Amazon SageMaker JumpStart

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customer challenges

Getting started with ML takes

too long...

- Importing publicly available algorithms and models into SageMaker
- Maintaining and updating SageMaker-compatible scripts
- Setting up infrastructure
- Building NLP and vision models from scratch
- Model sharing and collaboration is done manually

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker JumpStart

ML hub with foundation models, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks

Machine learning hub

Browse through 400+ built-in algorithms with pretrained foundation models, solutions, and example notebooks

Pre-built training and inference scripts

Compatible with SageMaker and configurable with custom dataset

UI as well as API-based

Use the user interface for single click model deployment or API for the Python SDK-based workflow

Notebooks with examples

Jump into notebooks to use selected model with examples to guide you through the entire ML workflow

Share and collaborate within your organization

Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why use foundation models on SageMaker JumpStart

1

Choose foundation models offered by model providers

Meta AI

AI21 labs

LightOn
We bring Light to AI

stability.ai

co:here

alexa

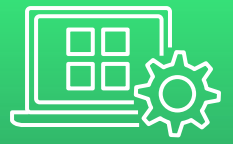
2

Deploy model

Deploy the model for inference using SageMaker hosting options includes single node

3

Fine tune model and automate ML workflow



Only selected models can be fine-tuned

Automate ML workflow

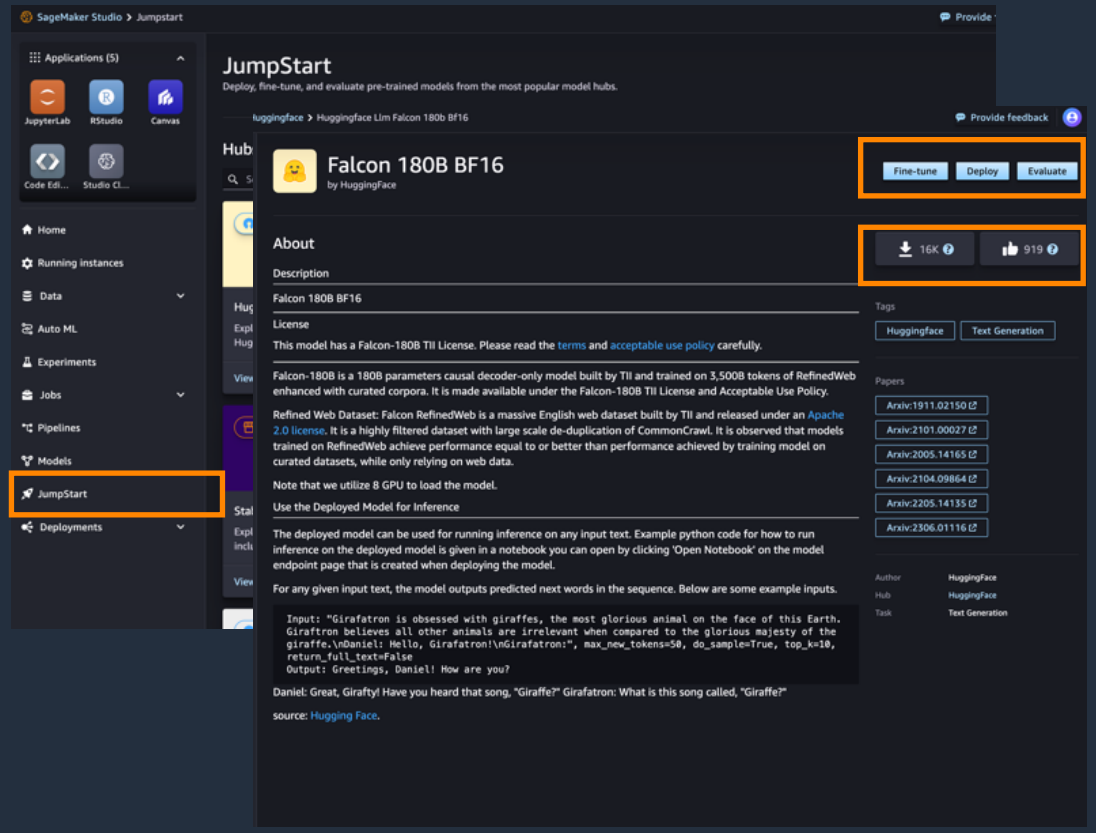
Data stays in your account including model, instances, logs, model inputs, model outputs

Fully integrated with Amazon SageMaker features

aws

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker JumpStart is easiest way to use latest models and features



- Available from Studio V2 New design with filters, sorts, download and like counts
- More deployment configs including environment variable support, timeout config etc.
- Qualitative model evaluation with SageMaker Clarify Deploy multiple models at once during deployment

All new features launching at re:Invent are available from JumpStart model detail and deployment page

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Build with the broadest selection of foundation models

AVAILABLE IN AMAZON SAGEMAKER JUMPSTART

Meta AI

Models
Llama 2 7B, 13B, 70B
Code Llama 7B, 13B, 34B

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

AI21 labs

Models
Jurassic-2 Ultra, Mid

Tasks
Contextual answers
Summarize
Paraphrase
Grammatical error correction

Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction
Question answering
Classification

co:here

Models
Command
Command Light

Tasks
Text generation
Information extraction
Question answering
Summarization

Models
Falcon-7B, 40B, 180B
Mistral 7B
Open LlaMA
RedPajama
MPT-7B, Dolly
BloomZ 176B
Flan T-5 models (8 variants)
DistilGPT2
GPT NeoXT
Bloom models

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

stability.ai

Models
Stable Diffusion XL 1.0
2.1 base
Upscaling
Inpainting

Tasks
Generate photo-realistic images from text input
Improve quality of generated images

Features
Fine-tuning on Stable Diffusion 2.1 base model

LightOn

Models
Lyra-Fr
10B, Mini

Tasks
Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification

alexa

Models
AlexaTM 20B

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key factors in decision making

Cost

Optimize for cost with a variety of models, size, and instance for your needs with AWS pay-as-you go pricing

Accuracy

Use highly accurate models per HELM benchmarks

Speed (latency)

Optimize for performance with different model sizes and instance types

Ease of use

Instantly try models in playground; deploy with SageMaker using managed inference scripts

Data security

Host models on customer-dedicated endpoints inside your VPC

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Comprehensive data protection and privacy

Your data used with SageMaker JumpStart is not used for service improvement and not shared with third party model providers

Private connectivity between Amazon SageMaker and your virtual private cloud (VPC)

Your data is encrypted in transit and at rest

Customize FMs privately, retaining control over how your data is used and encrypted

Deploy FMs on single tenant endpoints provisioned for your inference use

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Algorithms and Model Packages

Easily and securely customize third-party models by fine-tuning them on your private data.

End-to-End Security

Private data and incremental model weights learnt from that data are never exposed

Multiple Training Techniques

Choose from a library of ready-to-use training techniques, including fine-tuning

Easy to use

Use familiar interfaces like the SageMaker Python SDK

Secure Model Inference

Deploy the customized model to a secure endpoint

Variety of models

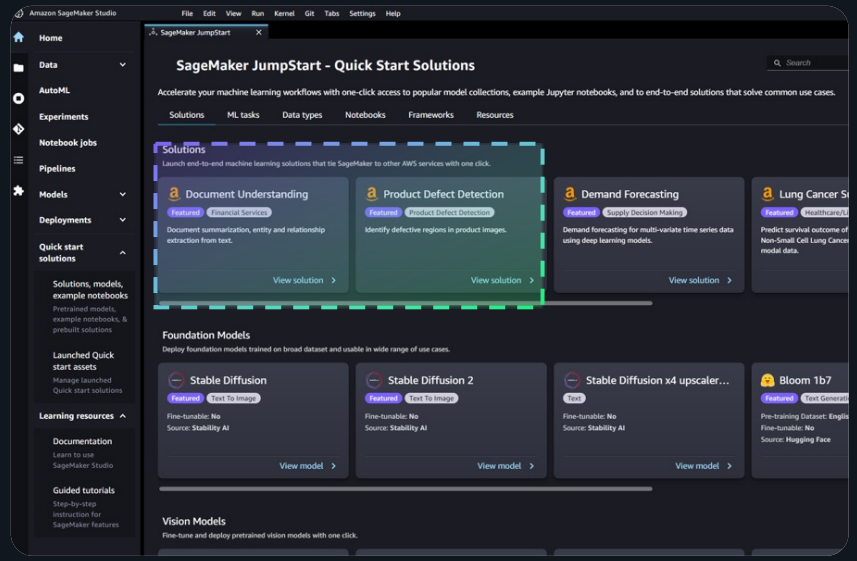
Select from a range of proprietary third-party models.

aws

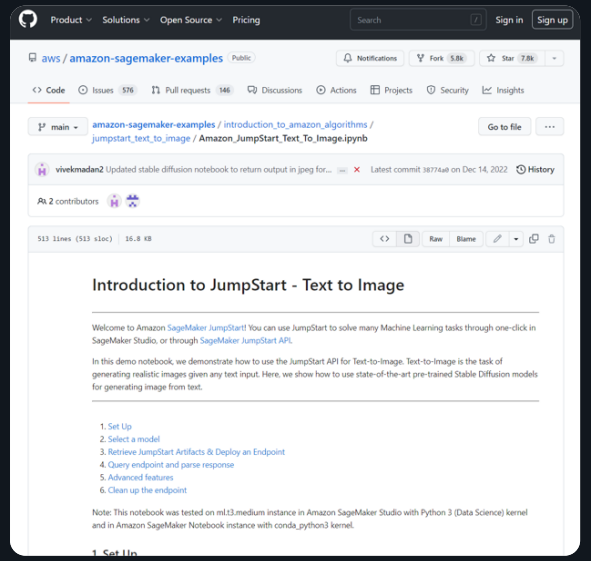
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3 ways to use foundation models with SageMaker JumpStart

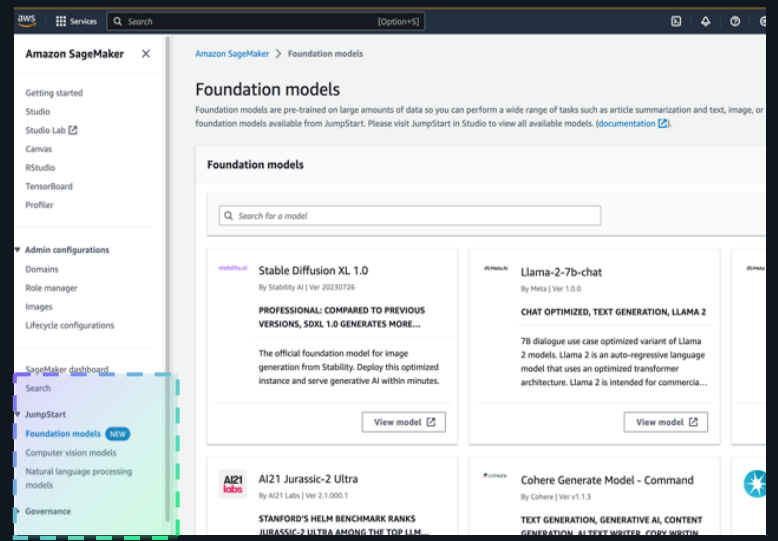
SageMaker Studio
One-click deploy



SageMaker SDK

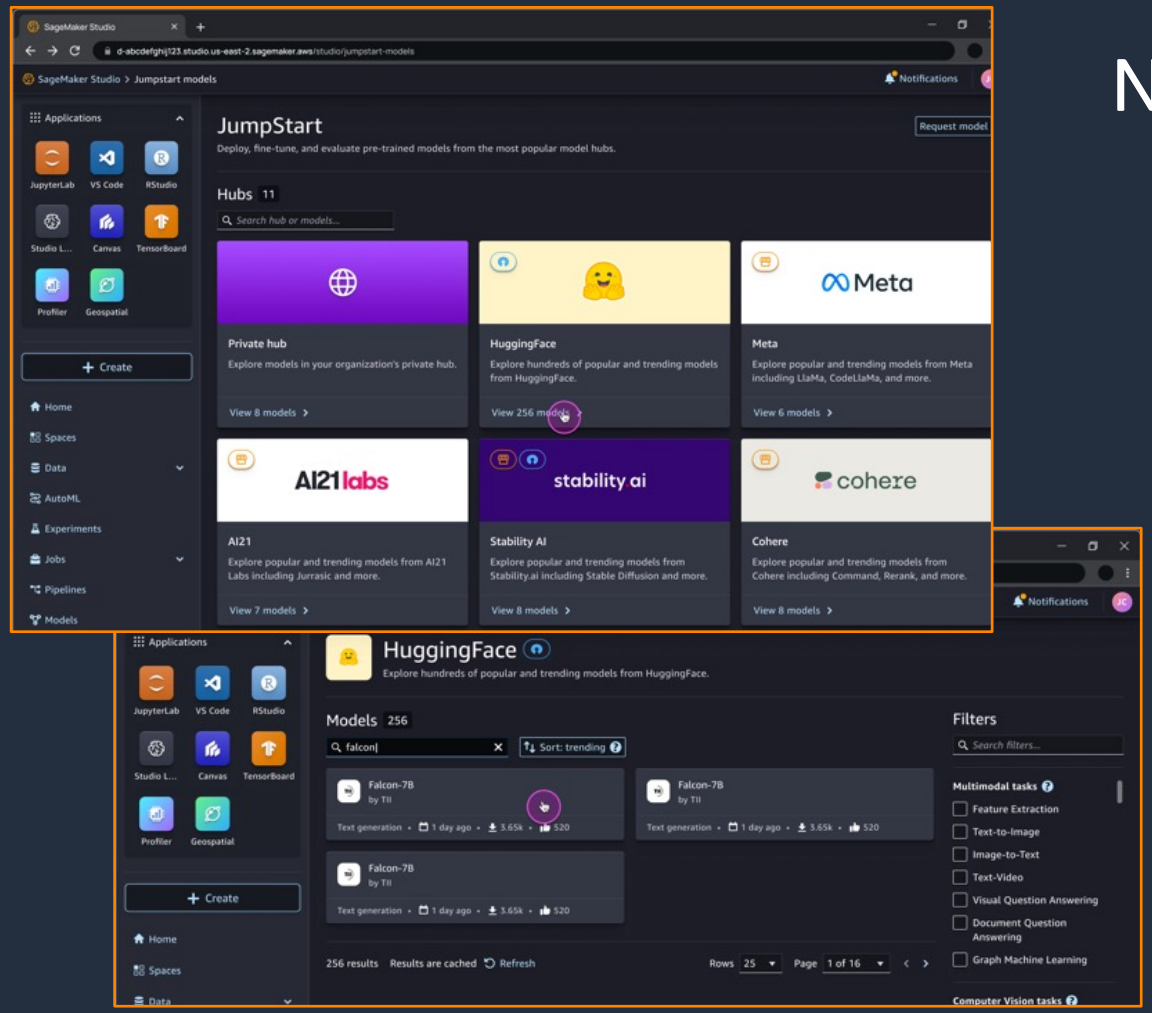


SageMaker console



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

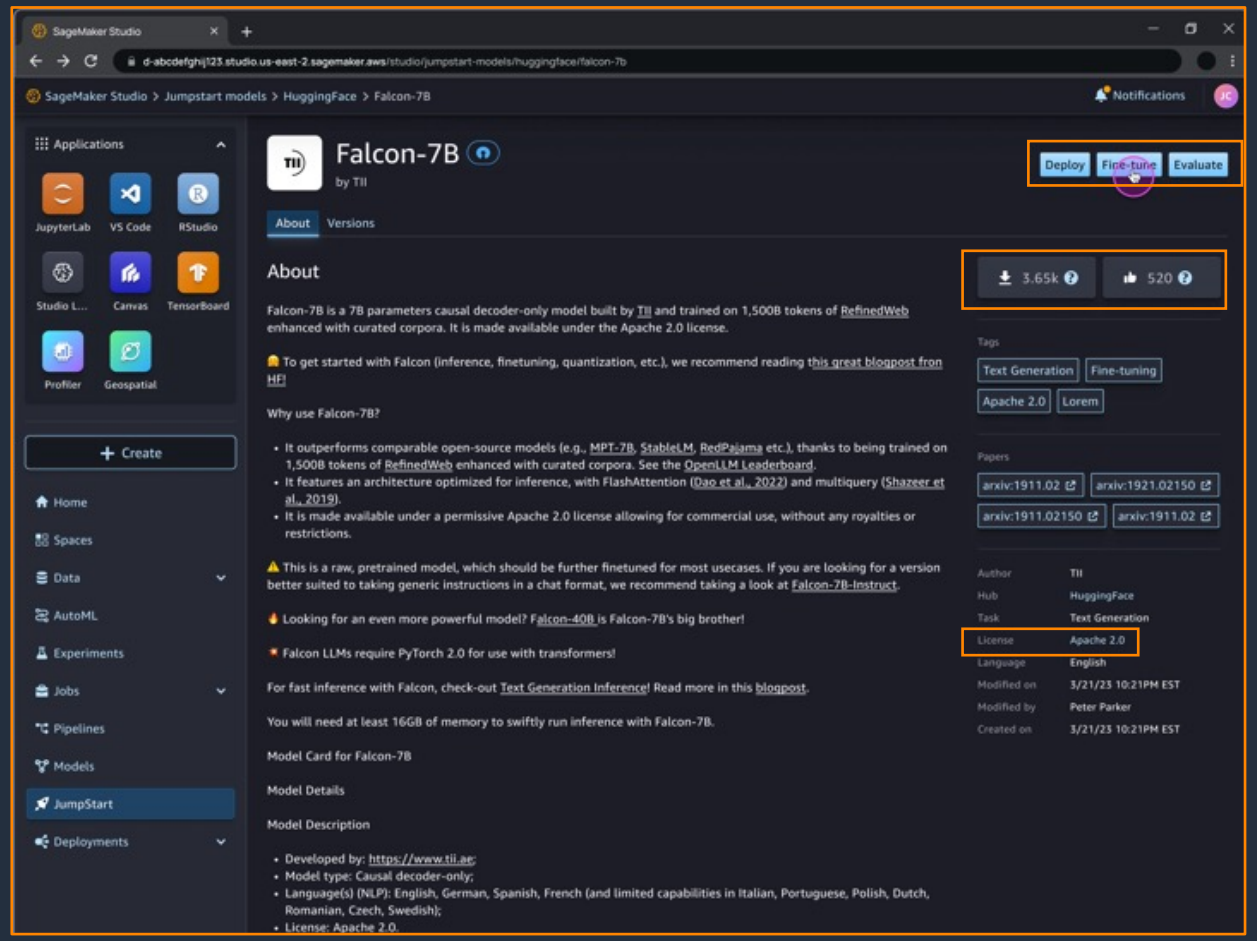
Browsing models in JumpStart with new experience



- New experience in SageMaker Studio
  - Browse models from various providers in single place
  - Filters by use cases, modalities, and license types
  - Sort by popularity (likes and downloads)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deep diving into each model for next steps



- Details from model provider:
  - Model size and description
  - License info
  - Use cases and how-to use model
- Available actions users can take: Deploy, Train, and Evaluate

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Evaluation Capabilities

ACCURACY, ROBUSTNESS, STEREOTYPING, TOXICITY, FACTUAL KNOWLEDGE

UI Evaluation Wizard
&
FM Eval Library

Human-in-the-Loop Evaluations

Task Templates
General, Summarization, Q&A, Classification

Built-In &
BYO Prompt Datasets

Built-In & BYO Scoring Algorithms

LLM Evaluation Reports

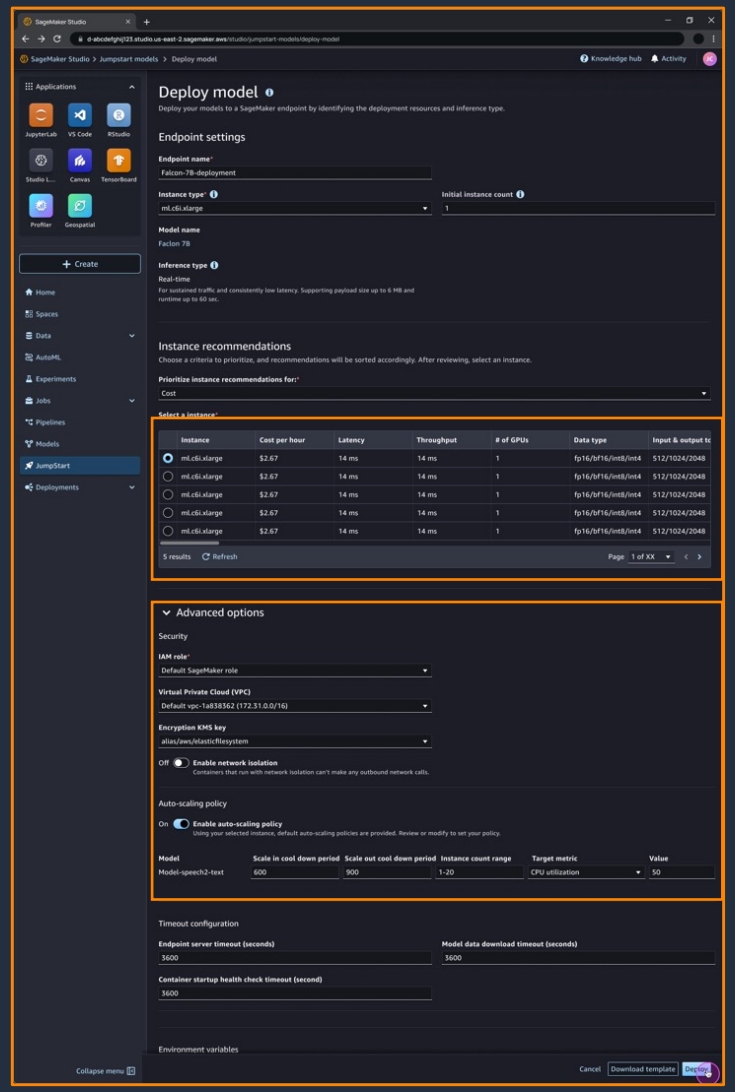
© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Metric, Algorithms and Datasets

Task	Eval Dimension	Algorithm	Dataset
General / Text Generation	Prompt Stereotyping	Is Biased, Log Probability Difference	CrowS-Pairs
	Toxicity	Detoxify, Toxigen (amount of toxic content)	RealToxicityPrompts, BOLD
	Factual Knoweldge	Percentage of correctly retrieved real-world facts	TREX
	Semantic Robustness	Performance change	BOLD, TREX prompts, WikiText, English Wikipedia
Text Summarization	Accuracy	Rouge-N	Government Report Dataset Gigaword., XSUM
		Meteor	
		BERTScore	
	Toxicity	Detoxify, Toxigen	
	Semantic Robustness	Performance change	
Questions & Answering	Accuracy	Exact match	BoolQ, NaturalQuestions, TriviaQA
		Quasi exact match	
		F1-over-words	
	Toxicity	Detoxify, Toxigen	
	Semantic Robustness	Performance change	
Text Classification	Accuracy	Classification accuracy	Women's Ecommerce Clothing Reviews
		Balanced classification accuracy	
		Precision	
		Recall	
	Semantic Robustness	Performance change	

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deploy with SageMaker deploy experience



- Deploy with defaults with no further inputs
- Choose instance that meets your need
- Deploy with special options such as VPC settings
- Auto-scale set up for scalability

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker JumpStart SDK experience

Deploy pre-trained JumpStart Model

```
from sagemaker.serve.builder.model_builder import ModelBuilder

model_builder = ModelBuilder(
      model="huggingface-llm-falcon-7b-bf16",
      schema_builder=SchemaBuilder(sample_input, sample_output),
      role_arn=role
)

sagemaker_model = model_builder.build()

predictor = sagemaker_model.deploy()

prediction = predictor.predict(sample_input)
```

Fine-tuning

```
from sagemaker.jumpstart.estimator import JumpStartEstimator

domain_adaptation_estimator = JumpStartEstimator(
      model_id="huggingface-llm-falcon-7b-bf16",
      hyperparameters=my_hyperparameters,
      instance_type="ml.p4dn.24xlarge",
)

domain_adaptation_estimator.fit(
      {
            "train": training_dataset_s3_path,
            "validation": validation_dataset_s3_path},
            logs=True
)
```

Retrieval Augmented generation (RAG)

```
from langchain.llms.sagemaker_endpoint import SagemakerEndpoint

sm_llm = SagemakerEndpoint(

   endpoint_name="huggingface-llm-falcon-7b-bf1",
   region_name=aws_region,
   model_kwargs=parameters,
   content_handler=content_handler,
)

PROMPT = PromptTemplate(template=prompt_template, input_variables=["context",
"question"])
chain = load_qa_chain(llm=sm_llm, prompt=PROMPT)

result = chain({"input_documents": docs, "question": question},
```

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Choosing the right instance for hosting

Size of model (# of parameters)	3B–10B	11B–20B	100B+*
Task Type	Image generation Simple text classification (Short form)	Natural language understanding (NLU)	Natural language generation (NLG) (long form)
Minimum instance required	p3.2xlarge g5.2xlarge	p3.8xlarge g5.12xlarge	p4de.24xlarge p4d.24xlarge
Pricing	$4/hr $2/hr	$15/hr $9/hr	$47/hr $38/hr

Scale vertically (larger instances) to improve latency

Scale horizontally (more instances) to support higher traffic

*P4d instances will have limited availability, escalate to S-Team for support

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model Showcase

- Llama-2
- Code Llama
- Mistral 7B
- SD XL 1.0

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Llama-2 models are highly performant OSS models that are approved for commercial use

Performant	Llama 2 on the leaderboard on many external benchmarks, including reasoning, coding, proficiency, and knowledge tests. Uses supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety
Flexible	4K context length; available in 7B, 13B, and 70B parameters sizes; fine-tuned for chat use cases
Open	Open weights and available for research and commercial use[1]
Responsible	Fine-tuned models have been red-teamed — tested for safety — through internal and external efforts[2]

1/ License: https://ai.meta.com/llama/license/
2/ Responsible use policy: https://ai.meta.com/llama/use-policy/

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Update to Llama 2 – Nov'23

Ability to customize deployment config: set different environment variables (rolling batch, max concurrency, max window)

Faster deployment time and faster inference

Support for streaming responses

Support for inf2/trn1 hardware in us-east-2, us-west-2, us-east-1

Expanding to new regions where instances are supported

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Code Llama models

Performant	Capable of generating code, and natural language about code, from both code and natural language prompts. State-of-the-art performance in Python, C++, Java, PHP, C#, TypeScript, and Bash.
Flexible	Offers a higher context window of 16k, compared to the Llama 2 context length of 4k
Open	Open weights and available for research and commercial use[1]
Responsible	7B and 13B base and instruct variants support infilling based on surrounding content, making them ideal for code assistant applications

1/ License: https://ai.meta.com/llama/license/
2/ Responsible use policy: https://ai.meta.com/llama/use-policy/

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Mistral 7B models

Performant	Supports English text and code generation abilities. Strong throughput and latency performance when compared to larger model alternatives, with low memory requirement at a 7B model size.
Flexible	8k token context length and easily customizable. Mistral 7B-instruct model for chat use cases, fine-tuned using a variety of publicly available conversation datasets.
Open	Open weights and available for research and commercial use[1]

1/ License: https://www.apache.org/licenses/LICENSE-2.0

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SDXL 1.0 model prompts

- Highlights: Improved photorealism effect, artistic intelligence, simpler prompting and more accurate.

- Prompt guidance: You can change each of the three individually to generate variations of an image:
  - type of image
  - description
  - (realistic/artistic/type of art etc.)

- Negative prompts can be used to direct the model to avoid certain objects, colors, styles, attributes ,etc.

Images generated by StabilityAI SDXL 1.0

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Foundation Models on SageMaker resources

- Jumpstart First Call Deck
- Jumpstart GenAI labs (4 hours)
- Generative AI on Amazon SageMaker Workshop (1 day)
- Github examples from WWSO SA
  (self paced learning)
- Generative AI with LLMs in Coursera (new)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Additional References

RAG and text summarization use-cases along with associated workshop:

- https://github.com/aws-samples/amazon-sagemaker-generativeai/tree/main/usecases
- https://catalog.us-east-1.prod.workshops.aws/workshops/972fd252-36e5-4eed-8608-743e84957f8e/en-US/lab4

Text summarization to image example:

- https://github.com/aws-samples/amazon-sagemaker-generativeai/tree/main/usecases/text_summarization_to_image

Blogs (new)

- Exploring summarization options for Healthcare with Amazon SageMaker
- Use generative AI foundation models in VPC mode with no internet connectivity using Amazon SageMaker JumpStart
- Use Stable Diffusion XL with Amazon SageMaker JumpStart in Amazon SageMaker Studio
- Llama 2 foundation models from Meta are now available in Amazon SageMaker JumpStart
- Safe image generation and diffusion models with Amazon AI content moderation services
- Use proprietary foundation models from Amazon SageMaker JumpStart in Amazon SageMaker Studio

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Built-in algorithms and pretrained models

400+ ALGORITHMS AND PRE-TRAINED, STATE-OF-THE-ART, PUBLICLY AVAILABLE MODELS FROM PYTORCH HUB, TENSORFLOW HUB, AND HUGGING FACE HUB THAT ARE SECURELY STORED IN AN AWS-OWNED ENVIRONMENT

	Tasks		Algorithms/models
Tabular	Classification, regression, time-series		LightGBM, CatBoost, AutoGluon, TabTransformer, XGBoost, DeepAR
Vision	Image classification Image embedding	Object detection Semantic segmentation	ResNet, Inception, MobileNet, SSD, Faster RCNN, YOLO, and more
Text	Sentence classification Text classification Question answering	Summarization Text generation, translation, Named-entity recognition	AlexaTM, Bloom, Stable Diffusion 2.0, BERT, RoBERTa, DistilBERT, Distillbart xsum, GPT2, ELECTRA, and more
Audio	Audio embedding		TRILL, TRILLsson, TRILL-Distilled, FRILL

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

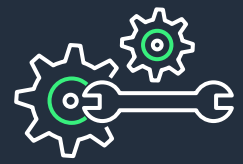
Built in algorithms with pretrained models: how it works

Amazon SageMaker JumpStart

Built-in algorithms with pretrained ML models, including foundation models

Access and browse

Browse through hundreds of built-in algorithms with pretrained models from popular model hubs and foundation models



Select and fine-tune

Select pretrained model or an algorithm for your ML task (vision, text, tabular, audio) and fine-tune with your data

Deploy

Deploy model with a single click or with the API, and use example to further experiment

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Solutions with SageMaker JumpStart

Extract & analyze data from documents

Document understanding

Handwriting recognition

Intelligently fill in missing form data

Privacy-based NLP

Forecasting & optimization

Deep demand forecasting

Price optimization

Purchase modelling

Filling missing value

Lung cancer survival prediction

Classification

Detect malicious users and transactions

Fraud detection in financial transactions

Financial payment classification

Privacy sentiment classification

Credit risk prediction

Corporate credit rating prediction

Graph-based credit scoring

Explain credit decisions

Predictive maintenance

Detecting potential equipment failure for manufacturing

Predictive maintenance for vehicle fleets

Computer vision

Product defect detection

Bird species object detection

Autonomous driving

Visual perception with active learning

Reinforcement learning for battlesnake AI

Personalized recommendations

Entity resolution for smart advertising

Purchase modeling

Churn prediction

Churn prediction

Churn prediction for mobile phone customers

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Solutions: how it works

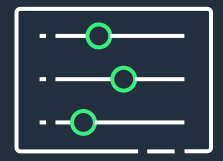
Amazon SageMaker Studio

Amazon SageMaker JumpStart

Prebuilt ML solutions that you can deploy quickly

Access and browse

Browse prebuilt solution templates using AWS CloudFormation for common business use cases



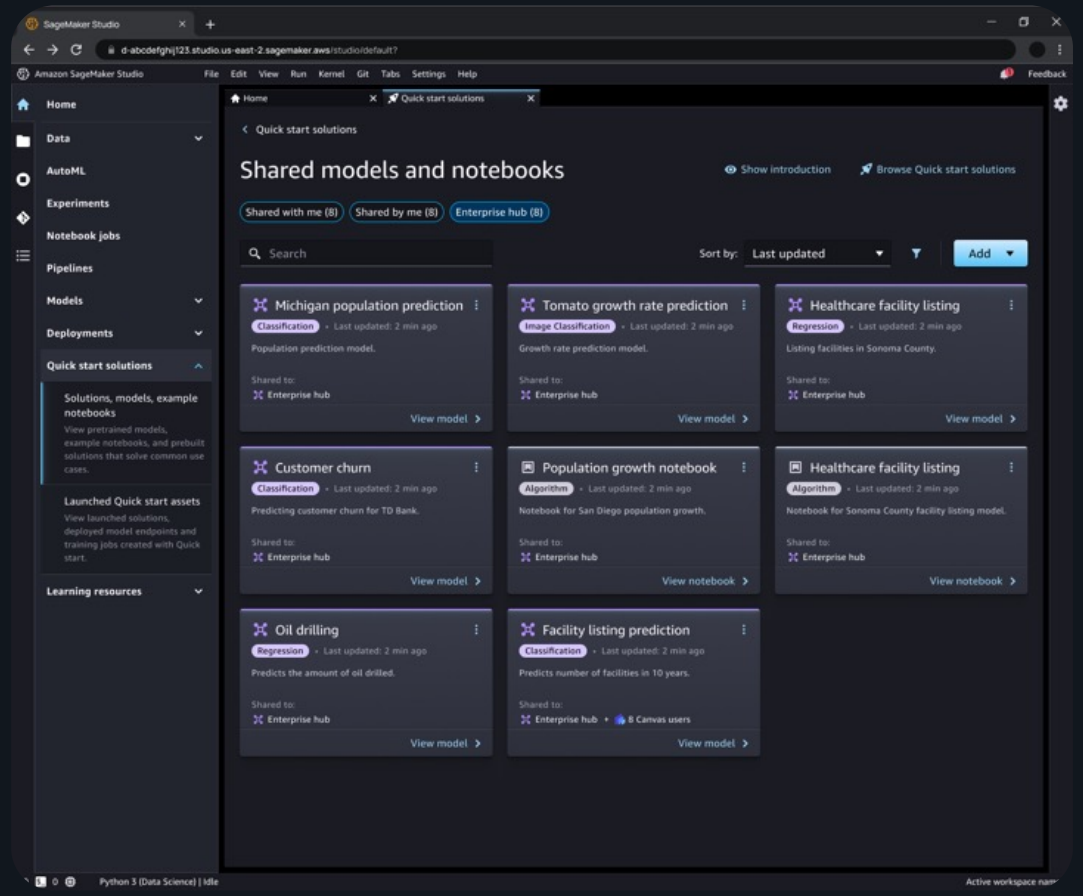
Select and customize

Select a template solution, which includes example datasets, and customize for your use cases using your own data

Deploy

Deploy solution with just a few clicks

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



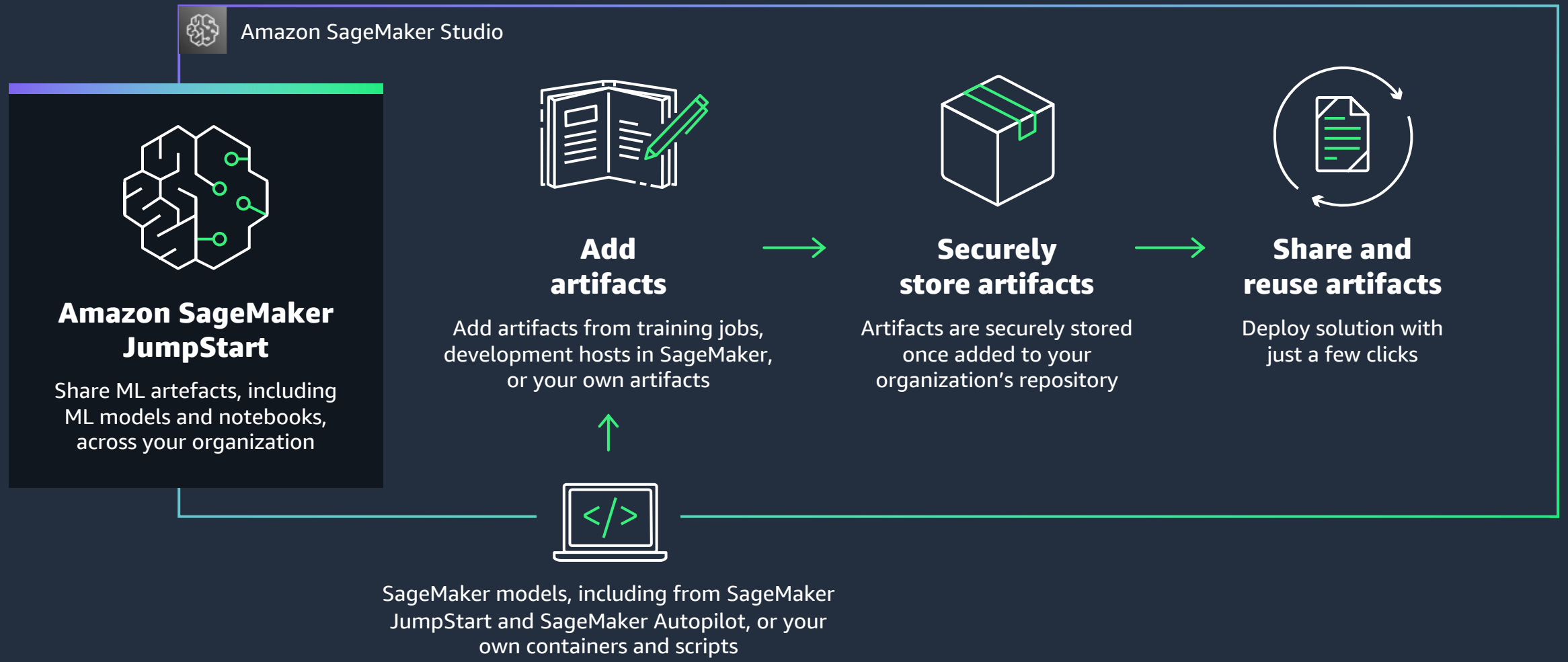
ML artifact sharing in SageMaker JumpStart

Enables data scientists to share ML artifacts securely within the enterprise and reuse alongside SageMaker built-in content

- Share with other users in your organization
- Discover shared contents easily and start fine-tuning with your own data
- Monitor and control what contents are shared

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ML artifact sharing: How it works



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker JumpStart pricing and availability

Generally available

Available to all SageMaker Studio customers **at no additional cost**

Pay only for resources used

Available in all regions where SageMaker Studio is available

(SageMaker JumpStart console experience in preview)

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Amazon Web Services, AWS, the Powered by AWS logo, and all AWS service names used in this slide deck are trademarks of Amazon.com, Inc. or its affiliates.