Model Evaluation Using Amazon Bedrock & Amazon SageMaker

Ray Strickland (he/his/him)

Sr PSA - WW Data & AI
6/4/24

Agenda:

Introduction to LLM Model Evaluation

Amazon’s Evaluation Tools

Selecting the Best Fit LLM

Deployment and Monitoring

Responsible AI Considerations

Continuous Evaluation and Refinement

Next Steps

The Current State of LLM Adoption

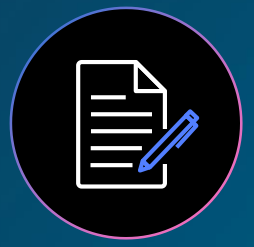
Capabilities

Generation

Question answering

Summarization

Translation



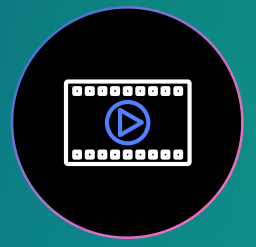
Correction

Classification

Input/Output

Text

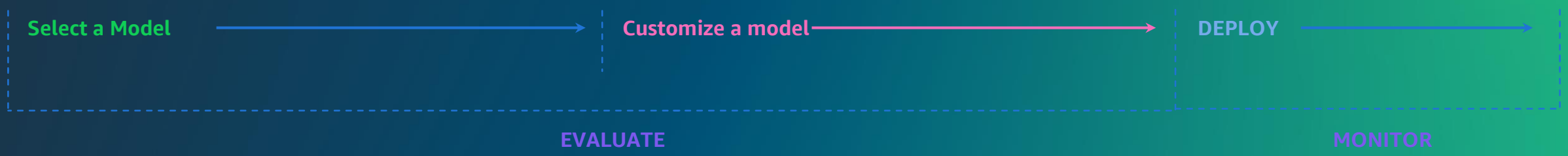
Images



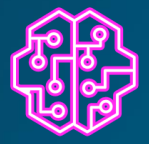
Video/Audio

Code

Challenges in Selecting the Optimal LLM



Hundreds of possible LLMs

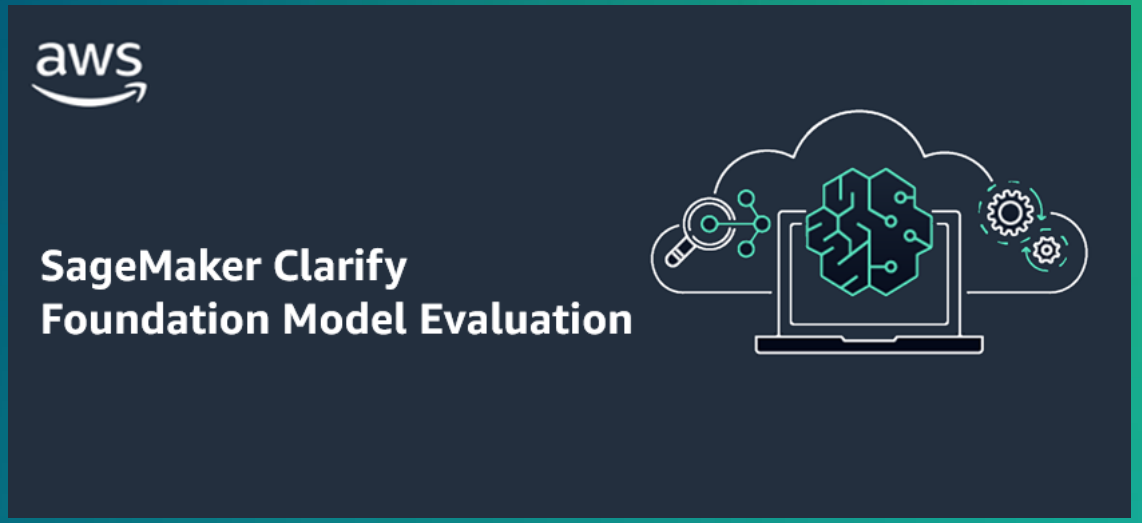


How to select evaluation data

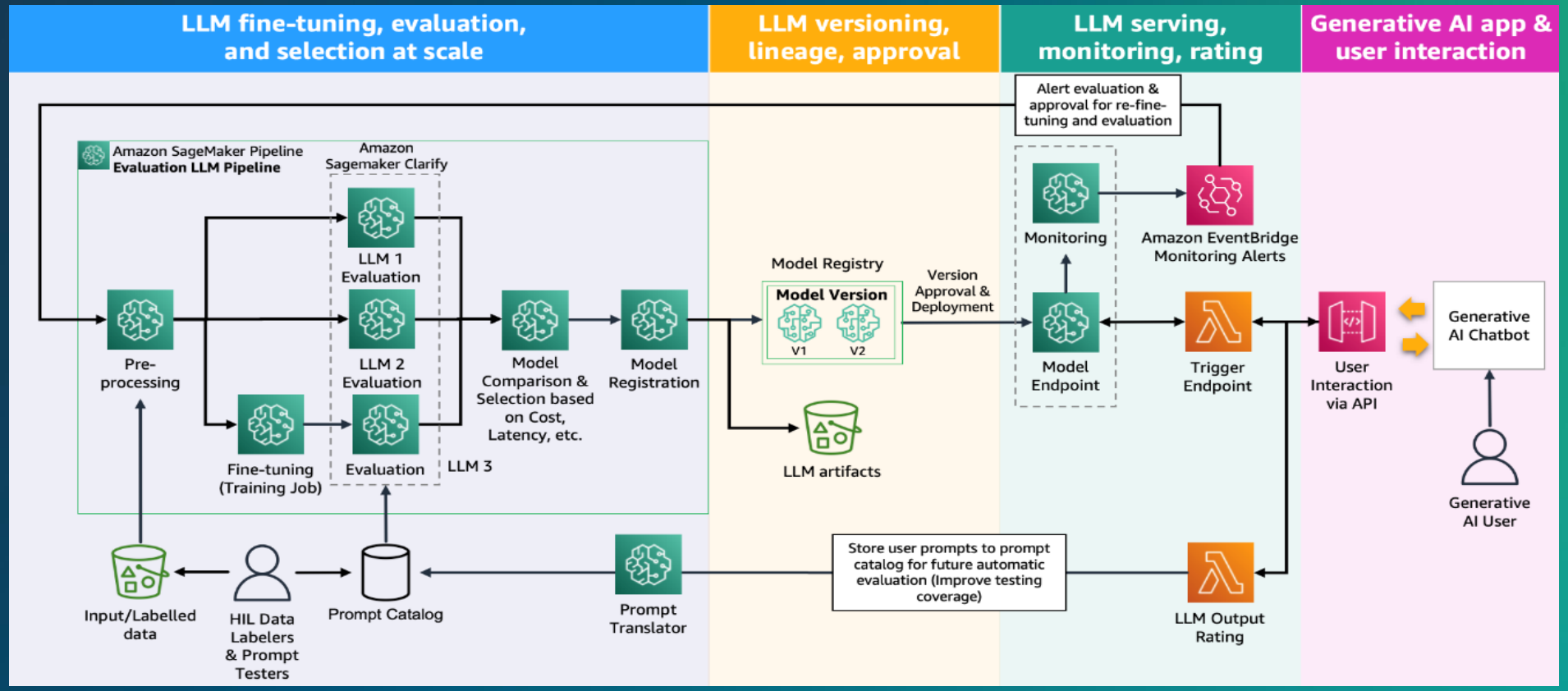
Defining criteria and deploying tools is time-consuming and expensive

How to monitor and assure responsible AI

Amazon SageMaker Clarify and Bedrock's Model Evaluation Tool



The Comprehensive Evaluation Architecture



Selecting the Best Fit LLM

Step 1: Define Your Requirements

- Use case and desired outcomes
- Model quality requirements
- Responsible AI
- Latency requirements
- Cost and budget considerations
- Other non-functional requirements

Selecting the Best Fit LLM

Step 2: Select Your Evaluation Data and Tasks

- Curating a representative dataset
- Defining relevant evaluation tasks
- Establishing appropriate metrics
- Considering data privacy and compliance
- Aligning with real-world usage

Selecting the Best Fit LLM

Step 3: Utilize Bedrock's Model Evaluation Tool

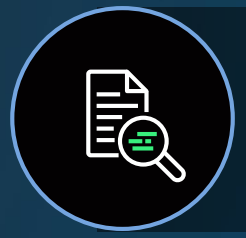
Benefits

- Automated benchmarking
- Cost comparison
- Complementary to Amazon SageMaker Clarify
- Integration with other AWS services

Bedrock evaluation demo

Selecting the Best Fit LLM

Step 4: Leverage Amazon SageMaker Clarify



Comprehensive model evaluation

Objective and repeatable evaluation

Integration with other AWS services

Scalability and flexibility

Selecting the Best Fit LLM

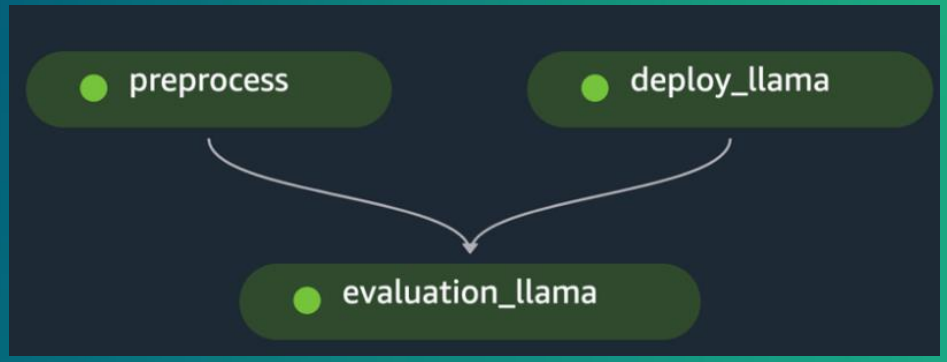
SageMaker evaluation demo

Selecting the Best Fit LLM

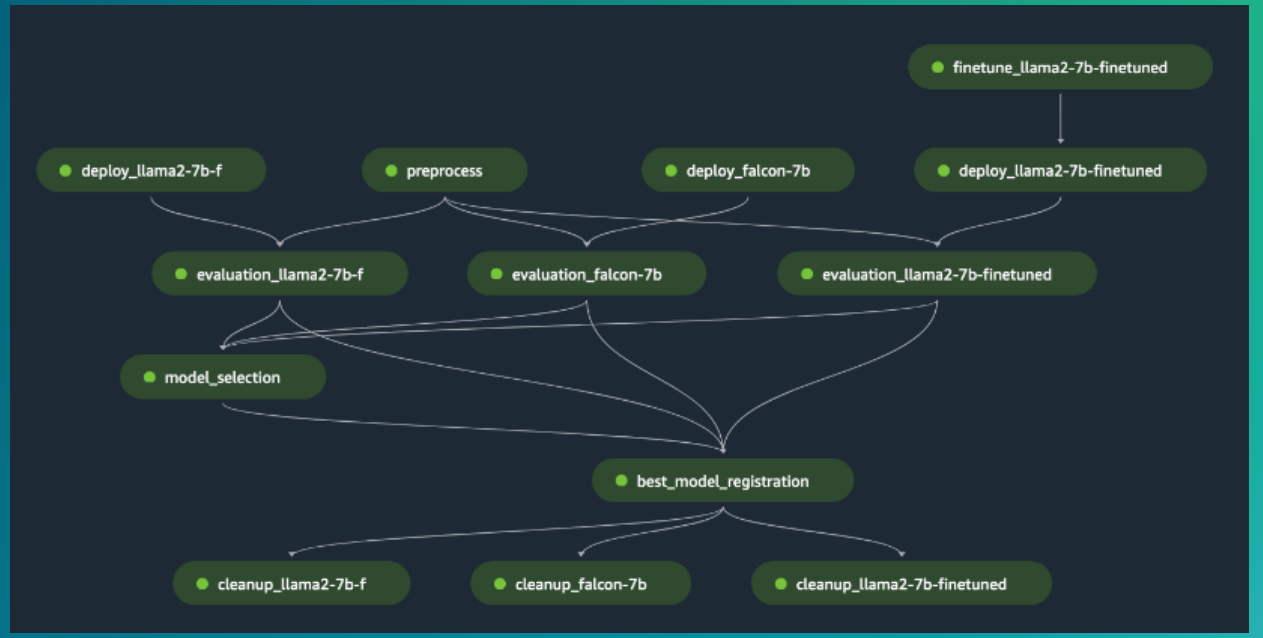
Optional: Leverage Amazon SageMaker Clarify

- Use the fmeval library to run automatic evaluations and customize your workflow
- Supports models on Jumpstart, Bedrock and even 3P models (eg: ChatGPT3.5 on HF)
- Supports built-in or custom datasets
- Supports Text generation, Summarization, Q&A and Classification
- Operationalize FM evaluation at Scale by combining with Amazon SageMaker Pipelines (MLOps)

Single model evaluation



Multi-model evaluation



Selecting the Best Fit LLM

Step 5: Interpret Results and Select the Best Fit

- Compare model performance across quality, latency, and cost metrics
- Consider trade-offs based on your priorities
- Align selection with your defined use case
- Consider broader perspectives
- Document the selection process and rationale

Deployment and Monitoring

Deploying Your GenAI Solution

Consider deployment option

Deploy the LLM as an Endpoint

Implement integration with CI/CD Pipelines

Leverage SageMaker's Security Features

Deployment and Monitoring

Implementing Comprehensive Monitoring

- Leverage Amazon CloudWatch for Monitoring and Observability
- Define Custom Metrics and Alarms
- Leverage CloudWatch Logs for Detailed Logging
- Implement Centralized Dashboards and Reporting
- Integrate with Other AWS Services

Establishing Responsible AI Practices

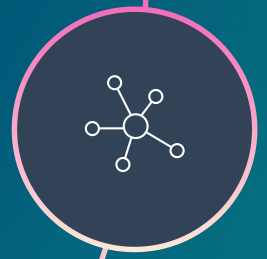
- Define use cases—the more specific & narrow, the better
- Prioritize education & diversity in your workforce
- Match processes to risk with a performance evaluation
- Test, test, test
- Distinguish application performance by dataset
- Share responsibility upstream and downstream

Guardrails for Amazon Bedrock

Implement safeguards customized to your application requirements and responsible AI policies

Apply guardrails to multiple foundation models and Agents for Amazon Bedrock

Configure harmful content filtering based on your responsible AI policies



Define and disallow denied topics with short natural language descriptions

Redact sensitive PII information in FM responses

Key Takeaways

- Leverage Amazon SageMaker and Bedrock for a Robust Evaluation Framework
- Establish Rigorous Monitoring and Observability
- Prioritize Responsible AI Practices
- Embrace a Flexible and Adaptable Approach
- Leverage the Power of AWS Services

Next Steps

- Explore the tools we discussed today
- Join the Amazon Partner Network (APN)
- Reach out to your AWS representative to schedule a one-on-one discussion
- Continue learning

Bedrock Model evaluation

Amazon SageMaker Clarify

FMEval GitHub Repository & Notebooks

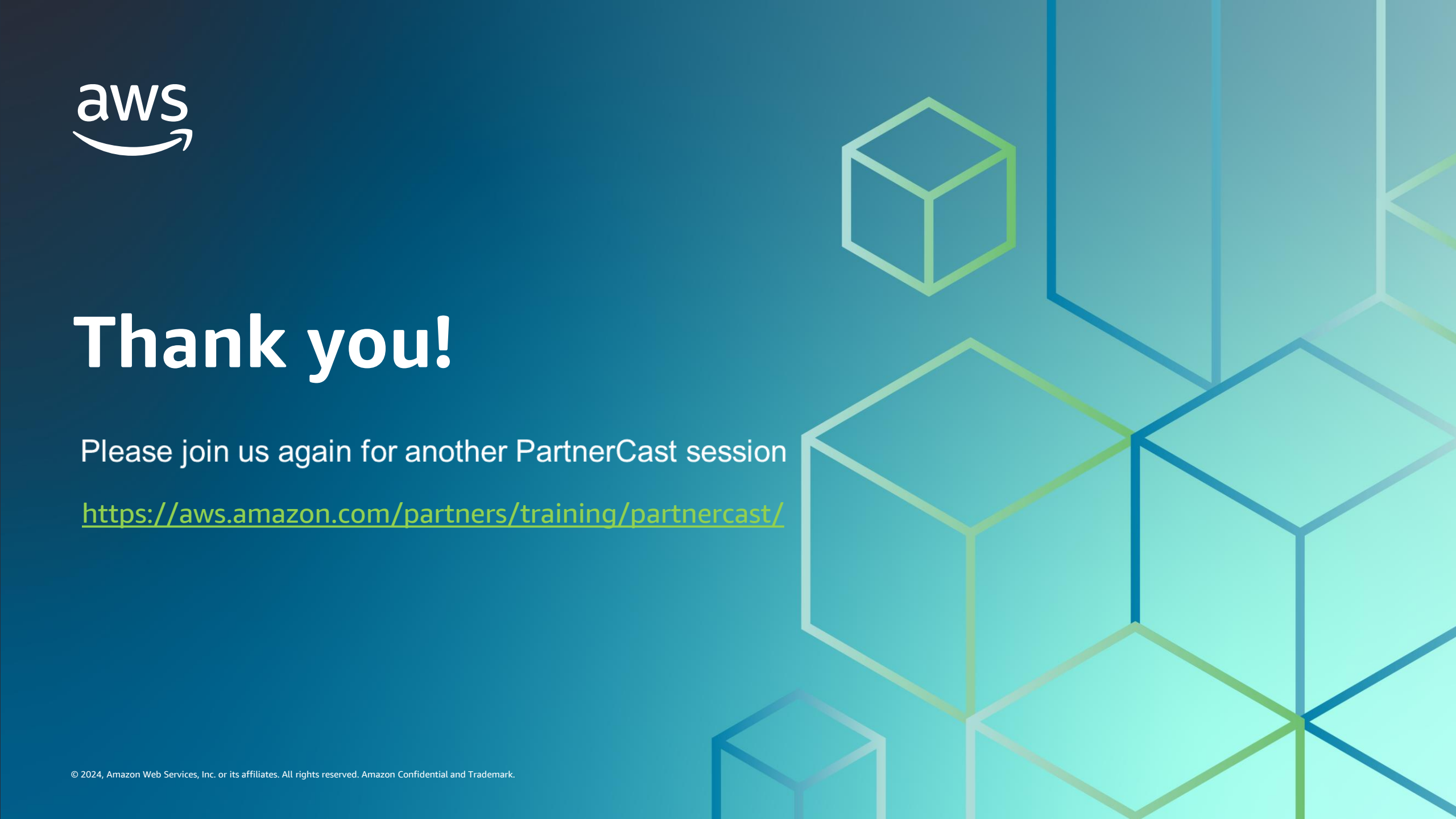
Generative AI for Builders on Amazon SageMaker

Help Us Improve Our Sessions & Delivery!

Feedback helps us plan upcoming PartnerCast sessions, modify content & become better presenters.

Scan QR code to leave feedback now:

You will also be automatically re-directed at the end of the webinar to the feedback survey!



Thank you!

Please join us again for another PartnerCast session

https://aws.amazon.com/partners/training/partnercast/

Q&A