Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
A Generative AI Engineer is designing a RAG application for answering user questions on technical regulations as they learn a new sport.
What are the steps needed to build this RAG application and deploy it?

A. Ingest documents from a source –> Index the documents and saves to Vector Search–> User submits queries against an LLM –> LLM retrieves relevant documents –> Evaluate model –> LLM generates a response –> Deploy it using Model Serving
B. Ingest documents from a source –> Index the documents and save to Vector Search –> User submits queries against an LLM –> LLM retrieves relevant documents –> LLM generates a response -> Evaluate model –> Deploy it using Model Serving
C. Ingest documents from a source –> Index the documents and save to Vector Search –> Evaluate model –> Deploy it using Model Serving
D. User submits queries against an LLM –> Ingest documents from a source –> Index the documents and save to Vector Search –> LLM retrieves relevant documents –> LLM generates a response –> Evaluate model –> Deploy it using Model Serving

Answer: B

NEW QUESTION 2
A Generative AI Engineer is tasked with improving the RAG quality by addressing its inflammatory outputs.
Which action would be most effective in mitigating the problem of offensive text outputs?

A. Increase the frequency of upstream data updates
B. Inform the user of the expected RAG behavior
C. Restrict access to the data sources to a limited number of users
D. Curate upstream data properly that includes manual review before it is fed into the RAG system

Answer: D

NEW QUESTION 3
Which TWO chain components are required for building a basic LLM-enabled chat application that includes conversational capabilities, knowledge retrieval, and contextual memory?

A. (Q)
B. Vector Stores
C. Conversation Buffer Memory
D. External tools
E. Chat loaders
F. React Components

Answer: BC

NEW QUESTION 4
A Generative AI Engineer is developing a RAG system for their company to perform internal document Q&A for structured HR policies, but the answers returned are frequently incomplete and unstructured It seems that the retriever is not returning all relevant context The Generative AI Engineer has experimented with different embedding and response generating LLMs but that did not improve results.
Which TWO options could be used to improve the response quality? Choose 2 answers

A. Add the section header as a prefix to chunks
B. Increase the document chunk size
C. Split the document by sentence
D. Use a larger embedding model
E. Fine tune the response generation model

Answer: AB

NEW QUESTION 5
A Generative AI Engineer is building an LLM to generate article summaries in the form of a type of poem, such as a haiku, given the article content. However, the initial output from the LLM does not match the desired tone or style.
Which approach will NOT improve the LLM??s response to achieve the desired response?

A. Provide the LLM with a prompt that explicitly instructs it to generate text in the desired tone and style
B. Use a neutralizer to normalize the tone and style of the underlying documents
C. Include few-shot examples in the prompt to the LLM
D. Fine-tune the LLM on a dataset of desired tone and style

Answer: B

NEW QUESTION 6
A Generative AI Engineer is developing an LLM application that users can use to generate personalized birthday poems based on their names.
Which technique would be most effective in safeguarding the application, given the potential for malicious user inputs?

A. Implement a safety filter that detects any harmful inputs and ask the LLM to respond that it is unable to assist
B. Reduce the time that the users can interact with the LLM
C. Ask the LLM to remind the user that the input is malicious but continue the conversation with the user
D. Increase the amount of compute that powers the LLM to process input faster

Answer: A

NEW QUESTION 7
A Generative AI Engineer is creating an LLM-powered application that will need access to up-to-date news articles and stock prices.
The design requires the use of stock prices which are stored in Delta tables and finding the latest relevant news articles by searching the internet.
How should the Generative AI Engineer architect their LLM system?

A. Use an LLM to summarize the latest news articles and lookup stock tickers from the summaries to find stock prices.
B. Query the Delta table for volatile stock prices and use an LLM to generate a search query to investigate potential causes of the stock volatility.
C. Download and store news articles and stock price information in a vector stor
D. Use a RAG architecture to retrieve and generate at runtime.
E. Create an agent with tools for SQL querying of Delta tables and web searching, provide retrieved values to an LLM for generation of response.

Answer: D

NEW QUESTION 8
A Generative AI Engineer is tasked with developing an application that is based on an open source large language model (LLM). They need a foundation LLM with a large context window.
Which model fits this need?

A. DistilBERT
B. MPT-30B
C. Llama2-70B
D. DBRX

Answer: C

NEW QUESTION 9
A Generative AI Engineer is building a production-ready LLM system which replies directly to customers. The solution makes use of the Foundation Model API via provisioned throughput. They are concerned that the LLM could potentially respond in a toxic or otherwise unsafe way. They also wish to perform this with the least amount of effort.
Which approach will do this?

A. Host Llama Guard on Foundation Model API and use it to detect unsafe responses
B. Add some LLM calls to their chain to detect unsafe content before returning text
C. Add a regex expression on inputs and outputs to detect unsafe responses.
D. Ask users to report unsafe responses

Answer: A

NEW QUESTION 10
Generative AI Engineer at an electronics company just deployed a RAG application for customers to ask questions about products that the company carries. However, they received feedback that the RAG response often returns information about an irrelevant product.
What can the engineer do to improve the relevance of the RAG??s response?

A. Assess the quality of the retrieved context
B. Implement caching for frequently asked questions
C. Use a different LLM to improve the generated response
D. Use a different semantic similarity search algorithm

Answer: A

NEW QUESTION 10
A Generative AI Engineer is building a system which will answer questions on latest stock news articles.
Which will NOT help with ensuring the outputs are relevant to financial news?

A. Implement a comprehensive guardrail framework that includes policies for content filters tailored to the finance sector.
B. Increase the compute to improve processing speed of questions to allow greater relevancy analysisC Implement a profanity filter to screen out offensive language
C. Incorporate manual reviews to correct any problematic outputs prior to sending to the users

Answer: B

NEW QUESTION 11
A Generative AI Engineer is building an LLM-based application that has an important transcription (speech-to-text) task. Speed is essential for the success of the application
Which open Generative AI models should be used?

A. L!ama-2-70b-chat-hf
B. MPT-30B-Instruct
C. DBRX
D. whisper-large-v3 (1.6B)

Answer: D

NEW QUESTION 15
Which indicator should be considered to evaluate the safety of the LLM outputs when qualitatively assessing LLM responses for a translation use case?

A. The ability to generate responses in code

B. The similarity to the previous language
C. The latency of the response and the length of text generated
D. The accuracy and relevance of the responses

Answer: D

NEW QUESTION 16
A small and cost-conscious startup in the cancer research field wants to build a RAG application using Foundation Model APIs.
Which strategy would allow the startup to build a good-quality RAG application while being cost-conscious and able to cater to customer needs?

A. Limit the number of relevant documents available for the RAG application to retrieve from
B. Pick a smaller LLM that is domain-specific
C. Limit the number of queries a customer can send per day
D. Use the largest LLM possible because that gives the best performance for any general queries

Answer: B

NEW QUESTION 20
A Generative AI Engineer is building a Generative AI system that suggests the best matched employee team member to newly scoped projects. The team member is selected from a very large team. The match should be based upon project date availability and how well their employee profile matches the project scope. Both the employee profile and project scope are unstructured text.
How should the Generative AI Engineer architect their system?

A. Create a tool for finding available team members given project date
B. Embed all project scopes into a vector store, perform a retrieval using team member profiles to find the best team member.
C. Create a tool for finding team member availability given project dates, and another tool that uses an LLM to extract keywords from project scope
D. Iterate through available team members?? profiles and perform keyword matching to find the best available team member.
E. Create a tool to find available team members given project date
F. Create a second tool that can calculate a similarity score for a combination of team member profile and the project scop
G. Iterate through the team members and rank by best score to select a team member.
H. Create a tool for finding available team members given project date
I. Embed team profiles into a vector store and use the project scope and filtering to perform retrieval to find the available best matched team members.

Answer: D

NEW QUESTION 23
A Generative AI Engineer has already trained an LLM on Databricks and it is now ready to be deployed.
Which of the following steps correctly outlines the easiest process for deploying a model on Databricks?

A. Log the model as a pickle object, upload the object to Unity Catalog Volume, register it to Unity Catalog using MLflow, and start a serving endpoint
B. Log the model using MLflow during training, directly register the model to Unity Catalog using the MLflow API, and start a serving endpoint
C. Save the model along with its dependencies in a local directory, build the Docker image, and run the Docker container
D. Wrap the LLM??s prediction function into a Flask application and serve using Gunicorn

Answer: B

NEW QUESTION 25
A Generative AI Engineer is developing a patient-facing healthcare-focused chatbot. If the patient??s question is not a medical emergency, the chatbot should solicit more information from the patient to pass to the doctor??s office and suggest a few relevant pre-approved medical articles for reading. If the patient??s question is urgent, direct the patient to calling their local emergency services.
Given the following user input:
??I have been experiencing severe headaches and dizziness for the past two days.?? Which response is most appropriate for the chatbot to generate?

A. Here are a few relevant articles for your browsin
B. Let me know if you have questions after reading them.
C. Please call your local emergency services.
D. Headaches can be toug
E. Hope you feel better soon!
F. Please provide your age, recent activities, and any other symptoms you have noticed along with your headaches and dizziness.

Answer: B

NEW QUESTION 28
A Generative AI Engineer is building a RAG application that will rely on context retrieved from source documents that are currently in PDF format. These PDFs can contain both text and images. They want to develop a solution using the least amount of lines of code.
Which Python package should be used to extract the text from the source documents?

A. flask
B. beautifulsoup
C. unstructured
D. numpy

Answer: B

NEW QUESTION 30
A Generative AI Engineer wants to build an LLM-based solution to help a restaurant improve its online customer experience with bookings by automatically handling common customer inquiries. The goal of the solution is to minimize escalations to human intervention and phone calls while maintaining a personalized

interaction. To design the solution, the Generative AI Engineer needs to define the input data to the LLM and the task it should perform.
Which input/output pair will support their goal?

A. Input: Online chat logs; Output: Group the chat logs by users, followed by summarizing each user??s interactions
B. Input: Online chat logs; Output: Buttons that represent choices for booking details
C. Input: Customer reviews; Output: Classify review sentiment
D. Input: Online chat logs; Output: Cancellation options

Answer: B

NEW QUESTION 34
A Generative AI Engineer is helping a cinema extend its website's chat bot to be able to respond to questions about specific showtimes for movies currently playing at their local theater. They already have the location of the user provided by location services to their agent, and a Delta table which is continually updated with the latest showtime information by location. They want to implement this new capability In their RAG application.
Which option will do this with the least effort and in the most performant way?

A. Create a Feature Serving Endpoint from a FeatureSpec that references an online store synced from the Delta tabl
B. Query the Feature Serving Endpoint as part of the agent logic/ tool implementation.
C. Query the Delta table directly via a SQL query constructed from the user's input using a text-to-SQL LLM in the agent logic / tool
D. implementatio
E. Write the Delta table contents to a text column.then embed those texts using an embedding model and store these in the vector index Lookup the information based on the embedding as part of the agent logic / tool implementation.
F. Set up a task in Databricks Workflows to write the information in the Delta table periodically to an external database such as MySQL and query the information from there as part of the agent logic / tool implementation.

Answer: A

NEW QUESTION 37
A Generative AI Engineer is developing a RAG application and would like to experiment with different embedding models to improve the application performance.
Which strategy for picking an embedding model should they choose?

A. Pick an embedding model trained on related domain knowledge
B. Pick the most recent and most performant open LLM released at the time
C. pick the embedding model ranked highest on the Massive Text Embedding Benchmark (MTEB) leaderboard hosted by HuggingFace
D. Pick an embedding model with multilingual support to support potential multilingual user questions

Answer: A

NEW QUESTION 40
A Generative AI Engineer just deployed an LLM application at a digital marketing company that assists with answering customer service inquiries.
Which metric should they monitor for their customer service LLM application in production?

A. Number of customer inquiries processed per unit of time
B. Energy usage per query
C. Final perplexity scores for the training of the model
D. HuggingFace Leaderboard values for the base LLM

Answer: A

NEW QUESTION 42
A Generative AI Engineer is testing a simple prompt template in LangChain using the code below, but is getting an error.

```
from langchain.chains import LLMChain
from langchain_community.llms import OpenAI
from langchain_core.prompts import PromptTemplate

prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
    input_variables=["adjective"],
    template=prompt_template
)

llm = LLMChain(prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

Assuming the API key was properly defined, what change does the Generative AI Engineer need to make to fix their chain?

A)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
     input_variables=["adjective"],
     template=prompt_template
)

llm = LLMChain(prompt=prompt)
llm.generate("funny")
```

B)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
     input_variables=["adjective"],
     template=prompt_template
)

llm = LLMChain(prompt=prompt.format("funny"))
llm.generate()
```

C)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
     input_variables=["adjective"],
     template=prompt_template
     llm=OpenAI()
)

llm = LLMChain(prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

D)

```
prompt_template = "Tell me a {adjective} joke"

prompt = PromptTemplate(
     input_variables=["adjective"],
     template=prompt_template
)

llm = LLMChain(llm=OpenAI(), prompt=prompt)
llm.generate([{"adjective": "funny"}])
```

A. Option A
B. Option B
C. Option C
D. Option D

Answer: C

NEW QUESTION 46

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

Databricks-Generative-AI-Engineer-Associate Practice Exam Features:

* Databricks-Generative-AI-Engineer-Associate Questions and Answers Updated Frequently

* Databricks-Generative-AI-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff

* Databricks-Generative-AI-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* Databricks-Generative-AI-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year