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
What is the most suitable library for building a multi-step LLM-based workflow?

A. Pandas
B. TensorFlow
C. PySpark
D. LangChain

Answer: D

NEW QUESTION 3
A Generative Al 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 4
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 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 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 7
A Generative Al Engineer is creating an LLM-based application. The documents for its retriever have been chunked to a maximum of 512 tokens each. The

Generative AI Engineer knows that cost and latency are more important than quality for this application. They have several context length levels to choose from. Which will fulfill their need?

A. context length 514; smallest model is 0.44GB and embedding dimension 768
B. context length 2048: smallest model is 11GB and embedding dimension 2560
C. context length 32768: smallest model is 14GB and embedding dimension 4096
D. context length 512: smallest model is 0.13GB and embedding dimension 384

Answer: D

NEW QUESTION 8
A Generative AI Engineer has been asked to design an LLM-based application that accomplishes the following business objective: answer employee HR questions using HR PDF documentation.
Which set of high level tasks should the Generative AI Engineer's system perform?

A. Calculate averaged embeddings for each HR document, compare embeddings to user query to find the best documen
B. Pass the best document with the user query into an LLM with a large context window to generate a response to the employee.
C. Use an LLM to summarize HR documentatio
D. Provide summaries of documentation and user query into an LLM with a large context window to generate a response to the user.
E. Create an interaction matrix of historical employee questions and HR documentatio
F. Use ALS to factorize the matrix and create embedding
G. Calculate the embeddings of new queries and use them to find the best HR documentatio
H. Use an LLM to generate a response to the employee question based upon the documentation retrieved.
I. Split HR documentation into chunks and embed into a vector stor
J. Use the employee question to retrieve best matched chunks of documentation, and use the LLM to generate a response to the employee based upon the documentation retrieved.

Answer: D

NEW QUESTION 9
A Generative AI Engineer I using the code below to test setting up a vector store:

```
from databricks.vector_search.client import VectorSearchClient

vsc = VectorSearchClient()

vsc.create_endpoint(
     name= "vector_search_test",
     endpoint_type= "STANDARD"
)
```

Assuming they intend to use Databricks managed embeddings with the default embedding model, what should be the next logical function call?

A. vsc.get_index()
B. vsc.create_delta_sync_index()
C. vsc.create_direct_access_index()
D. vsc.similarity_search()

Answer: B

NEW QUESTION 10
A Generative Al Engineer interfaces with an LLM with prompt/response behavior that has been trained on customer calls inquiring about product availability. The LLM is designed to output ??In Stock?? if the product is available or only the term ??Out of Stock?? if not.
Which prompt will work to allow the engineer to respond to call classification labels correctly?

A. Respond with ??In Stock?? if the customer asks for a product.
B. You will be given a customer call transcript where the customer asks about product availabilit
C. The outputs are either ??In Stock?? or ??Out of Stock??. Format the output in JSON, for example: {??call_id??: ??123??, ??label??: ??In Stock??}.
D. Respond with ??Out of Stock?? if the customer asks for a product.
E. You will be given a customer call transcript where the customer inquires about product availabilit
F. Respond with ??In Stock?? if the product is available or ??Out of Stock?? if not.

Answer: B

NEW QUESTION 10
A Generative AI Engineer developed an LLM application using the provisioned throughput Foundation Model API. Now that the application is ready to be deployed, they realize their volume of requests are not sufficiently high enough to create their own provisioned throughput endpoint. They want to choose a strategy that ensures the best cost- effectiveness for their application.
What strategy should the Generative AI Engineer use?

A. Switch to using External Models instead
B. Deploy the model using pay-per-token throughput as it comes with cost guarantees
C. Change to a model with a fewer number of parameters in order to reduce hardware constraint issues

D. Throttle the incoming batch of requests manually to avoid rate limiting issues

Answer: B

NEW QUESTION 14
A Generative AI Engineer is designing an LLM-powered live sports commentary platform. The platform provides real-time updates and LLM-generated analyses for any users who would like to have live summaries, rather than reading a series of potentially outdated news articles.
Which tool below will give the platform access to real-time data for generating game analyses based on the latest game scores?

A. DatabrickslQ
B. Foundation Model APIs
C. Feature Serving
D. AutoML

Answer: C

NEW QUESTION 17
A Generative AI Engineer is developing a chatbot designed to assist users with insurance- related queries. The chatbot is built on a large language model (LLM) and is conversational. However, to maintain the chatbot??s focus and to comply with company policy, it must not provide responses to questions about politics. Instead, when presented with political inquiries, the chatbot should respond with a standard message:
??Sorry, I cannot answer that. I am a chatbot that can only answer questions around insurance.??
Which framework type should be implemented to solve this?

A. Safety Guardrail
B. Security Guardrail
C. Contextual Guardrail
D. Compliance Guardrail

Answer: A

NEW QUESTION 22
A Generative Al Engineer has developed an LLM application to answer questions about internal company policies. The Generative AI Engineer must ensure that the application doesn??t hallucinate or leak confidential data.
Which approach should NOT be used to mitigate hallucination or confidential data leakage?

A. Add guardrails to filter outputs from the LLM before it is shown to the user
B. Fine-tune the model on your data, hoping it will learn what is appropriate and not
C. Limit the data available based on the user??s access level
D. Use a strong system prompt to ensure the model aligns with your needs.

Answer: B

NEW QUESTION 24
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 26
A Generative Al Engineer is working with a retail company that wants to enhance its customer experience by automatically handling common customer inquiries. They are working on an LLM-powered Al solution that should improve response times while maintaining a personalized interaction. They want to define the appropriate input and LLM task to do this.
Which input/output pair will do this?

A. Input: Customer reviews; Output Group the reviews by users and aggregate per-user average rating, then respond
B. Input: Customer service chat logs; Output Group the chat logs by users, followed by summarizing each user's interactions, then respond
C. Input: Customer service chat logs; Output: Find the answers to similar questions and respond with a summary
D. Input: Customer reviews: Output Classify review sentiment

Answer: C

NEW QUESTION 31
A Generative Al 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 33
What is an effective method to preprocess prompts using custom code before sending them to an LLM?

A. Directly modify the LLM??s internal architecture to include preprocessing steps
B. It is better not to introduce custom code to preprocess prompts as the LLM has not been trained with examples of the preprocessed prompts
C. Rather than preprocessing prompts, it??s more effective to postprocess the LLM outputs to align the outputs to desired outcomes
D. Write a MLflow PyFunc model that has a separate function to process the prompts

Answer: D

NEW QUESTION 37
A Generative AI Engineer has built an LLM-based system that will automatically translate user text between two languages. They now want to benchmark multiple LLM's on this task and pick the best one. They have an evaluation set with known high quality translation examples. They want to evaluate each LLM using the evaluation set with a performant metric.
Which metric should they choose for this evaluation?

A. ROUGE metric
B. BLEU metric
C. NDCG metric
D. RECALL metric

Answer: B

NEW QUESTION 42
When developing an LLM application, it??s crucial to ensure that the data used for training the model complies with licensing requirements to avoid legal risks.
Which action is NOT appropriate to avoid legal risks?

A. Reach out to the data curators directly before you have started using the trained model to let them know.
B. Use any available data you personally created which is completely original and you can decide what license to use.
C. Only use data explicitly labeled with an open license and ensure the license terms are followed.
D. Reach out to the data curators directly after you have started using the trained model to let them know.

Answer: D

NEW QUESTION 46
A Generative AI Engineer is designing a chatbot for a gaming company that aims to engage users on its platform while its users play online video games.
Which metric would help them increase user engagement and retention for their platform?

A. Randomness
B. Diversity of responses
C. Lack of relevance
D. Repetition of responses

Answer: B

NEW QUESTION 51
A Generative AI Engineer would like an LLM to generate formatted JSON from emails. This will require parsing and extracting the following information: order ID, date, and sender email. Here??s a sample email:

```
Date: April 23, 2024
Time: 4:22 PM
From: anjali.thayer@computex.org
To: cust_service@realtek.com
Subject: Shipment details

Hey there,

I have a shipment (order ID is CD34RFT) can you please send me an update?

Thank you,
Anjali
```

They will need to write a prompt that will extract the relevant information in JSON format with the highest level of output accuracy.
Which prompt will do that?

A. You will receive customer emails and need to extract date, sender email, and order I
B. You should return the date, sender email, and order ID information in JSON format.
C. You will receive customer emails and need to extract date, sender email, and order I
D. Return the extracted information in JSON format.Here??s an example: {??date??: ??April 16, 2024??, ??sender_email??: ??sarah.lee925@gmail.com??,??order_id??: ??RE987D??}
E. You will receive customer emails and need to extract date, sender email, and order I
F. Return the extracted information in a human-readable format.
G. You will receive customer emails and need to extract date, sender email, and order I

H. Return the extracted information in JSON format.

Answer: B

NEW QUESTION 52
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 53
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 55
A Generative Al 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 58
A Generative AI Engineer has a provisioned throughput model serving endpoint as part of a RAG application and would like to monitor the serving endpoint??s incoming requests and outgoing responses. The current approach is to include a micro-service in between the endpoint and the user interface to write logs to a remote server.
Which Databricks feature should they use instead which will perform the same task?

A. Vector Search
B. Lakeview
C. DBSQL
D. Inference Tables

Answer: D

NEW QUESTION 61
A Generative Al 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 62
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 65
.....

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