Starting Now

<packt>

Optimizing the Fine-Tuning and Deployment for GenAI

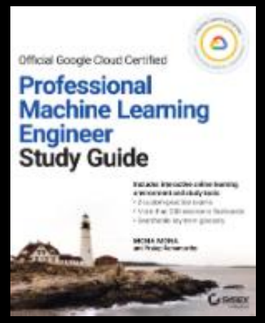
Learn how you can fine tune LLM on Amazon SageMaker AI

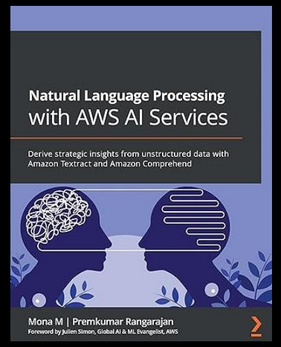
Mona Mona

Sr Gen AI Solutions Architect, AWS

Optimizing the Fine-Tuning and Deployment for GenAI

Learn how you can fine tune LLM on Amazon SageMaker AI





Mona Mona

Sr Gen AI Solutions Architect, AWS

About Mona Mona

- Mona currently works as a Sr WW Gen AI Specialist Solutions Architect at Amazon focusing on Gen AI Solutions.
  She was a Lead Generative AI specialist in Google Public Sector at Google before joining Amazon.
- Mona is a published author of two books: Natural Language Processing with AWS AI Services: Derive strategic insights from unstructured data with Amazon Textract and Amazon Comprehend and Google Cloud Certified Professional Machine Learning Study Guide.
- Her first book is on the course curriculum of Georgia State University's Masters in Computer Information Systems program, where she has been a guest lecturer.
- She has authored 20+ blogs on AI/ML and cloud technology and a co-author on a research paper on CORD19 Neural Search which won an award for Best Research Paper at the prestigious AAAI (Association for the Advancement of Artificial Intelligence) conference.
- Mona Mona is a member of SWE (Society for women engineers) and she mentors early career graduates looking to pursue careers in AI and Cloud Computing. She is also a frequent speaker at multiple conferences such as ISMB 2022, AWS Re:Invent 2020,2019,2018 and AWS DC Summit.

Agenda

01 Generative AI use cases

02 Why fine-tuning?

03 Fine tuning and optimization technique

04 How AWS can help

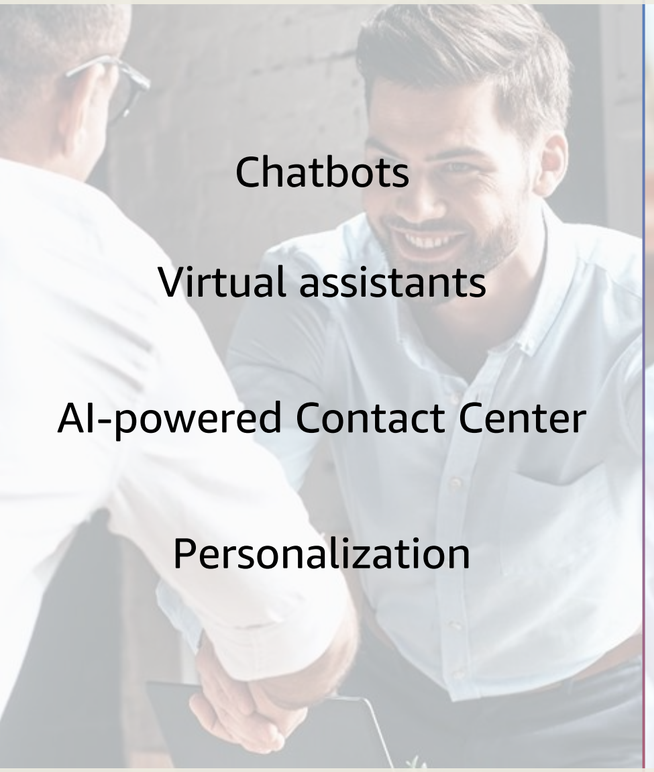
05 LLM Training on AWS

06 LLM Deployment options on AWS

07 Demo

Section 1: Generative AI use cases

Generative AI is powering multiple use cases



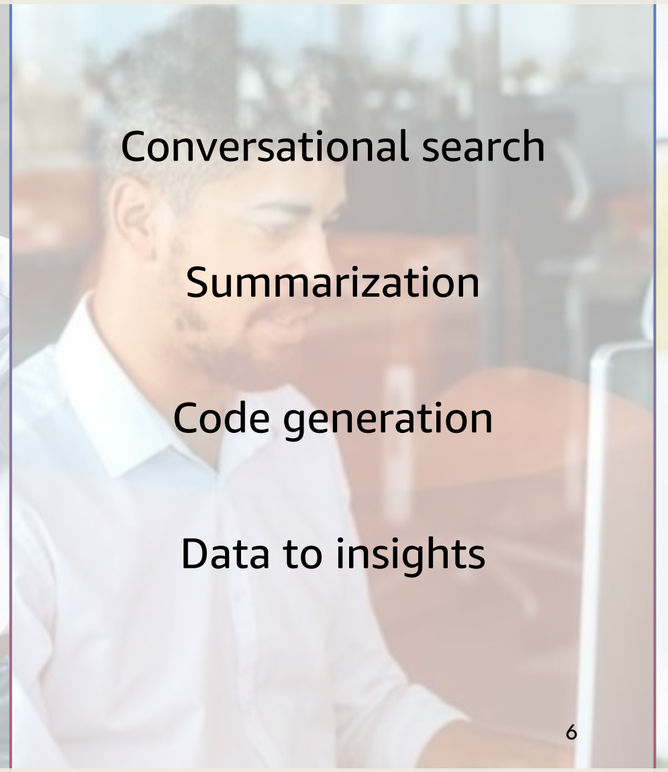
Chatbots

Virtual assistants

AI-powered Contact Center

Personalization

Enhance customer experience



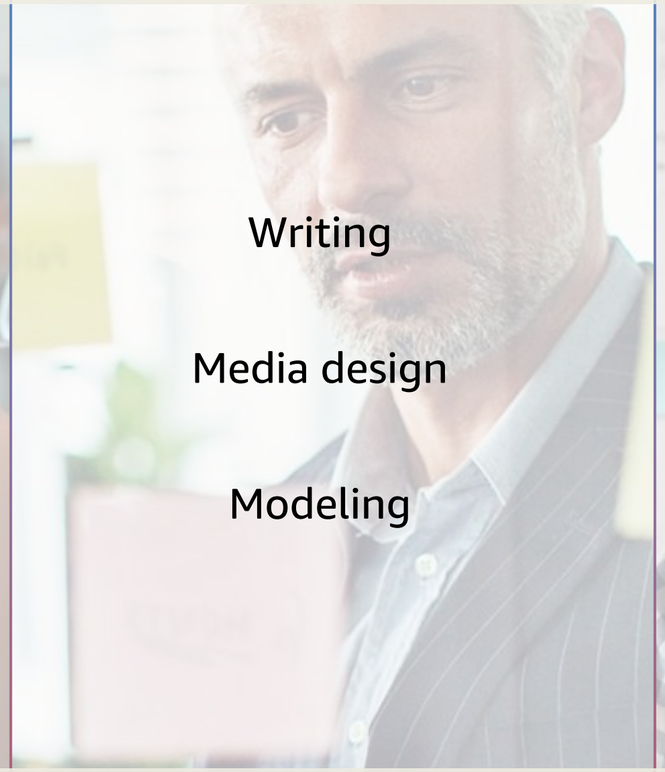
Conversational search

Summarization

Code generation

Data to insights

Boost employee productivity

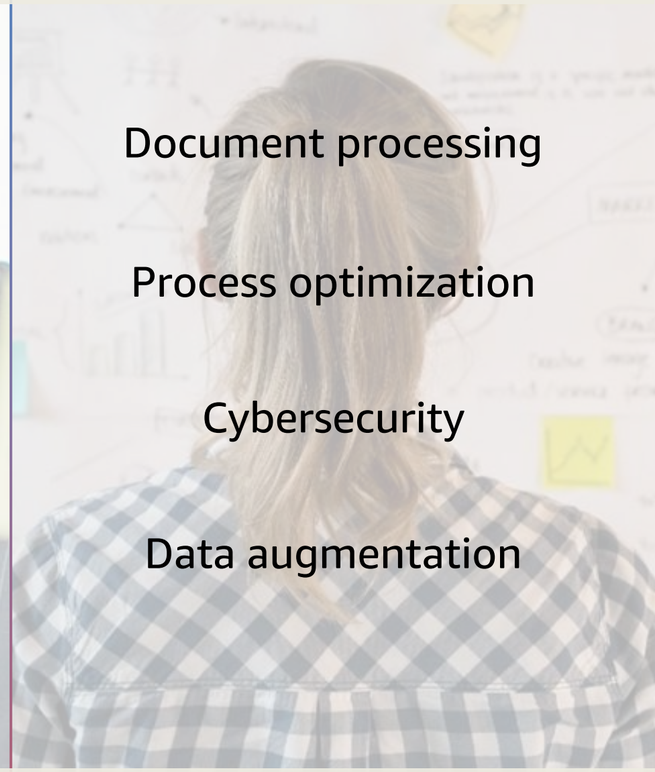


Writing

Media design

Modeling

Creativity and content creation



Document processing

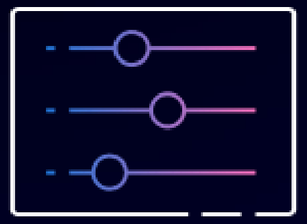
Process optimization

Cybersecurity

Data augmentation

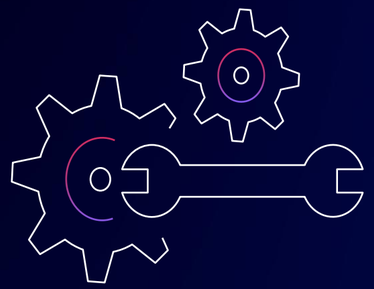
Improve business operations

How do customers implement these use cases ?



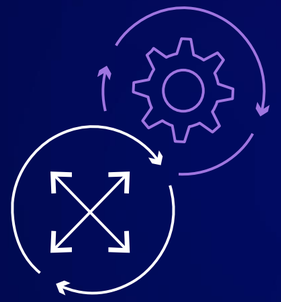
Foundation

Choose the right Foundation Model (FM) for your use case



Customization

Connect models to business data through RAG and/or fine-tuning



Orchestration

Build end-to-end workflows with orchestration tools

Deployment

Balance performance, cost, and efficiency at scale

Section 2: Why fine-tuning?

Fine-Tuning

What

Fine-tuning allows you to customize an FM for specific tasks or to adapt to specific domains

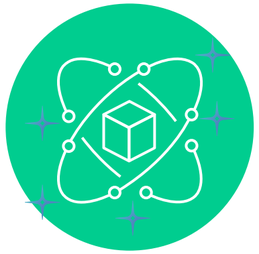
Who

It’s intended to be used mostly by ML Practitioners

When

1) Improve performance on a specific task.
2) Customize outputs characteristics.
3) Adapt to new data .

Key benefits of customizing FMs



Tailored results

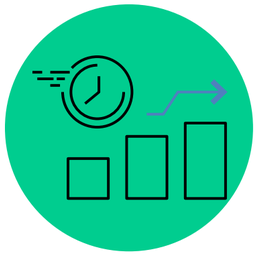
Enhances model quality and relevancy specific to specific use cases, industries, preferences and styles

Mitigates bias and risks

Mitigates risk of non-compliance, toxicity, and hallucination

Improved performance

Improves model inference, performance, and cost



Optimization and scale

Critical part of the GenAI model lifecycle management

Customizing FMs is critical to ensuring model quality and relevancy

In-Context Learning
Does not alter model weights/parameters

Fine-tuning
Alters model weights/parameters

In context learning	Retrieval Augmented Generation (RAG)	Domain Adaptation Fine-Tuning (DAFT)	Re-Training
Using an FM with few/no additional prompts	Augmenting prompts with context retrieved from large domain-specific databases	Develop more precise context-aware LLMs based on un/semi-supervised learning	Parameter Efficient re-training of model parameters with new training data

Prompt Engineering	Supervised Fine-Tuning (SFT)	Instruction Fine-Tuning (IT)
Design prompts to make the outputs accurate, relevant to adapt model for your needs	Adjusting pre-trained FMs on a specific dataset with labeled examples	Model is fine-tuned on the data, with specific instructions or guidelines, and formats

Customer Benefits: Fine-Tuning Open Source FMs on Amazon SageMaker AI

Boost Employee Productivity

50%

lower costs for hosting FMs

Boost Employee Productivity

7 months

reduced time-to-value from 12-18 months

Enhance Customer Experience

80%

reduction in inference latency

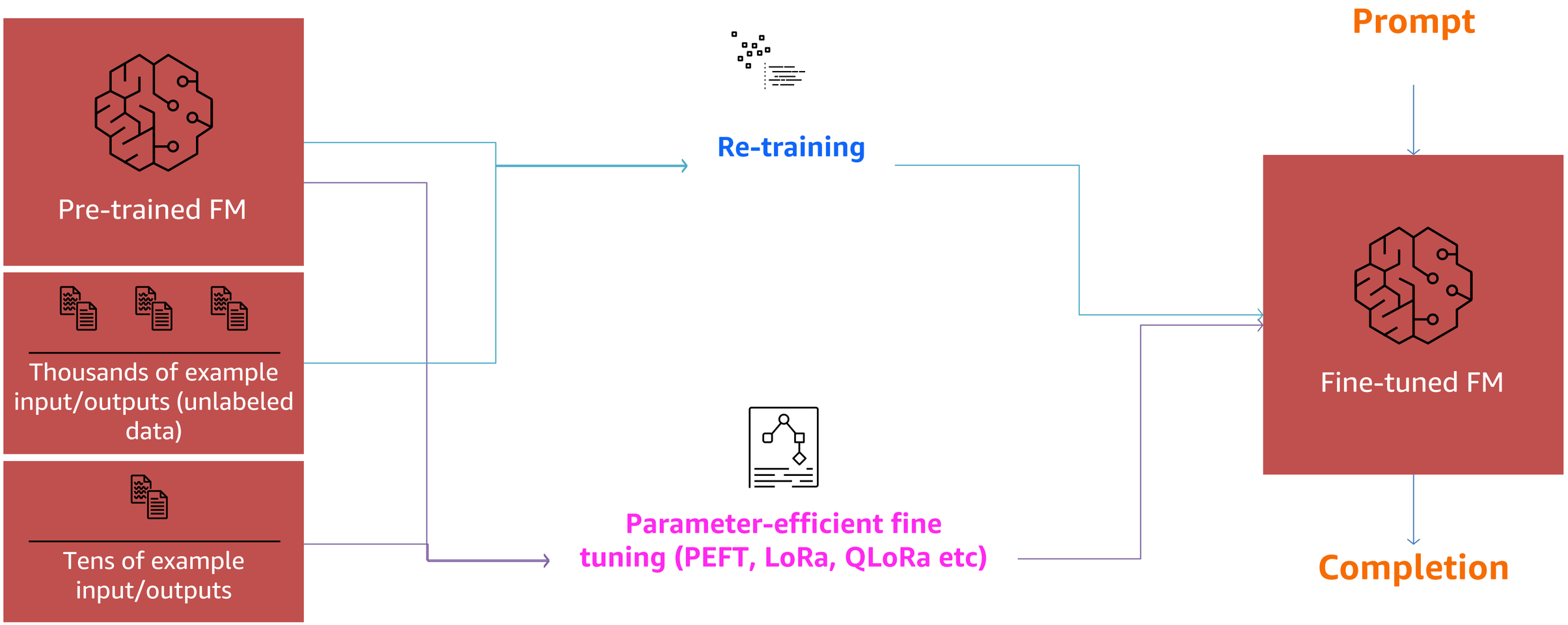
Streamline Business Processes

66%

cost savings with GPU utilization

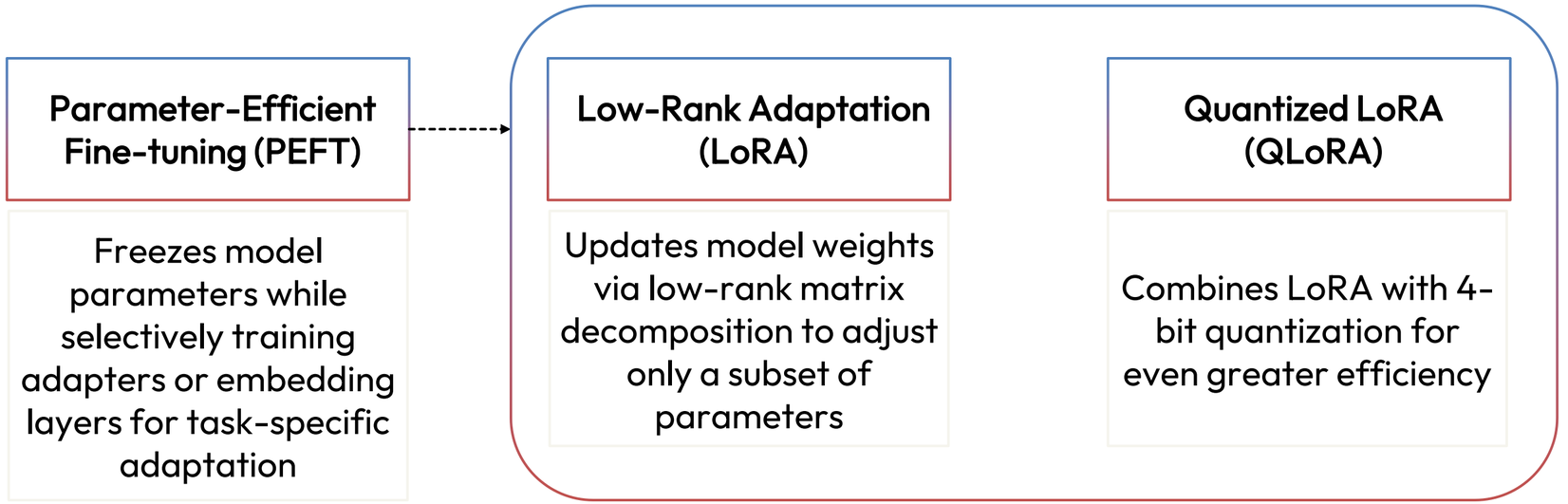
Section 3: Fine-tuning and optimization techniques

Full fine tuning (domain adaptation) vs Advanced fine-tuning (PEFT)



Optimization techniques

Streamline LLM fine-tuning/retraining by reducing computational costs, memory usage, and parameter updates while maintaining or enhancing model performance



Optimization techniques enable efficient LLM adaptation through methods like LoRA (parameter reduction), QLoRA (quantized memory savings), and PEFT (targeted updates) for **domain specialization or RLHF (alignment tuning), quantization (speed boosts), and knowledge distillation (model compression)**

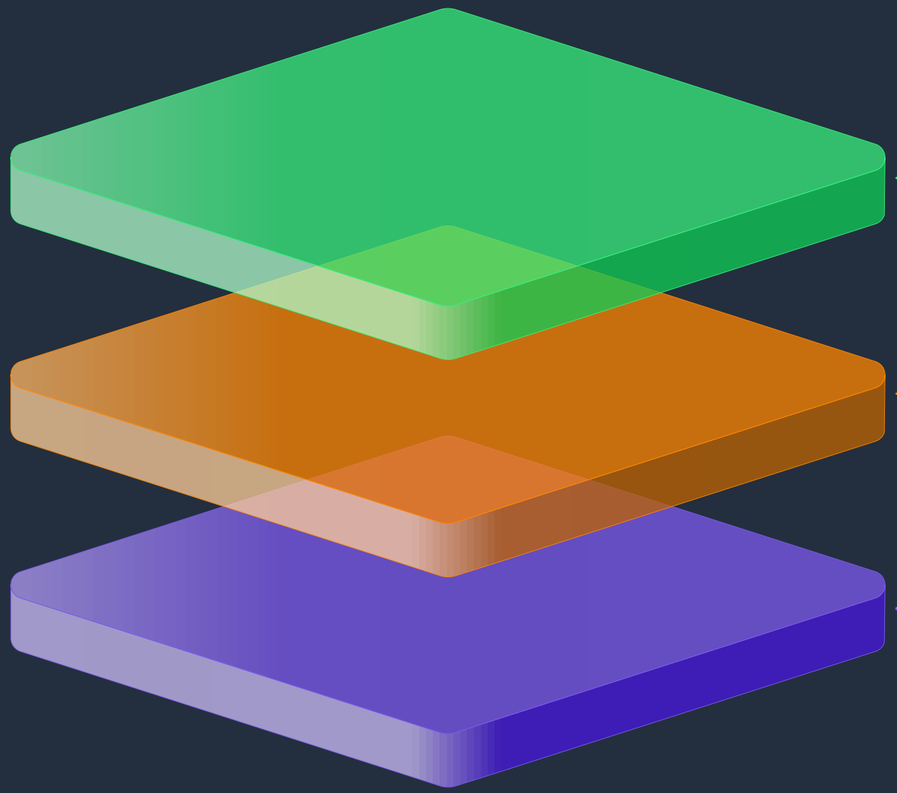
Optimization techniques

Decision criteria

Technique	Computational Efficiency	Memory Usage	Resource Constraints	When to Use
LoRA	High (updates only a small number of parameters)	Moderate (low-rank matrices add some overhead)	Moderate	Use when fine-tuning is needed with moderate resources.
QLoRA	Very High (adds 4-bit quantization for efficiency)	Very Low	Very Low	Use in resource-constrained environments, where both memory and computational efficiency are critical.

Section 4: How AWS can help

AWS Generative AI Stack



APPLICATIONS TO BOOST PRODUCTIVITY

Amazon Q Business
INSIGHTS AND AUTOMATION

Amazon Q Developer
SOFTWARE DEVELOPMENT LIFECYCLE

MODELS AND TOOLS TO BUILD GENERATIVE AI APPS

Amazon Bedrock
AMAZON MODELS | PARTNER MODELS

INFRASTRUCTURE TO BUILD AND TRAIN AI MODELS

Amazon SageMaker AI
MANAGED INFRASTRUCTURE

AWS Trainium
AWS Inferentia

GPUs

HIGH PERFORMANCE COMPUTE

Section 5: LLM training on Amazon SageMaker

Approaches for Fine-Tuning FMs on SageMaker

Flexibility to choose the way you want to build

Amazon SageMaker Jumpstart

SKILLSET

- Low Code

WHY YOU USE IT

- Faster time to market
- Maximizing accuracy for specific tasks

CONSIDERATIONS

- Experimentation

Amazon SageMaker Studio

SKILLSET

- Interactive Notebook

WHY YOU USE IT

- Can immediately see the response
- Quicker Debugging

CONSIDERATIONS

- Single node
- Good for testing

Amazon SageMaker Training

SKILLSET

- Code

WHY YOU USE IT

- Scale
- Resiliency to maintain the cluster

CONSIDERATIONS

- Need to reserve a cluster

Amazon SageMaker Training options

Large scale model training and fine-tuning

SageMaker Training Jobs

Fully managed resilient infrastructure for large-scale and cost-effective training

- Focus on model building rather than IT
- Provide access to flexible on-demand GPU cluster with a pay as you go option

SageMaker HyperPod

Resilient and self orchestration infrastructure for maximum resource control

- Customize and manage cluster orchestration (Slurm or EKS)
- Schedule workloads to maximize cluster utilization across teams

Fine-Tuner Personas and SageMaker Solutions

Fully-managed

Primary focus on model architecture, they need flexibility in experimentation, training, and optimization techniques

Amazon SageMaker Training

Access any open-source model, build your own model with the most popular open-source framework, and optimization techniques

Managed

Primary focus on research, requires a highly scalable and resilient environment for large training workloads

Amazon SageMaker Hyperpod

Highly scalable and resilient training environments, suited for large-scale machine learning workloads. It provides persistent cluster infrastructure across thousands of AI accelerators

Fine-tune with SageMaker AI with 3 Steps

Decide training framework

Choose framework and reserve capacity

Run fine-tuning at scale

SageMaker training jobs

Monitor your training

MLflow

1 Choose framework and reserve capacity

OPTIMIZED DISTRIBUTED TRAINING LIBRARIES & FRAMEWORKS

TensorFlow	PyTorch	Hugging Face	SageMaker Distributed Training Libraries	Bring your own library (e.g., Nemo Megatron, FSDP)

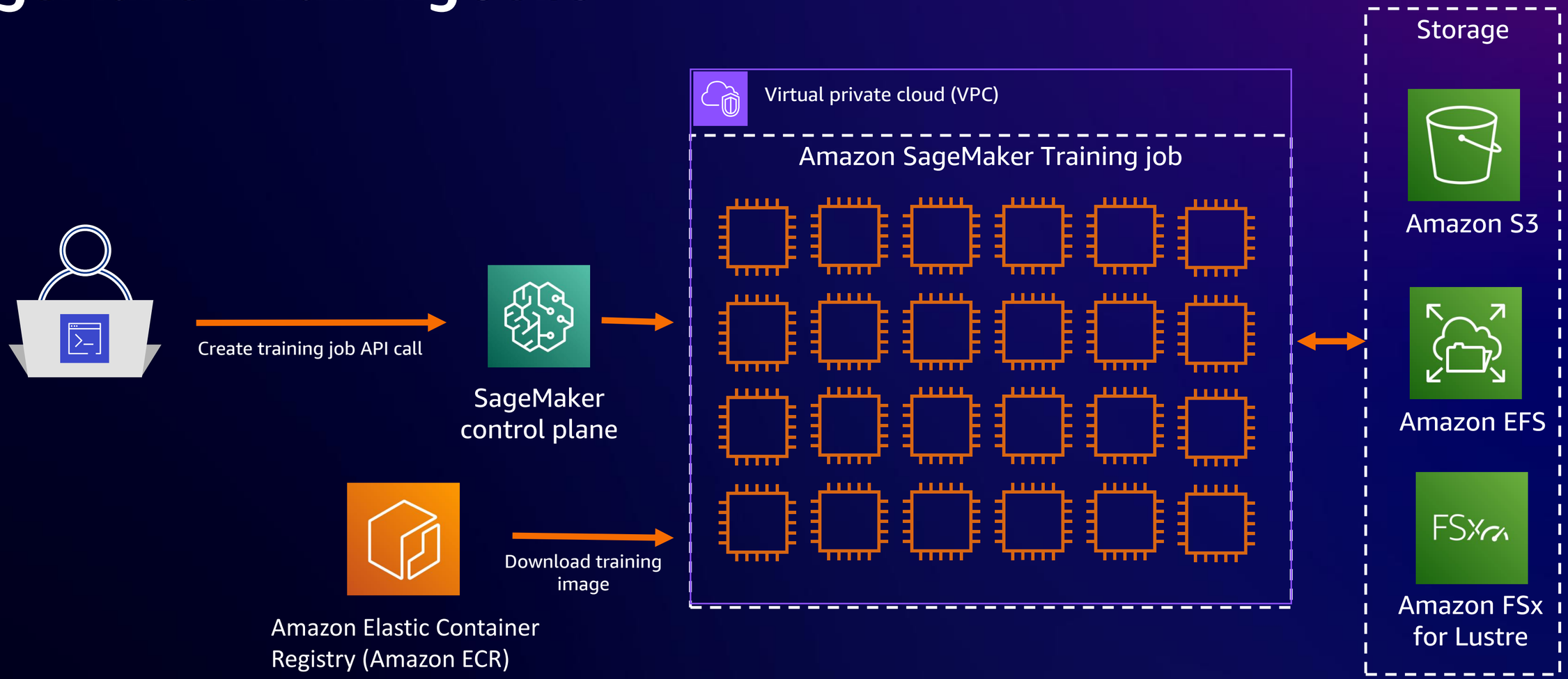
SAGEMAKER HYPERPOD FLEXIBLE TRAINING PLANS

Save weeks of training time and help meet timelines and budgets

ML COMPUTE INSTANCES & ACCELERATORS

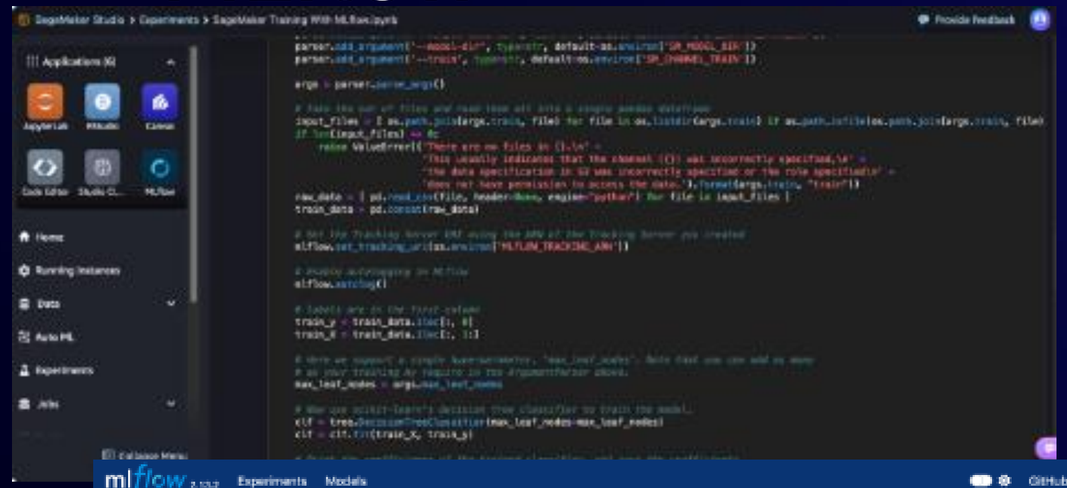
NVIDIA GPUS H100, A100, A10	AWS Trainium Trn1 and Trn1n	CPU instances	AWS Nitro System	EFA based networking

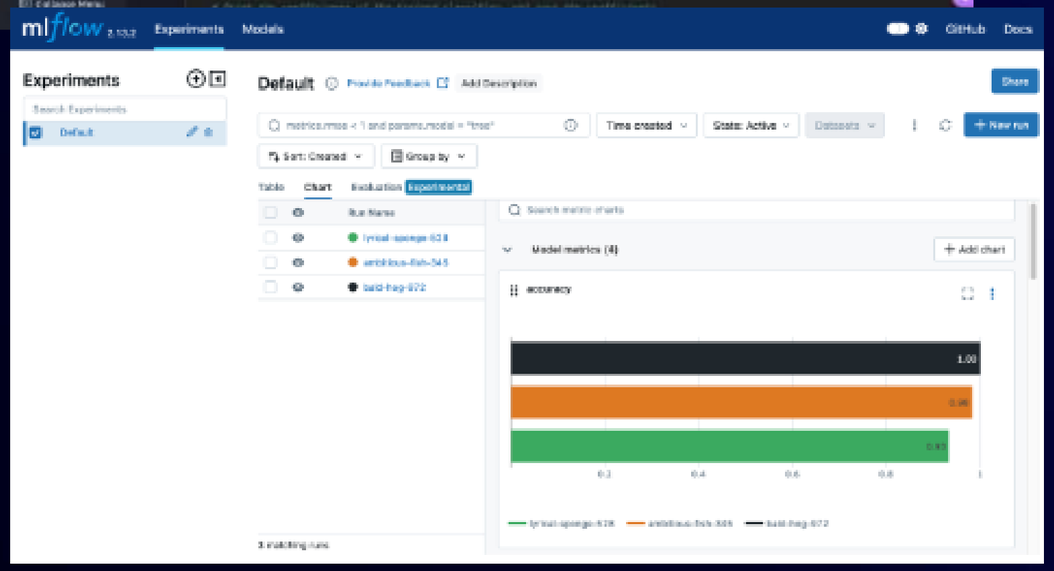
2 SageMaker Training Jobs



- **Managed training environment**
- **Resilient environment**
- **Streamline distributed training**

3 ML Experiment using Amazon SageMaker with MLflow



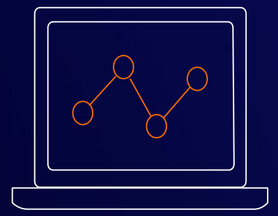


Experiment With ML models and FMs

Use MLflow to manage metrics and experiments

Zero Infrastructure Maintenance

Saves time and cost for setting up Data Science environments



Accelerate experimentation to production

Deploy models registered in MLflow to SageMaker without repackaging model artifacts

Embrace open source community

Benefit from open source innovation with infrastructure management provided by AWS

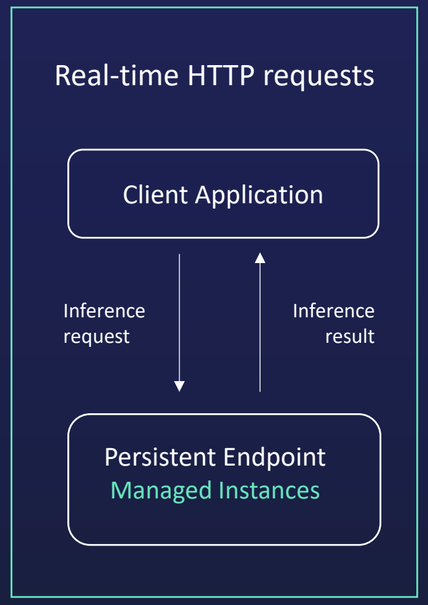
Section 6: LLM Deployment options on AWS

LLM Deployment: Solutions

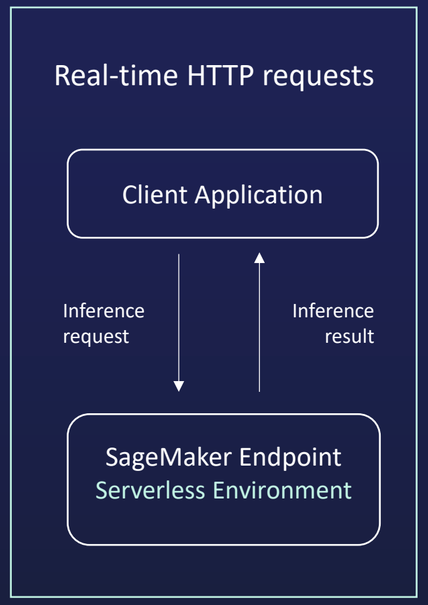
 Fully-managed	Specialized ML deployment options (Real-Time, Serverless, Batch, Asynchronous) with built-in infrastructure management, and scale down to zero option	Amazon SageMaker Inference
 Managed	Managed containerized ML deployment, with a full control of infrastructure, services, and ingress controllers	Amazon SageMaker Hyperpod
 Serverless	Fully serverless deployment with no infrastructure management. Import custom-trained or fine-tuned models for supported architectures for on-demand inference	Amazon Bedrock Custom Model Import

Amazon SageMaker Deployment Modes

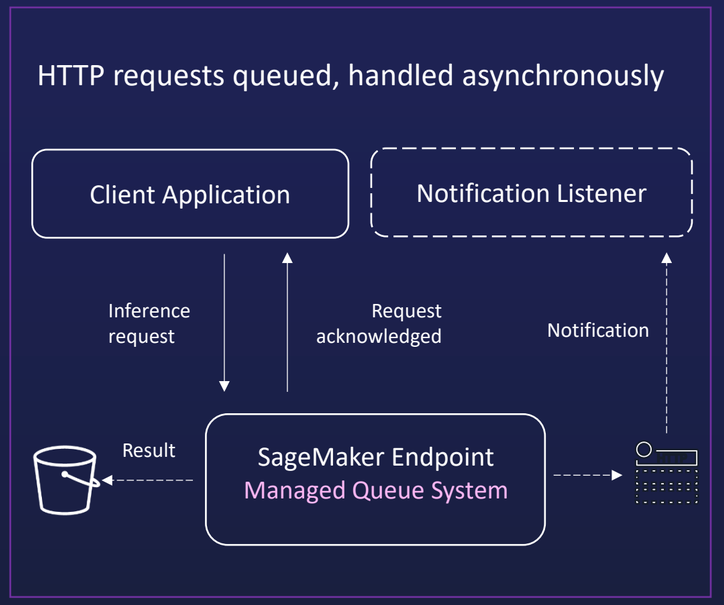
Real-time Inference



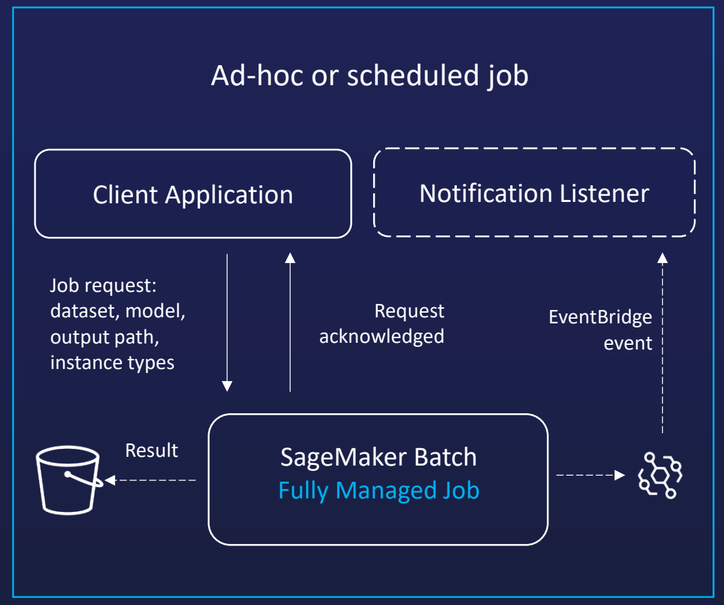
Serverless Inference



Asynchronous Inference



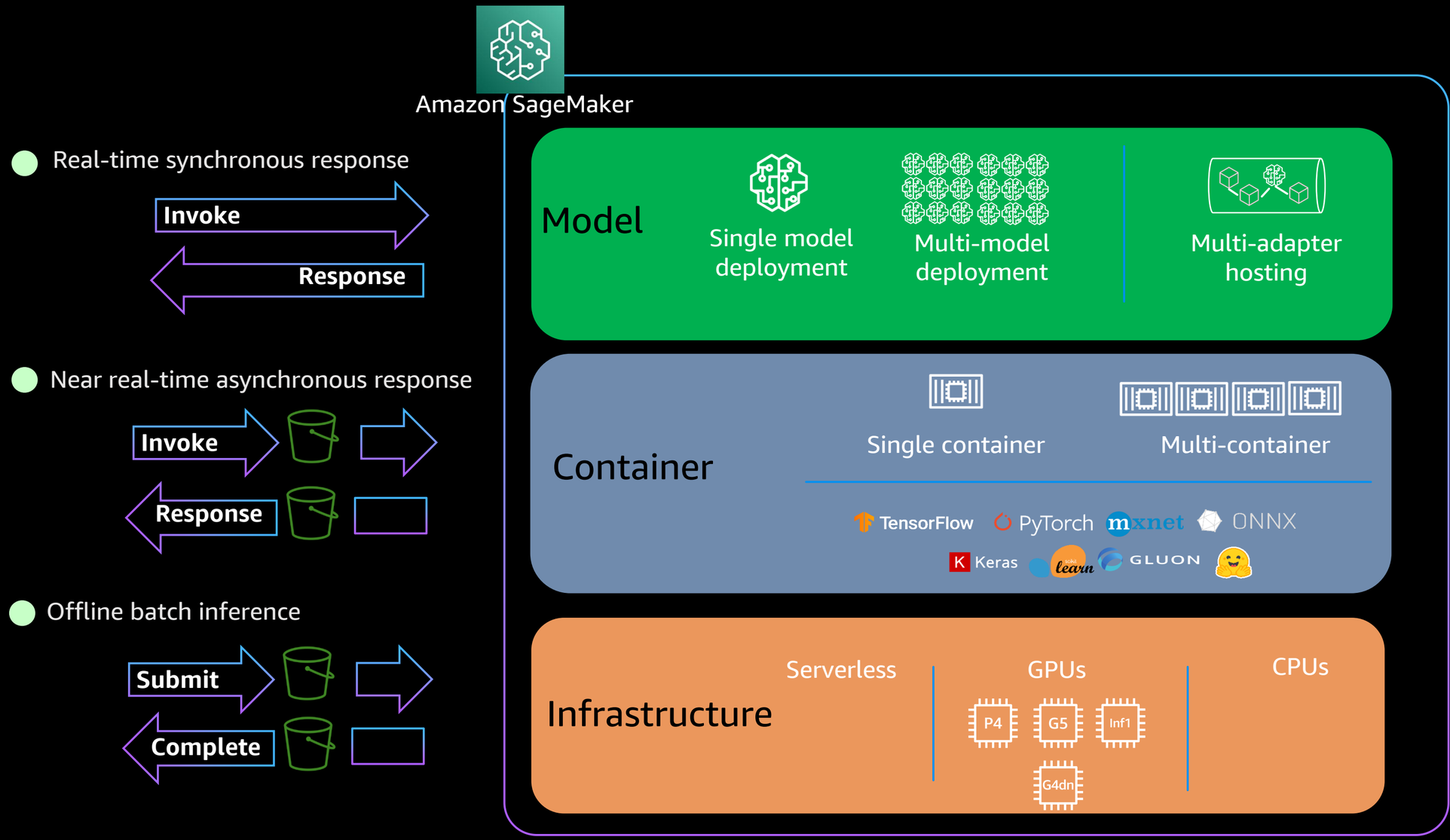
Batch Transform



Example use cases and technical considerations

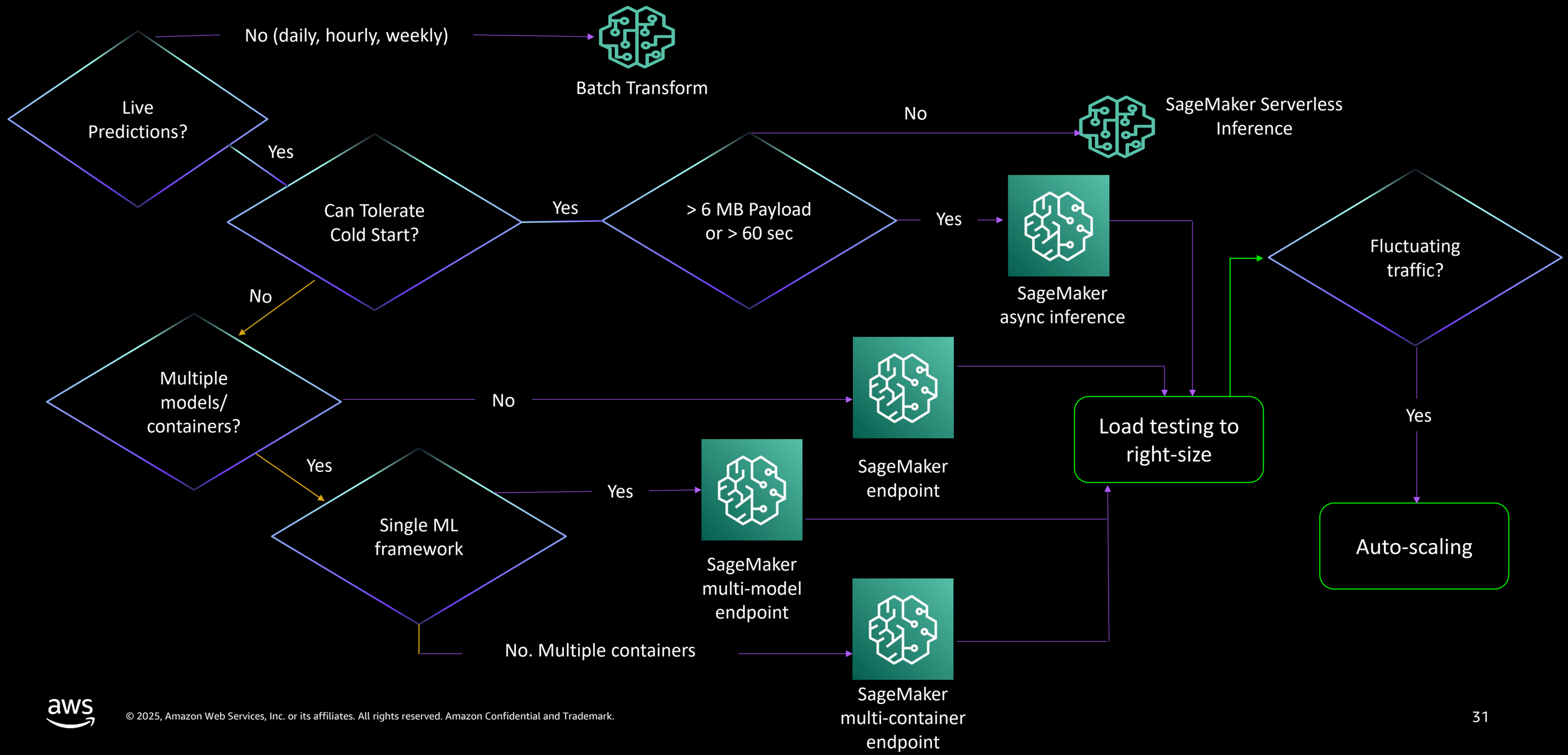
Ad serving, search, personalized recommendations, Generative AI	Extract data from documents, form processing, chatbots, model dev/test	Video processing, large image processing, decoupled applications and systems	Business forecasting, propensity modeling, churn prediction, predictive maintenance
Low latency, high throughput Supports multi-model endpoints	Automatically scale to accommodate unpredictable traffic (scales to zero)	Ability to scale resources to zero Responses can be near real-time	Suitable for periodic arrival of large datasets
Responses within milliseconds	Response times vary (warm/cold start)	Processing times vary: queue size, worker status	Jobs can be long running
Max request payload: 6 MB Timeout: 60 sec	Max request payload: 4 MB Timeout: 60 sec	Inference input: Pointers to S3 objects (1 GB max) Ideal for models with long processing times (15min max)	Ideal for large datasets (Batch Transform allows for splitting of datasets across multiple instances)

Model deployment on Amazon SageMaker



How to choose your Deployment Strategy

A decision tree



Section 7: Demo


Thank You