DATA+AI SUMMIT BY databricks

GET READY TO BE DATABRICKS CERTIFIED: Generative AI Engineer Associate

James Kantor & Corey Abshire

YOUR SPEAKERS

James Kantor

Sr. Certification Developer

Corey Abshire

Sr. Specialist Solutions Architect

DATABRICKS CERTIFIED GENERATIVE AI ENGINEER ASSOCIATE

CERTIFICATION BACKGROUND

- Get certified in today's GenAI and Databricks
  - Exam = cutting edge LLM/Chained/Agents/RAG solutions PLUS Databricks
- A new role path for Databricks learning: GenAI
  - ML path is separate and not a prerequisite
- One exam to certify
  - Standard 2 year validity

EXAM DETAILS

- 45 questions x 90 minutes
- Scenario-based questions
    - Multiple choice or multiple selection
- $200/attempt
- https://www.webassessor.com/databricks

WHAT IS THE SCOPE OF THIS EXAM?

In-Scope

- Problem Decomposition
- Selecting Models, Tools, and Approaches
    - Prompt engineering, current API's, evaluation and monitoring
- Databricks Technology
    - Vector Search, Model Serving, MLflow, Unity Catalog

Out-of-Scope

- Fine tuning, Pre-training from scratch, Continued pre-training, Model Architecture

THE EXAM TARGET (MQC)

- MQC is the Minimally Qualified Candidate
  - Ideal candidate for the exam
- MQC can design and implement LLM-enabled solutions using Databricks
- MQC has consumed all learning and has six months hands-on experience

THE EXAM GUIDE

databricks

Databricks Exam Guide

Databricks Certified Generative AI Engineer Associate

Provide Exam Guide Feedback

Purpose of this Exam Guide

The purpose of this exam guide is to give you an overview of the exam and what is covered on the exam to help you determine your exam readiness. This document will get updated anytime there are any changes to an exam (and when those changes will take effect on an exam) so that you can be prepared.

Audience Description

The Databricks Certified Generative AI Engineer Associate certification exam assesses an individual's ability to design and implement LLM-enabled solutions using Databricks. This includes problem decomposition to break down complex requirements into manageable tasks as well as choosing appropriate models, tools, and approaches from the current generative AI landscape for developing comprehensive solutions. It also assesses Databricks-specific tools such as Vector Search for semantic similarity searches, Model Serving for deploying models and solutions, MLflow for managing solution lifecycle, and Unity Catalog for data governance. Individuals who pass this exam can be expected to build and deploy performant RAG applications and LLM chains that take full advantage of Databricks and its toolset.

MAJOR SKILL AREAS

- Problem Decomposition
- Selecting Models, Tools, and Approaches
    - Prompt engineering, LangChain, Hugging Face, current API's, LLM Evaluation
- Databricks Technology
    - Vector Search
    - Model Serving
    - MLflow
    - Unity Catalog

EXAM SECTIONS

1. Design Applications
2. Data Preparation
3. Application Development
4. Assembling and Deploying Applications
5. Governance
6. Evaluation and Monitoring

EXAM SECTIONS 1-3

Design Applications - 14%

- Prompt design
- Model selection
- Chaining
- Function calling (tools)
- Multi-stage reasoning

Data Preparation - 14%

- Chunking documents
- Data preparation
- Content pipelines
- Data sources
- Retrieval evaluation

App Development - 30%

- Data retrieval
- Implementing guardrails
- Prompt augmentation
- Minimizing hallucinations
- Context length

EXAM SECTIONS 4-6

Assembling and Deploying Applications - 22%

- Deploying via MLflow
- Access control
- Models in Unity Catalog
- Vector search indexes
- Model serving endpoints

Governance - 8%

- Guardrails and masking
- Malicious prompt protection
- Problematic text in RAG data sources
- Legal and licensing concerns

Evaluation and Monitoring - 12%

- Choosing an LLM
- Monitoring and metrics
- Performance evaluation with MLflow
- Inference logging
- LLM cost and controls for RAG

SAMPLE QUESTION

After changing the response-generating LLM in a RAG pipeline from GPT-4 to a model with a shorter context length that the company self-hosts, the Generative AI Engineer is getting the following error:

```
{"error_code":"BAD_REQUEST","message":"Bad request: rpc error: code = InvalidArgument desc = prompt token count (4595) cannot exceed 4096..."}
```

What TWO solutions should the Generative AI Engineer implement without changing the response generating model?

A. Use a smaller embedding model to generate embeddings
B. Reduce the maximum output tokens of the new model
C. Retrain the response generating model using ALiBi
D. Decrease the chunk size of embedded documents
E. Reduce the number of records retrieved from the vector database

EXAM PREPARATION

1. The Exam Guide, The Exam Guide, and The Exam Guide
2. Databricks Courseware
3. Review today’s LLM-based API’s and tools
4. Databricks Documentation

WHY GET CERTIFIED ON GENAI WITH DATABRICKS?

WHY CERTIFY?

- Confirm your skills in the most in-demand tech
- Show your organization you meet the Databricks standard now
- Get on Databricks GenAI path now for the future

Q&A