## aws re: Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

aws

# Optimize cost performance using Amazon SageMaker inference

### **Rishabh Ray Chaudhury**

(he/him) Senior Product Manager, Al AWS

### Raghu Ramesha

(he/him) Senior GenAI/ML Specialist SA AWS

# Agenda

- Customer challenges with gen AI models for inference
- Amazon SageMaker inference optimization toolkit
- Performance improvements
- Inference optimization techniques on Amazon SageMaker
- How does it work?
- Demo

# Customer challenges with gen AI models for inference



# Amazon SageMaker inference optimization toolkit

### 1. Fully managed toolkit

Out-of-the-box support for latest optimization techniques - speculative decoding, compilation, and quantization

### 2. Reduced time-to-market

Optimize gen AI models for best priceperformance in hours, compared to months earlier

# **3. Increase throughput by up to 2x and reduce costs by up to 50%**

Improve cost/performance for a variety of gen AI models including Llama 3, Mistral, and Mixtral models

aws

Provide a state of the	Northerese Control of
	Intel Freenate Cest as //ine-taring Cestimated Meta Meta Net Consisten Agath 2.6
Control	Ram [Pines-taming] Opstimized Mercs Mercs Perc Devalution Aparts 2.0
<ul> <li>i pink</li> <li>i pin</li></ul>	Ram [Pines-taming] Opstimized Mercs Mercs Perc Devalution Aparts 2.0
<ul> <li>According the set of the</li></ul>	Optimized Nets Nets Text Detection Apache 2 II
Abc: Abc:	Optimized Nets Nets Text Detection Apache 2 II
<ul> <li>A control of the standard of the standard products the standard standa</li></ul>	Optimized Nets Nets Text Detection Apache 2 II
A series of the series of t	Merco Merco Text Demonstram Agactor 2 G
<ul> <li>A construction</li> &lt;</ul>	Mess Text Generation Apache 215
<ul> <li>Construction</li> <li>Cons</li></ul>	
<ul> <li>Constant</li> <li>Constant</li></ul>	
Area Description   Area <	
Area Description   Area <	
Audie	
<ul> <li>we were were were were were were were w</li></ul>	
Andie     weener       Control     Control       Andie     Control	
Application       Application         Main       Applica	
<ul> <li>All and a start of a sta</li></ul>	
number       The physics of physics	
number       The physics of physics	
Performant     Performant </td <td></td>	
Performant     Performant </td <td></td>	
Action Normal Series In and series a low area (2003 and hybrid). Action Normal Series In Ac	
Performance       ●         Performance       ● <t< td=""><td></td></t<>	
Constraints 2 where a weak 2 argument are main threater threat where the threat and the thr	
Started ar Starts I was I harrow a quintization A particular Starts I was I harrow a quintization gabs A particular Starts I was I harrow a quintization gabs Constant I was I was I harrow a quintization gabs Constant I was I was I harrow a quintization gabs Constant I harrow a quintization gab	
A create inference optimization job Create inference optimization inference optinter optimization infer	
Create Interence optimization glob Create Interence optimization Create Interence Create Interence optimization Create Interence optimization Create Interence optimization Create Interence Create Interence optimization Create Interence optimization Create Interence optimization Create Interence Create Interence optimization Create Interence Creat	wledge hub 🛛 🤹 Activity
Status D.c.         Convert           P Outlob D.c.         End User License Agreement (ECLA)           P Outlob D.c.         By Global Scheme Agreement (ECLA)	
hy clicking factors by an appect ledg you are of this off-bare is highly to the station for an one clicking appropriate BAD of prove the clicking factors by the states the area with a clicking approximation by a clicking factors by an all obstates transmission of the states that areas the states the areas the states the states that areas the states and conditions. ▲ Insure	
COLON PRIOR     I accept the EULA and read the terms and conditions.	se of AWS
VarX moves     I accept the EULA and read the terms and conditions.	also agree and provider, as
n Home	
© Running Instances	
Cata     Inference optimization job	
2 AutoHL. Job name' Inference-optimization-job-001 ・	
a Experiments	
Jobs     Optimization configurations	
Inference optimization	
Training Optimization configurations	
Model evaluation Instance type: 0	
t Pipelines	
V Models Quantization 0	
Ø JunpSart	
Speculative decoding 0	
Cutput 0	
Output 55 URI	
0, s3://bucket-at/archive/Wename.sip X	
Advanced options	
Collapse menu 🔯	Brows 53

#### Inference optimization 0

#### Inference optimization jobs

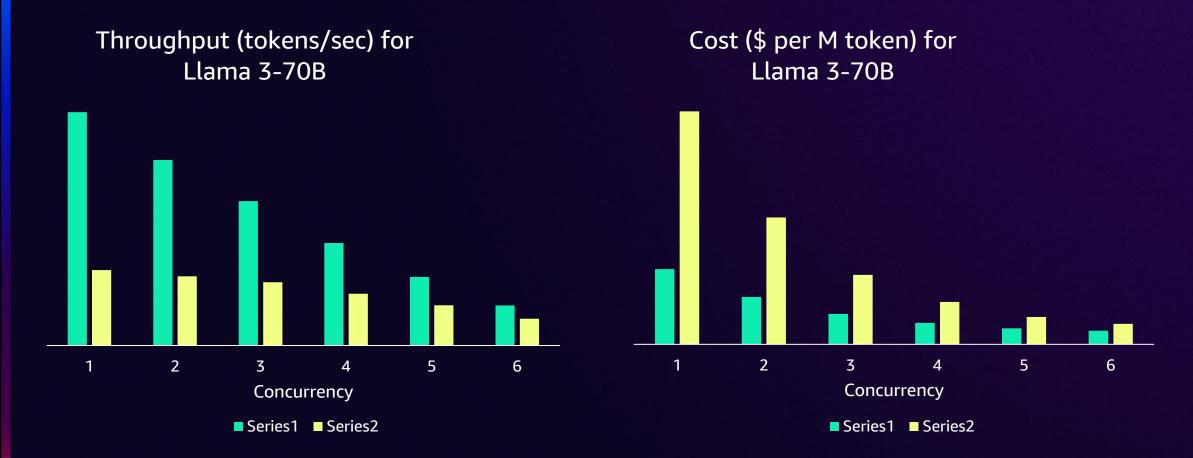
	Job name 🕹	Optimized model	Created on 4	Optimization configurations	Status ↓
0	Inference-optimization-job-001	llama-3-70b-184929	4/9/2024, 5:19:16 PM	Compilation	Completed
0	Inference-optimization-job-002	llama-3-70b-184912	4/9/2024, 5:19:16 PM	Quantization	🕑 Completed
0	Inference-optimization-job-003	llama-3-70b-184916	4/9/2024, 5:19:16 PM	Speculative decoding	🔀 Failed
0	Inference-optimization-job-004	llama-3-70b-184914	4/9/2024, 5:19:16 PM	Quantization	Stopped
0	Inference-optimization-job-005	llama-3-70b-184919	4/9/2024, 5:19:16 PM	Compilation	C Starting
0	Inference-optimization-job-006	llama-3-70b-184911	4/1/2024, 5:09:16 PM	Speculative decoding	C Stopping

#### Learn about inference optimization jobs

	Get started	Documentation	What's new	
	Deploy models for inference	Real-time inference	<ul> <li>Recent features launches</li> </ul>	
<ul> <li>SageMaker Inference explained: Which style i right for you?</li> </ul>	SageMaker Inference explained: Which style is	SageMaker pricing	<ul> <li>Engineering blog</li> </ul>	
	right for you?	Deploy a Machine Learning Model to a Real-Time Inference Endpoint		

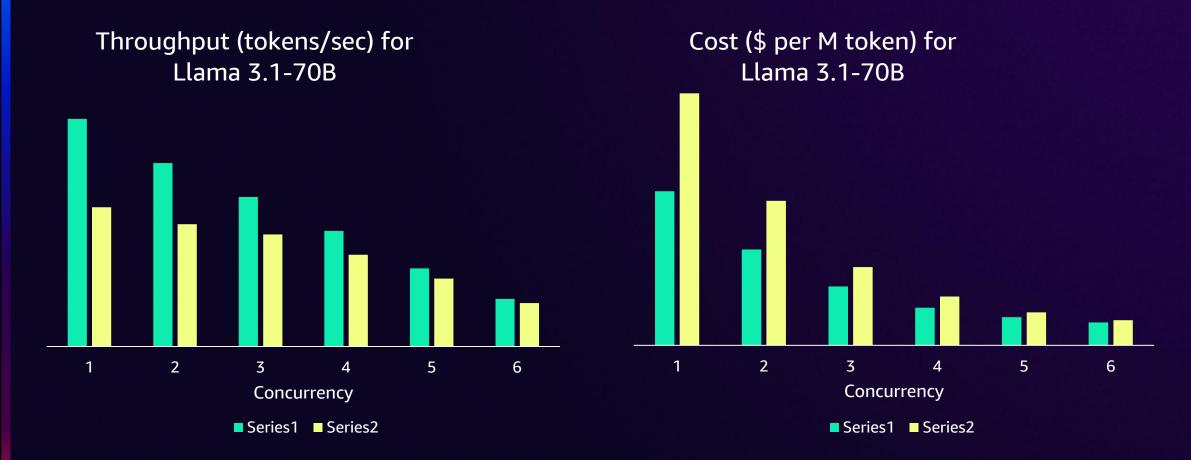
Endpoint settings			
Endpoint name*			
llama 3-706-184929-10-17-2024-6-15			
Deployment instance type" 🚯		Initial Instance count	
mi,mi,p4d 24starge (Default)			
Deployment instance type cannot be changed from the recon Optimized model setting. Changing the Deployment instance a loss of optimization value.	mended type will result in		
Inference type 🕕			
Real-time			
Models			
Piodels			
v 🖂 llama-3-706-184929 🖌 Inference Optimized			
Optimization			
Inference optimization job name Job 001		nization configurations	
300-001	1.00	ntization Speculative decoding	
Model settings			
Number of accelerators'	Mini	num number of copies' 🚯	
Minimum CPU memory'		num CPU memory	
Environment variables			
	Vale	51	
ENDPOINT_SERVER_TIMEOUT	360		
Timeout settings			
Model data download timeout (in seconds)	Com	tainer startup health check timeout (in seconds)	
3600	360		
Tags			
+ Add new tag			
+ Add new tag			
> Advanced settings			

# Up to 2x improvement in throughput and up to 50% lower costs for Llama 3 models



Internal benchmarks on ml.p4d.24xl using OpenOrca dataset

# Significant speed-ups with the latest Llama 3.1 models, as well



Internal benchmarks on ml.p4d.24xl using OpenOrca dataset

# Scaling improvements

### **Container caching**:

- Cached Amazon SageMaker deep learning containers for scaling speed-up
- Supported frameworks Large Model Inference (LMI), HuggingFace TGI, Nvidia Triton, PyTorch
- Reduce scaling latencies by up to **56%** for Llama3.1 70B

### Fast model loading:

- Stream model weights directly in GPU memory
- Reduce scaling latency by up to 20% for Llama3.1 70B

# Scale down to zero

Scale your inference endpoints to zero capacity when there is no traffic

• Endpoint waits 15 minutes before releasing the capacity

### Scale back up with scaling improvements

- Scaling up begins in 1 minute
- Models like Llama3.1 70B scale up and are ready to serve traffic in 6 minutes

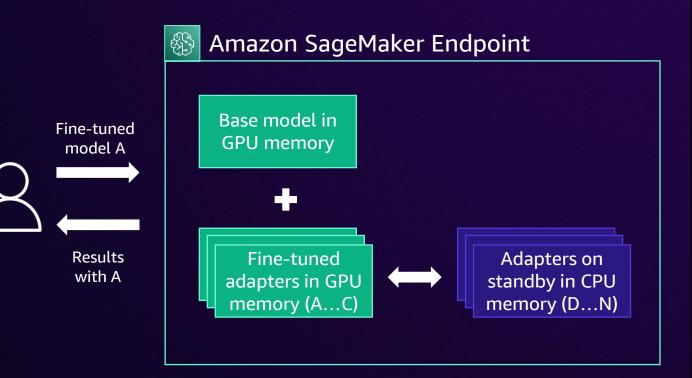
LLaMa3.1 70B					
Time to trigger step scaling policy (min)	Time to provision instance(s) (min)	Time to instantiate a new model copy (min)	Total time (min)		
1	3.018	1.984	6.002		



# Host hundreds of fine-tuned adapters

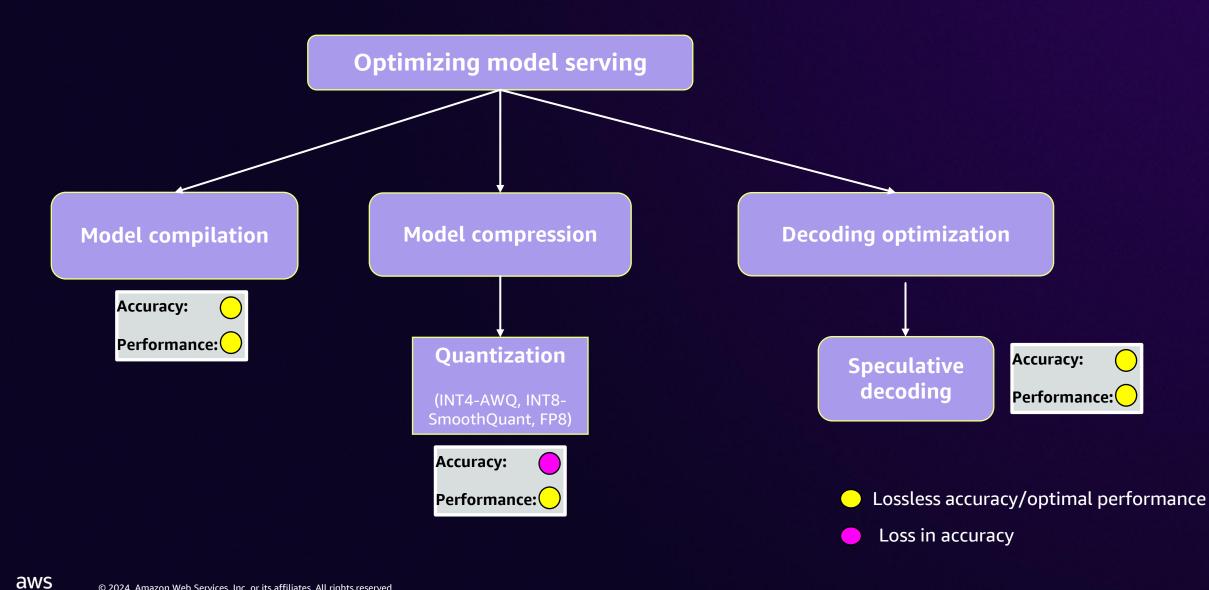
### Save costs

- Host multiple base models with their own fine-tuned adapters on the same endpoint and instances
- Ease of use
  - Manage adapters on the endpoint with lifecycle APIs
  - Monitor usage of each fine-tuned adapter
  - Auto scale base model up and down in response to traffic
- Low overhead latency
  - <1ms overhead to use adapters in GPU memory
  - <10ms overhead to load adapters from CPU to GPU memory for inference

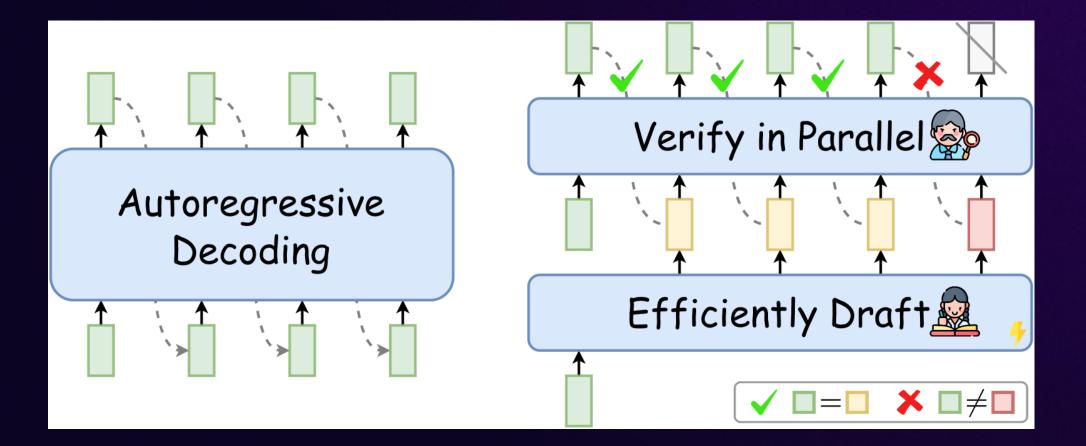




## Inference optimization techniques on Amazon SageMaker

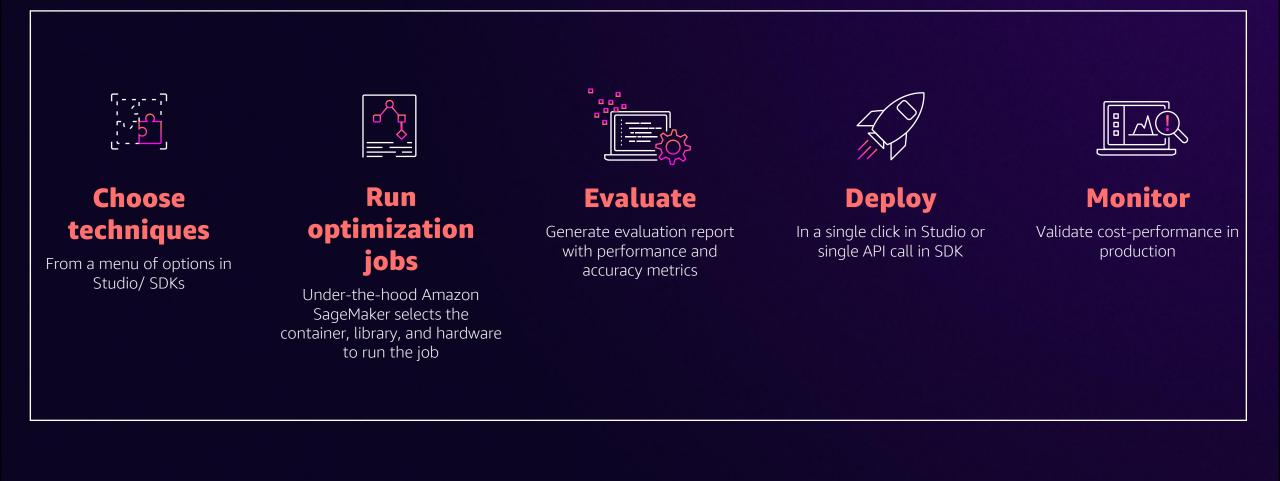


# Speculative decoding



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# How does it work?



# What are customers saying?

Large language models require expensive GPU based instances for hosting, so achieving a substantial cost reduction is immensely valuable. With the new inference optimization toolkit from Amazon SageMaker, based on our experimentation, we expect to reduce deployment costs of our self-hosted large language models by roughly 30% and to reduce latency by up to 25% for up to 8 concurrent request

### **FNU** Imran

aws

Machine Learning Engineer, Qualtrics

# What are customers saying?

The Scale to Zero feature for SageMaker Endpoints will be fundamental for iFood's Machine Learning Operations. Over the years, we've collaborated closely with the SageMaker team to enhance our inference capabilities. This feature represents a significant advancement, as it allows us to improve cost efficiency without compromising the performance and quality of our ML services, given that inference constitutes a substantial part of our infrastructure expenses.

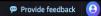
### Daniel Vieira

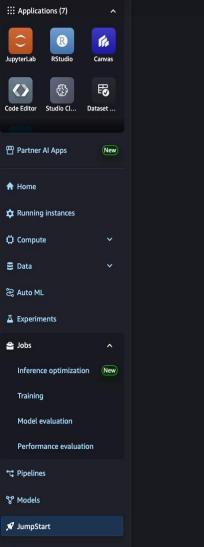
aws

MLOps Engineer Manager at iFoods

# Demo







#### What's New!

• Bedrock ready models You can now use JumpStart to deploy certain models directly into Bedrock and use its feature toolset like Playgrounds and Guardrails. Learn more 🖄

Introducing speculative decoding for Llama-3.1 models.

• Introducing Fast Model Loading for large models.

### All public models

Discover all popular pre-trained models offered by SageMaker

#### Providers 6

Q Search providers or models...

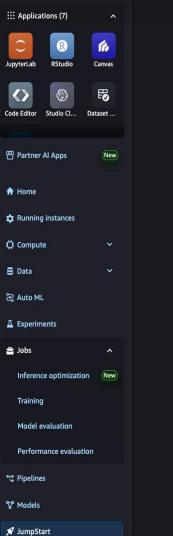
•	💿 🚫 Meta	• stability ai	•
HuggingFace Bedrock ready Explore hundreds of popular and trending	Meta Explore popular and trending models from	Stability Al Explore popular and trending models from	<b>TensorFlow</b> Explore popular and trending models from
models from HuggingFace.	Meta including Llama, Code Llama, and more.	Stability.ai including Stable Diffusion and more.	TensorFlow for computer vision and NLP tasks.
View 406 models >	View 66 models >	View 10 models >	View 346 models >
• PyTorch	• aws		
PyTorch	Amazon		
Explore popular and trending models from PyTorch for computer vision and NLP tasks.	Explore popular and trending models from AWS for computer vision, NLP, and tabular tasks.		
View 34 models >	View 39 models >		

📢 Deployments

Collapse Menu

P





< Deployments

Endpoints Manage deployed models

Projects Automate model building & deployment

#### He What's New!

• Bedrock ready models You can now use JumpStart to deploy certain models directly into Bedrock and use its feature toolset like Playgrounds and Guardrails. Learn more 💈

Introducing speculative decoding for Llama-3.1 models.

• Introducing Fast Model Loading for large models.

### All public models

Discover all popular pre-trained models offered by SageMaker

#### Providers 6

Q Search providers or models...

•	💿 🔊 Meta	• stability ai	•
HuggingFace Bedrock ready	Meta	Stability Al	TensorFlow
Explore hundreds of popular and trending models from HuggingFace.	Explore popular and trending models from Meta including Llama, Code Llama, and more.	Explore popular and trending models from Stability.ai including Stable Diffusion and more.	Explore popular and trending models from TensorFlow for computer vision and NLP tasks.
View 406 models >	View 66 models >	View 10 models >	View 346 models >
• PyTorch	aws		
PyTorch	Amazon		
Explore popular and trending models from PyTorch for computer vision and NLP tasks.	Explore popular and trending models from AWS for computer vision, NLP, and tabular tasks.		
View 34 models >	View 39 models >		

📶 Amazon SageMaker launches 🗙 🌕 🛞 SageMakerPublicHub - Sage 🗙 🔶 🕂 studio-d-9ltxbezjspmx.devo.studio.us-west-2.asfiovnxocqpcry.com/jumpstart/SageMakerPublicHub Ð New Chrome available C ŝ С 88 🗀 Useful 🗀 Personal 🗀 Sagemaker All Bookmarks 8 SageMaker Studio > Jumpstart > SageMakerPublicHub 🗭 Provide feedback 🏚 Running instances **What's New!** • Bedrock ready models You can now use JumpStart to deploy certain models directly into Bedrock and use its feature toolset like Playgrounds and Guardrails. Learn more 🖸 Compute ~ Introducing speculative decoding for Llama-3.1 models. • Introducing Fast Model Loading for large models. 🛢 Data  $\sim$ All public models auto ML Discover all popular pre-trained models offered by SageMaker Experiments 🚔 Jobs ~ **Providers** 6 Inference optimization (New) Q Search providers or models... Training  $(\mathbf{0})$ 0 ( 0 0 Model evaluation stability.ai 🔿 Meta Performance evaluation T Pipelines HuggingFace (Bedrock ready) Meta Stability AI **TensorFlow** Explore hundreds of popular and trending Explore popular and trending models from Explore popular and trending models from Explore popular and trending models from St Models models from HuggingFace. Meta including Llama, Code Llama, and Stability.ai including Stable Diffusion and TensorFlow for computer vision and NLP more. more. tasks. 🝠 JumpStart View 406 models > View 66 models > View 10 models > View 346 models > Contraction Contractica Contra ~ This provider contains open source models Endpoints Q Manage deployed models PyTorch aws Projects Automate model building & PyTorch Amazon Explore popular and trending models from Explore popular and trending models from Collapse Menu PyTorch for computer vision and NLP tasks. AWS for computer vision, NLP, and tabular

tasks.

~

ę

SageMaker Studio > Jumpstart > SageMakerPublicHub



🔒 Home

🏚 Running instances

🛱 Compute

🛢 Data

දි Auto ML

A Experiments

🚔 Jobs

Inference optimization New

Training

Model evaluation

Performance evaluation

⁼t; Pipelines

😵 Models

🝠 JumpStart

< Deployments

Endpoints Manage deployed models

Projects Automate model building & deployment

#### What's New!

• Bedrock ready models You can now use JumpStart to deploy certain models directly into Bedrock and use its feature toolset like Playgrounds and Guardrails. Learn more 🗹

- Introducing speculative decoding for Llama-3.1 models.
- Introducing Fast Model Loading for large models.

### All public models

Discover all popular pre-trained models offered by SageMaker

#### Providers 6

Q Search providers or models...

•	💿 🔿 Meta	• stability ai	•
HuggingFace (Bedrock ready)	Meta	Stability AI	TensorFlow
Explore hundreds of popular and trending models from HuggingFace.	Explore popular and trending models from Meta including Llama, Code Llama, and more.	Explore popular and trending models from Stability.ai including Stable Diffusion and more.	Explore popular and trending models from TensorFlow for computer vision and NLP tasks.
View 406 models >	View 66 models, >	View 10 models >	View 346 models >
• PyTorch	• aws		
PyTorch	Amazon		
Explore popular and trending models from PyTorch for computer vision and NLP tasks.	Explore popular and trending models from AWS for computer vision, NLP, and tabular tasks.		
View 34 models >	View 39 models >		

🗭 Provide feedback 🛛 😣

P

### How to optimize the Meta Llama-3.170B Amazon JumpStart model for inference using Amazon SageMaker model optimization jobs

Recommended kernel(s): This notebook can be run with any Amazon SageMaker Studio kernel.

In this notebook, you will learn how to apply state-of-the-art optimization techniques to an Amazon JumpStart model (JumpStart model ID: meta-textgeneration-llama-3-1-70b-instruct) using Amazon SageMaker ahead-of-time (AOT) model optimization capabilities. Each example includes the deployment of the optimized model to an Amazon SageMaker endpoint. In all cases, the inference image will be the SageMaker-managed LMI (Large Model Inference) Docker image. LMI images features a DJL serving stack powered by the Deep Java Library.

You will successively:

- 1. Deploy a pre-optimized variant of the Amazon JumpStart model with speculative decoding enabled (using SageMaker provided draft model). For popular models, the JumpStart team indeed selects and applies the best optimization configurations for you.
- 2. Customize the speculative decoding with open-source draft model.
- 3. Quantize the model weights using the FP8 algorithm.
- 4. Compile the model for a deployment TensorRT-LLM framework.

#### Notices:

- Make sure that the ml.p4d.24xlarge and ml.g5.12xlarge instance types required for this tutorial are available in your AWS Region.
- Make sure that the value of your "ml.p4d.24xlarge for endpoint usage" and "ml.g5.12xlarge for endpoint usage" Amazon SageMaker service quotas allow you to deploy at least one Amazon SageMaker endpoint using these instance types.

This notebook leverages the Model Builder Class within the sagemaker Python SDK to abstract out container and model server management/tuning. Via the Model Builder Class you can easily interact with JumpStart Models, HuggingFace Hub Models, and also custom models via pointing towards an S3 path with your Model Data. For this sample we will focus on the JumpStart Optimization path.

#### License agreement

- This model is under the Meta license, please refer to the original model card.
- This notebook is a sample notebook and not intended for production use.

#### Execution environment setup

This notebook requires the following third-party Python dependencies:

AWS sagemaker with a version greater than or equal to 2.232.2

Let's install or upgrade these dependencies using the following command:

[ ]: %pip install "sagemaker>=2.235.2" ---upgrade ---quiet ---no--warn-conflicts

#### Setup

[]: import sagemaker

from sagemaker.serve.builder.model\_builder import ModelBuilder
from sagemaker.serve.builder.schema\_builder import SchemaBuilder
from sagemaker.jumpstart.model import ModelAccessConfig

## Resources



Inference optimization toolkit

### Scale down to zero

### Faster model loading

Container caching



# Thank you!



Please complete the session survey in the mobile app

Rishabh Ray Chaudhury rishrayc@amazon.com

### Raghu Ramesha ragmesh@amazon.com

