger architectures

GenAI Solutions – The Paradigm Shift

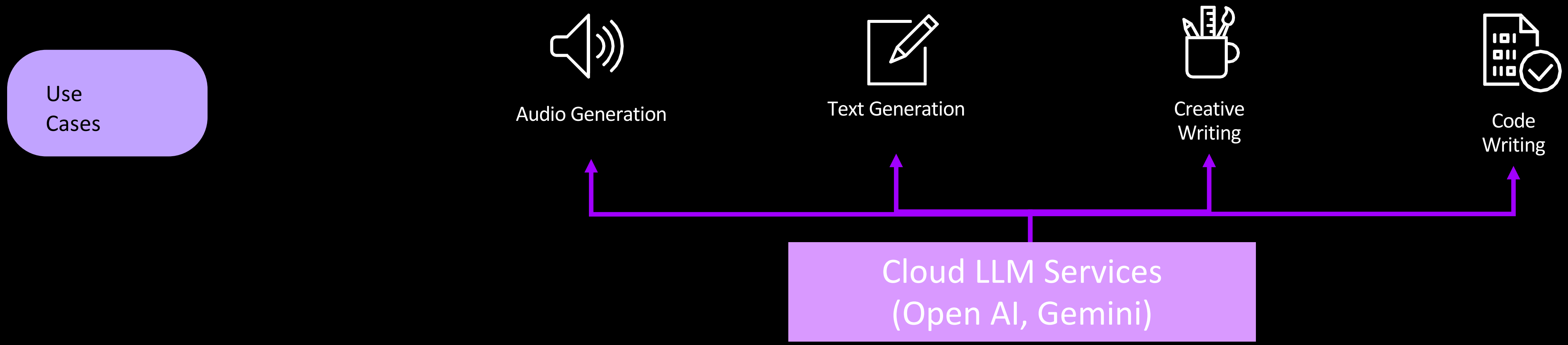
From API-first to Model Optimization Era

	As Is: API-first		To Be: Model Optimization
Mindset	**API Consumers:** we utilize external resources by a pay-per-token basis.	›	**System Architect:** we build our own solution by analyzing and optimizing our resources.
Goal	**A «One-solution-fits-all» approach:** through Prompt Engineering, we rely on Commercial Models to answer in the proper manner.	›	**Optimal Performance & Cost Efficiency:** we tailor our models and training data to our processes, building clusters of LLMs that solve very specific needs.
Control	**Black Box:** we have no control on the model behavior, except from Prompt Engineering, making the solution weak to model hallucinations.	›	**Glass Box:** through data manipulation and curation, we can train our models to complete tasks in the way we see fit.
Focus	**General Response:** we rely on the model to follow our prompt at the best of its capabilities.	›	**Specialized models:** we have full control on each node of our Client's processes.

It's about engineering an optimized solution—**balancing accuracy, performance, cost, and the ability to support Process reinvention through AI**

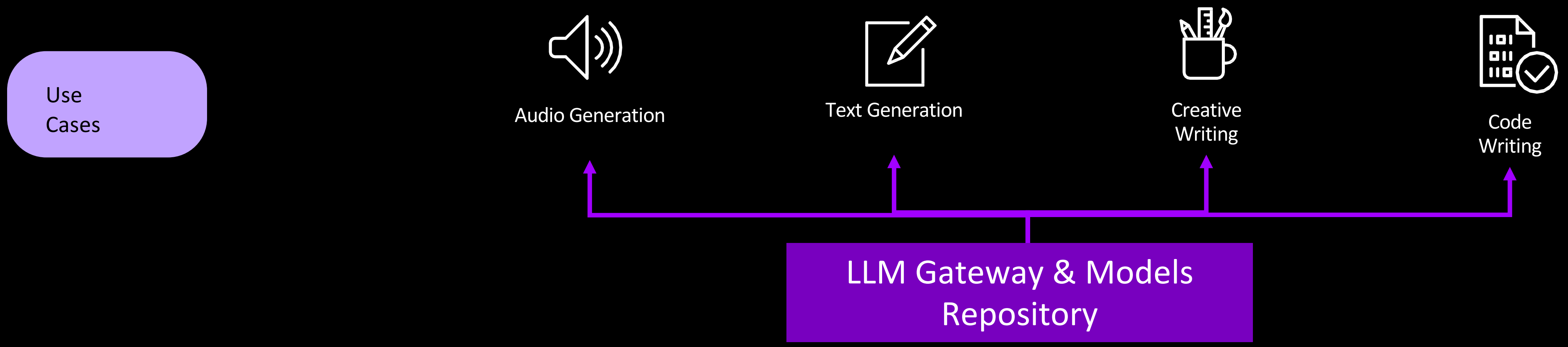
GenAI Solutions – Our Vision

The Model Optimization Era



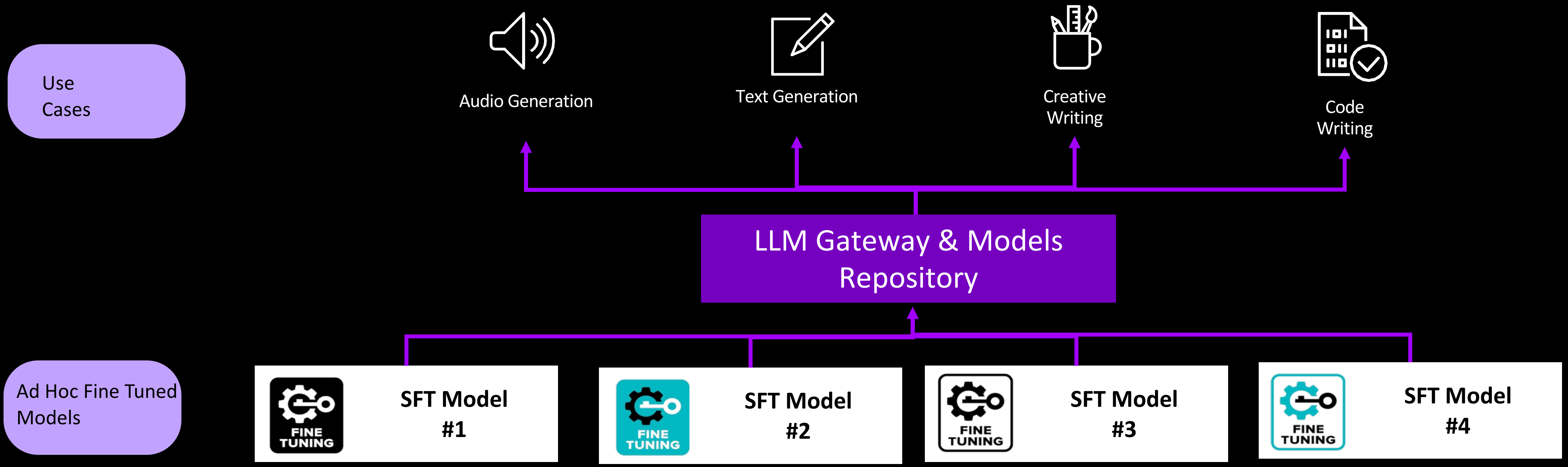
GenAI Solutions – Our Vision

The Model Optimization Era



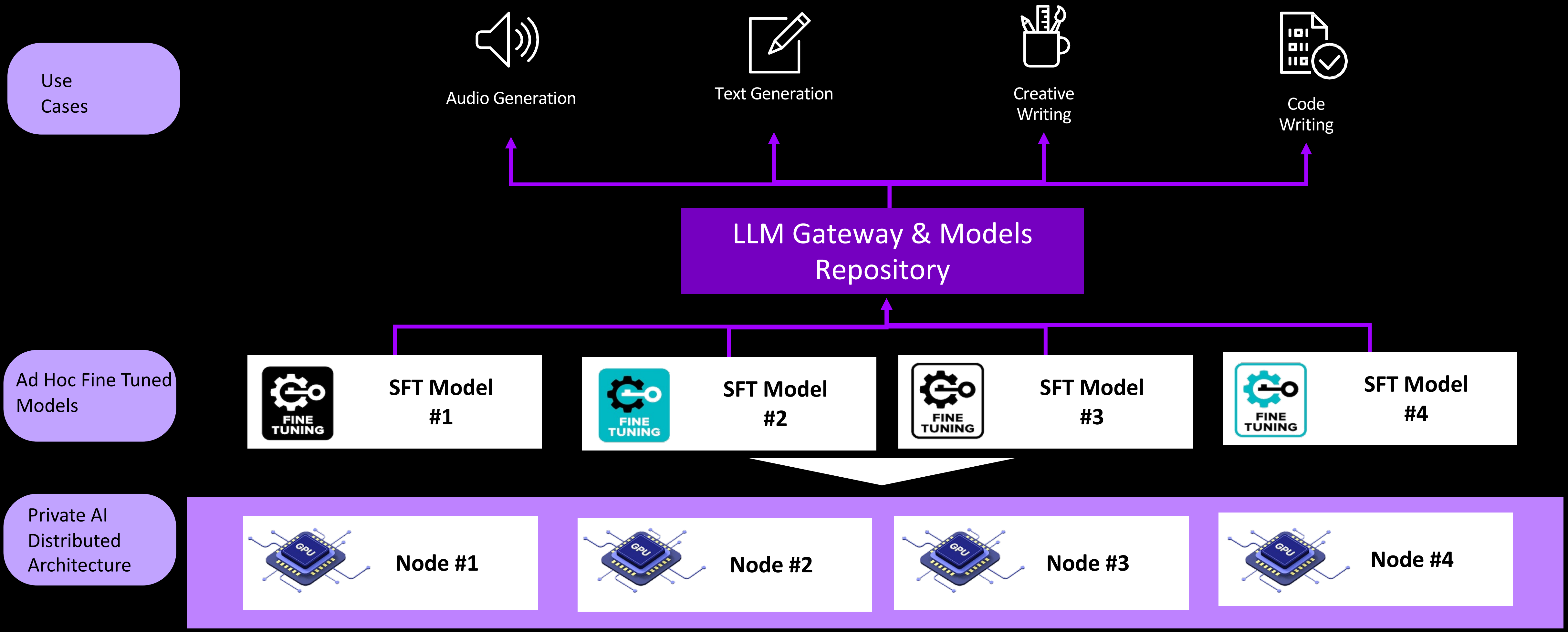
GenAI Solutions – Our Vision

The Model Optimization Era



GenAI Solutions – Our Vision

The Model Optimization Era



Thank you