

AIM219

Learn and experiment with LLMs in Amazon SageMaker Studio Lab

Michele Monclova

Principal Product Manager
Amazon Web Services

Mia Chang

AI/ML Specialist Solutions Architect
Amazon Web Services

Ioan Catana

Sr. AI/ML Specialist Solutions Architect
Amazon Web Services

Vadim Omeltchenko

Sr. AI/ML Specialist Solutions Architect
Amazon Web Services

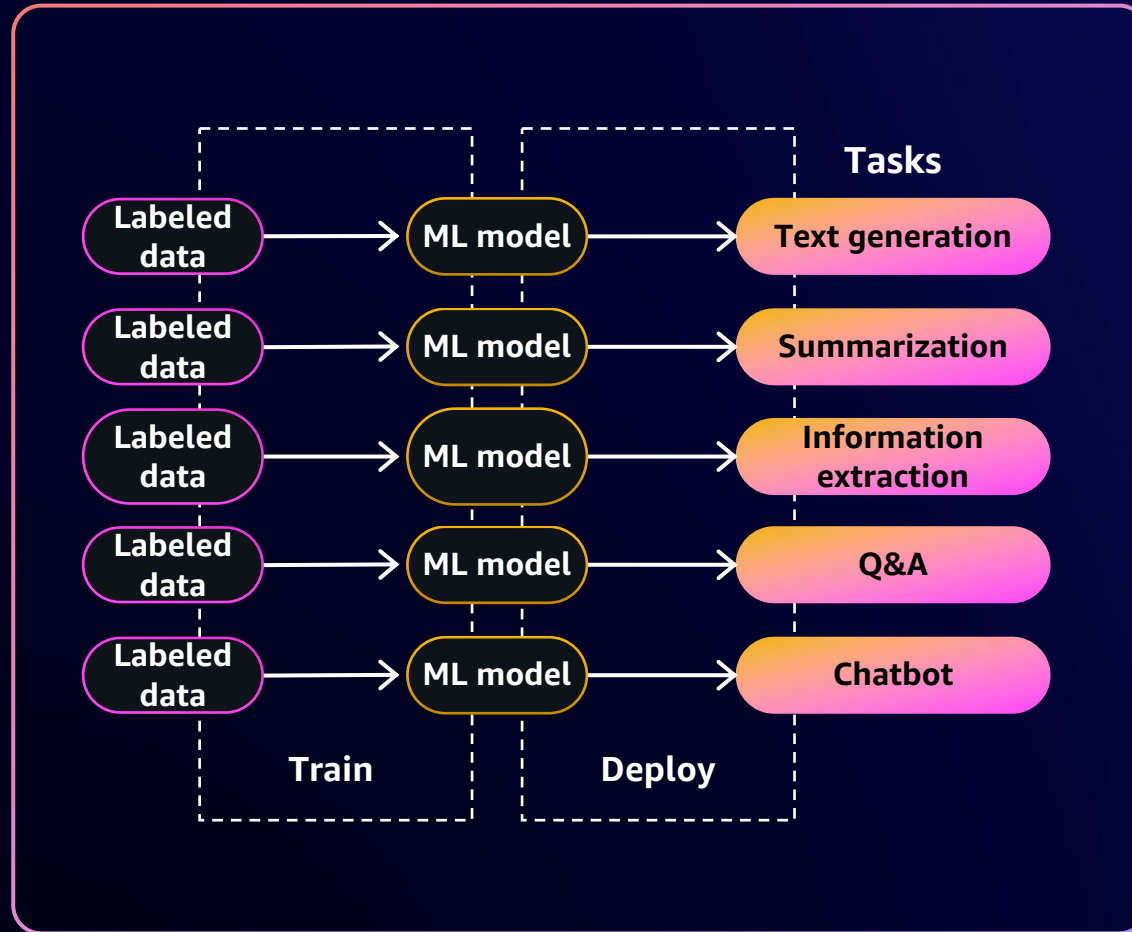


© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

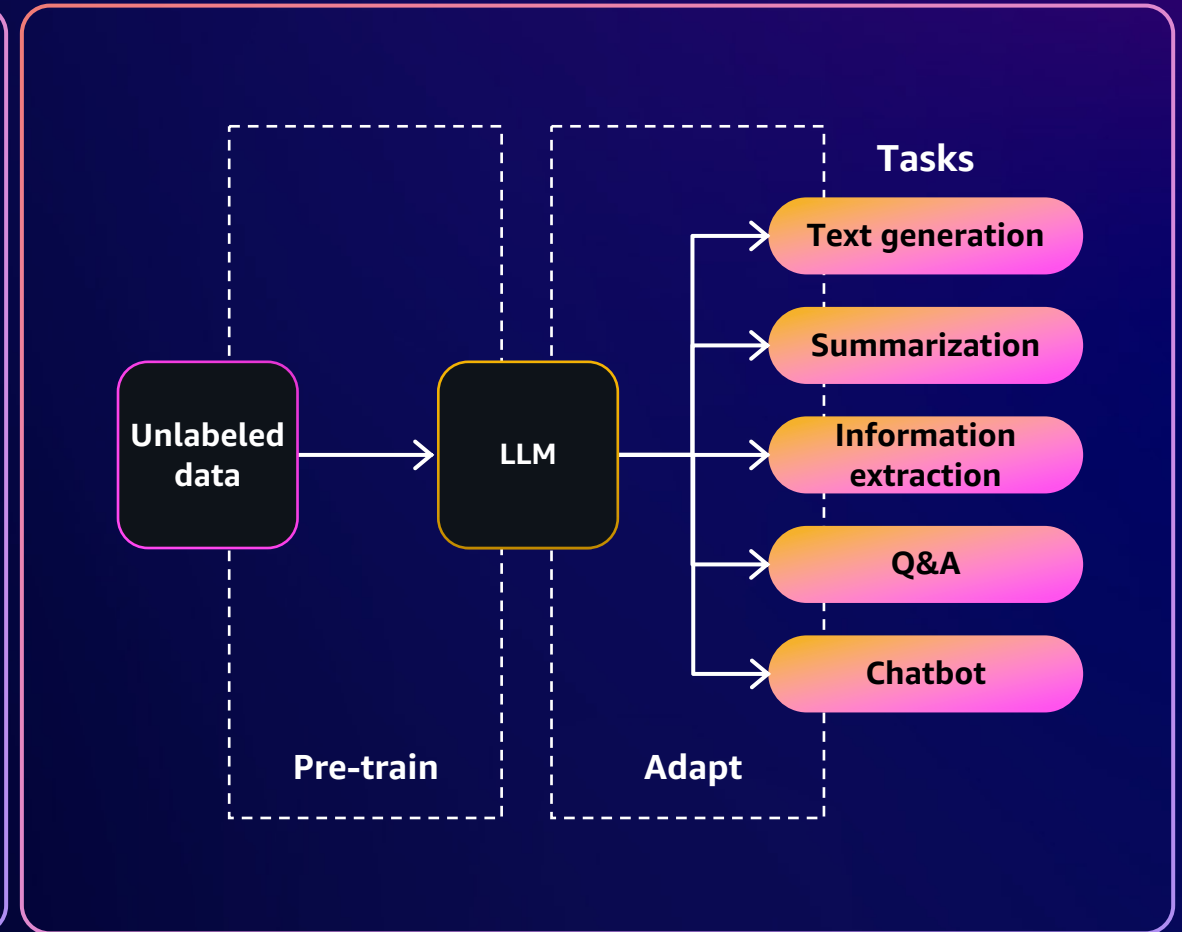
Today's session:

1. What are Large Language Models?
2. Overview of Amazon SageMaker Studio Lab
3. Account creation
4. Prompt engineering basics (hands on)
5. SageMaker vs SageMaker Studio Lab (demo)

How LLMs differ from other ML language models



Traditional ML Language Models



Large Language Models (LLMs)

How do LLMs work?

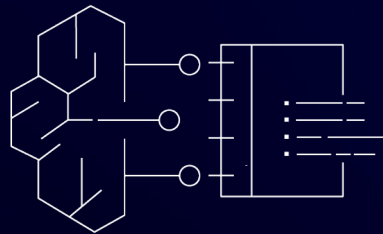


Studio Lab's Vision

Make machine learning and data science accessible to all customers (learners, developers, data scientists)

What is Amazon SageMaker Studio Lab

A JUPYTER NOTEBOOK SERVICE TO HELP CUSTOMERS MASTER THEIR SKILLS



Amazon SageMaker Studio Lab

A no-charge, no-configuration service that enable data scientists to learn and experiment with machine learning

Create an account with an email address – free

No setup or configuration required

15 GBs to save your work projects.

As many compute sessions as you need –
CPU (8 hrs)/GPU (4 hrs)

Access any notebook on GitHub

Migrate to SageMaker Studio when ready

Why Studio Lab for this exercise?

Access to variety of open source models

Privacy – it's your environment, you're in control

Ability to experiment with non-tuned and tuned models, compare, learn

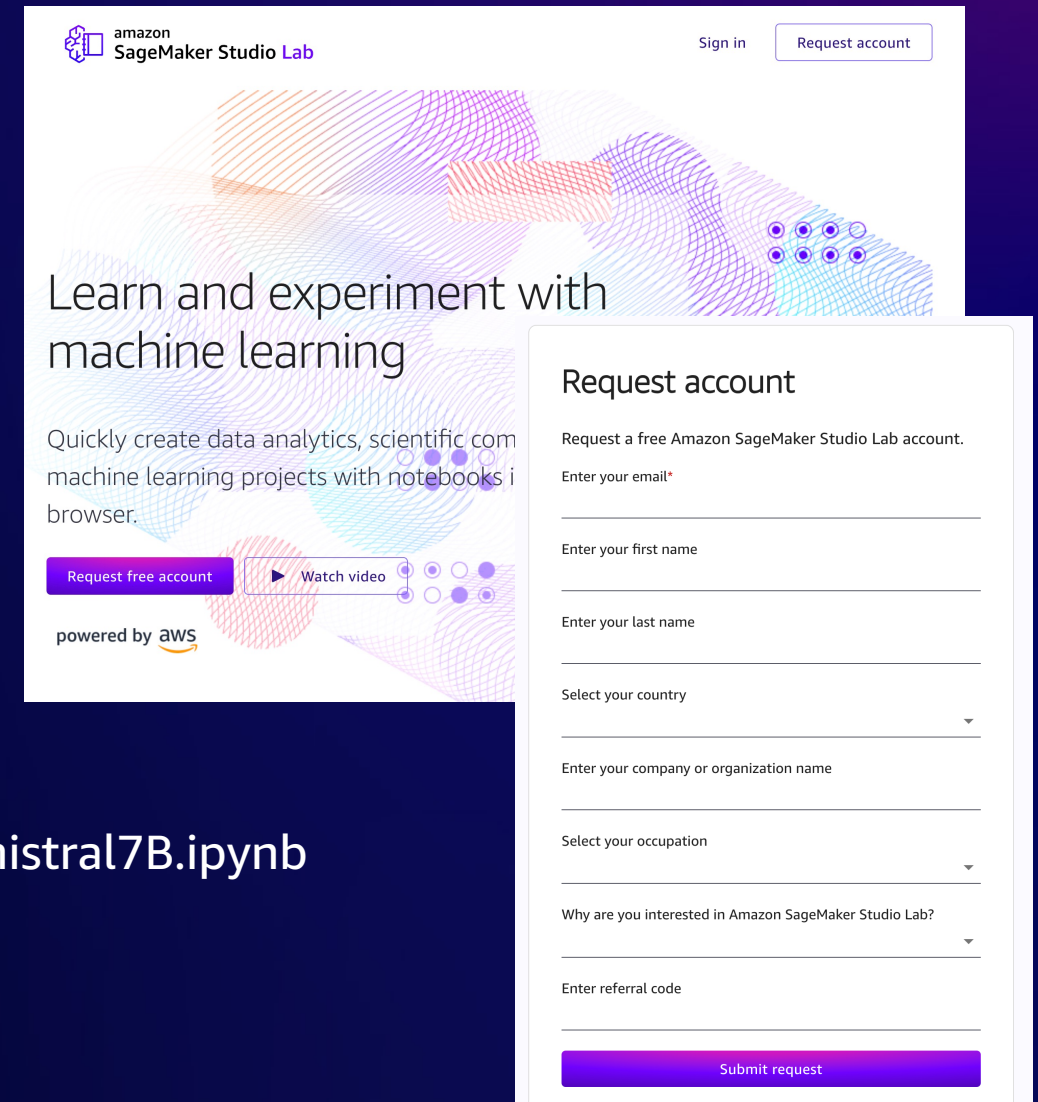
No cost environment

Start with Studio Lab, graduate to SageMaker to work with bigger models and more data

Signup for Studio Lab account

!! Use Referral Code !!

1. Go to <https://studiolab.sagemaker.aws/>
2. Request account
Use referral code: **reInvent23-875F6**
3. Confirm email
4. Sign into the account
5. git clone <https://github.com/aws/studio-lab-examples.git>
6. The notebook location:
`/studio-lab-examples/large-language-models/prompting-mistral7B.ipynb`



The screenshot shows the Amazon SageMaker Studio Lab website. At the top, there's a navigation bar with the Amazon SageMaker Studio Lab logo, a 'Sign in' link, and a 'Request account' button. The main heading is 'Learn and experiment with machine learning'. Below this, a subheading says 'Quickly create data analytics, scientific computing machine learning projects with notebooks in your browser.' There are two buttons: 'Request free account' and 'Watch video'. Below these is a 'powered by aws' logo. On the right side, there's a 'Request account' form. The form has the following fields: 'Enter your email*', 'Enter your first name', 'Enter your last name', 'Select your country' (a dropdown menu), 'Enter your company or organization name', 'Select your occupation' (a dropdown menu), 'Why are you interested in Amazon SageMaker Studio Lab?' (a dropdown menu), and 'Enter referral code'. At the bottom of the form is a 'Submit request' button.



<https://docs.aws.amazon.com/sagemaker/latest/dg/studio-lab-onboard.html>

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

Clean up for Studio Lab account

Q: How do I refresh my environment?

Switch to base conda environment:

```
conda activate base
```

List all conda environments available in your account:

```
conda list envs
```

Delete all the conda environments except environment named 'base':

```
conda remove -n "env_name" --all
```

Delete all sub-directories from user folder:

```
rm -rf /home/studio-lab-user/*
```

Restart project/runtime. Go to project page, stop instance and restart.

- * Make sure you have more than 6GB to start the lab
- * <https://studiolab.sagemaker.aws/faq>

Prompt Engineering: A lightspeed introduction

- What is a **Prompt**?
 - ✓ Text input provided to an AI system to elicit a response
- What is **Prompt Engineering**?
 - ✓ Using NLP techniques to craft prompts that steer FMs/LLMs towards desired responses
- Why is this important?
 - ✓ Enables fine-grained and strategic control over models' behavior
 - ✓ Targets desired capabilities
 - ✓ Mitigates risks



*NLP = Natural Language Processing
FM = Foundation Model
LLM = Large Language Model

Prompt Engineering: Zero-shot, One-shot, Few-shot



Zero shot learning

Generation: *Dear Mark, I would like to invest in your company with a minimum investment of \$100,000. John Write an email that kindly declines the offer:*

Summarization: *Summarize this article.... OUTPUT: Feds will not increase*

Code Generation: *"Create a sql to find all users who live in WA and have more than 3 cars"*



One-shot learning

Generation : *Task is to generate airport codes*

Text: "I want to fly form Los Angeles to Miami" Airport codes: LAX, MIA

Text: "I want to fly from Dallas to San Francisco" Airport codes:

Classification : *Tweet: " great pen" output GOOD Tweet: "great show" output*



Few-shot learning

Generation: *List the Country of origin of food. Pizza comes from Italy Burger comes from USA Curry comes from*

Classification: *Tweet: "I hate it when my phone battery dies.": Sentiment: Negative*

Tweet: "My day has been great": Sentiment: Positive

Tweet: "This is the link to the article": Sentiment: Neutral

Tweet: "This new music video was incredible" Sentiment:

Translation: *sea otter is loutre de mer <> cheese is <>*

Source : <https://huggingface.co/blog/few-shot-learning-gpt-neo-and-inference-api>



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

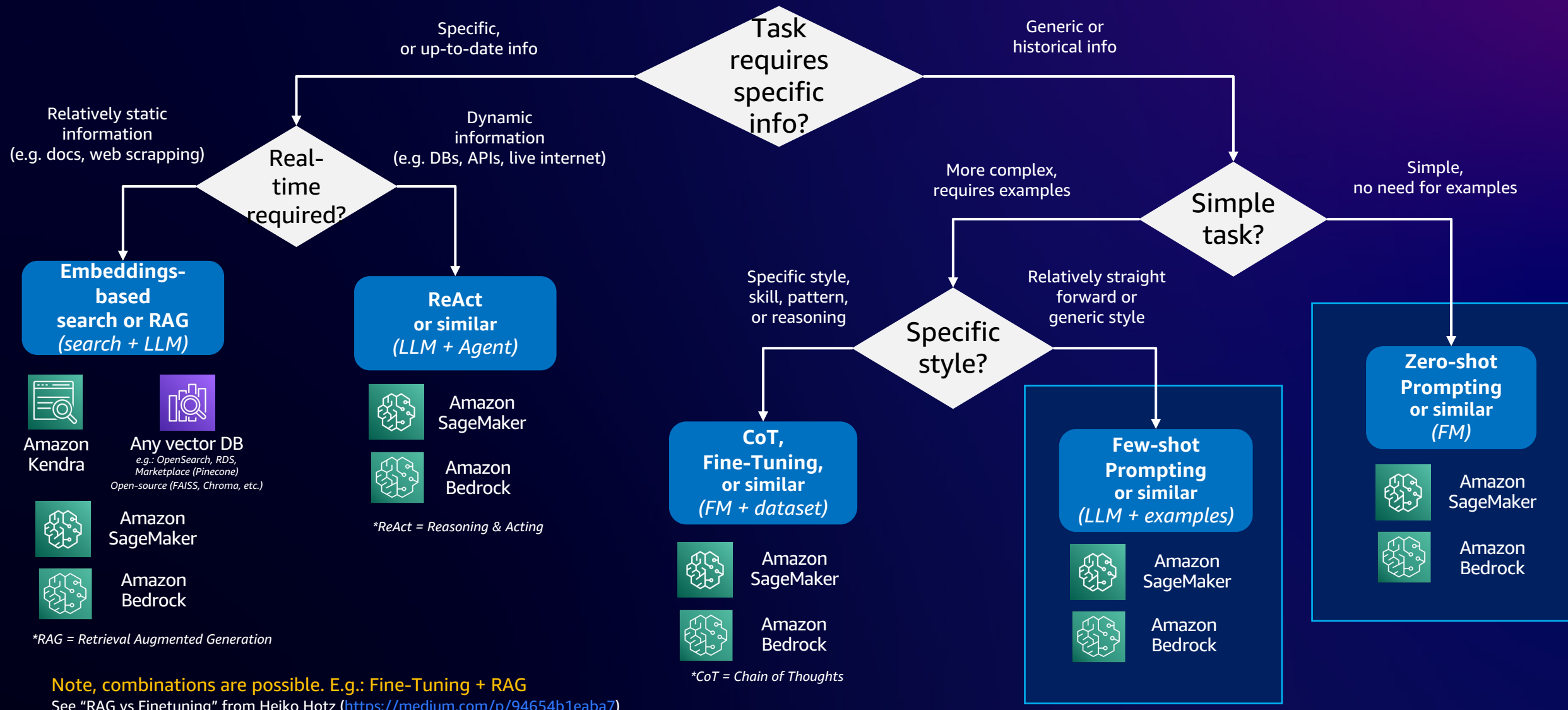
Tips for designing prompts

- Be clear and concise
- Include context if needed
- Use directives for the desired response type
- Consider the output in the prompt
- Provide an example response
- Use simple language
- Test and experiment

Graduation to SageMaker

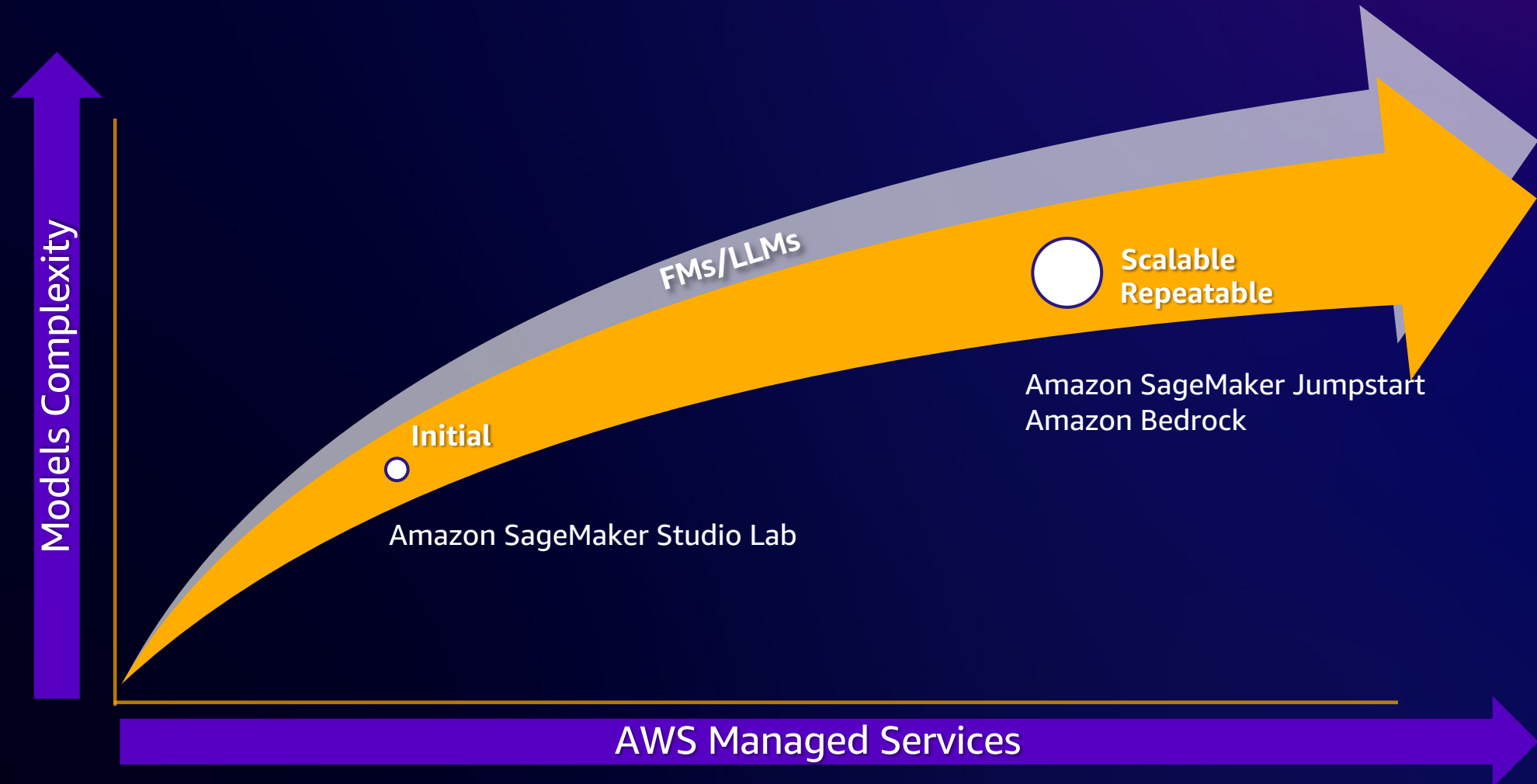


Choosing the right strategy for your GenAI task, with AWS



Note, combinations are possible. E.g.: Fine-Tuning + RAG
See "RAG vs Finetuning" from Heiko Hotz (<https://medium.com/p/94654b1eaba7>)

Graduation to SageMaker



SageMaker Studio Lab vs SageMaker



Amazon SageMaker Studio Lab

A no charge, no setup ML development environment

Amazon SageMaker

PREPARE

SageMaker Ground Truth

Label training data for machine learning

SageMaker Data Wrangler

Aggregate and prepare data for machine learning

SageMaker Processing

Built-in Python, BYO R/Spark

SageMaker Feature Store

Store, update, retrieve, and share features

SageMaker Clarify

Detect bias and understand model predictions

BUILD

SageMaker Studio Notebooks & Notebook Instances

Jupyter notebooks with elastic compute and sharing

Built-in and bring-your-own algorithms

Dozens of optimized algorithms or bring your own

Local mode

Test and prototype on your local machine

SageMaker Autopilot

Automatically create machine learning models with full visibility

SageMaker JumpStart

Pre-built solutions for common use cases

TRAIN & TUNE

Managed training

Distributed infrastructure management

SageMaker Experiments

Capture, organize, and compare every step

Automatic model tuning

Hyperparameter optimization

Distributed training libraries

Training for large datasets and models

SageMaker Debugger

Debug and profile training runs

Managed Spot training

Reduce training cost by 90%

DEPLOY & MANAGE

Managed deployment

Fully managed, ultra low latency, high throughput

Kubernetes and Kubeflow Integration

Simplify Kubernetes-based machine learning

Multi-model endpoints

Reduce cost by hosting multiple models per instance

SageMaker Model Monitor

Maintain accuracy of deployed models

SageMaker Edge Manager

Manage and monitor models on edge devices

SageMaker Pipelines

Workflow orchestration and automation

SageMaker Studio

Integrated development environment (IDE) for ML

Experiment with LLMs in Amazon SageMaker Studio Lab

Thank you!

<https://studiolab.sagemaker.aws/>



Please complete the session
survey in the mobile app