Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
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 2
A Generative AI Engineer is building a system that will answer questions on currently unfolding news topics. As such, it pulls information from a variety of sources including articles and social media posts. They are concerned about toxic posts on social media causing toxic outputs from their system.
Which guardrail will limit toxic outputs?

A. Use only approved social media and news accounts to prevent unexpected toxic data from getting to the LLM.
B. Implement rate limiting
C. Reduce the amount of context Items the system will Include in consideration for its response.
D. Log all LLM system responses and perform a batch toxicity analysis monthly.

Answer: A

NEW QUESTION 3
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 4
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 5
A Generative AI Engineer needs to design an LLM pipeline to conduct multi-stage reasoning that leverages external tools. To be effective at this, the LLM will need to plan and adapt actions while performing complex reasoning tasks.
Which approach will do this?

A. Tram the LLM to generate a single, comprehensive response without interacting with any external tools, relying solely on its pre-trained knowledge.
B. Implement a framework like ReAct which allows the LLM to generate reasoning traces and perform task-specific actions that leverage external tools if necessary.
C. Encourage the LLM to make multiple API calls in sequence without planning or structuring the calls, allowing the LLM to decide when and how to use external tools spontaneously.
D. Use a Chain-of-Thought (CoT) prompting technique to guide the LLM through a series of reasoning steps, then manually input the results from external tools for the final answer.

Answer: B

NEW QUESTION 6
A Generative AI 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 7
A Generative Al 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 8
A Generative Al Engineer is using an LLM to classify species of edible mushrooms based on text descriptions of certain features. The model is returning accurate responses in testing and the Generative Al Engineer is confident they have the correct list of possible labels, but the output frequently contains additional reasoning in the answer when the Generative Al Engineer only wants to return the label with no additional text.
Which action should they take to elicit the desired behavior from this LLM?

A. Use few snot prompting to instruct the model on expected output format
B. Use zero shot prompting to instruct the model on expected output format
C. Use zero shot chain-of-thought prompting to prevent a verbose output format
D. Use a system prompt to instruct the model to be succinct in its answer

Answer: D

NEW QUESTION 9
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 10
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
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? (Choose two.)

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

Answer: CD

NEW QUESTION 13
A Generative Al Engineer has developed an LLM application to answer questions about internal company policies. The Generative Al 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 16
A Generative AI Engineer received the following business requirements for an external chatbot.
The chatbot needs to know what types of questions the user asks and routes to appropriate models to answer the questions. For example, the user might ask about upcoming event details. Another user might ask about purchasing tickets for a particular event.
What is an ideal workflow for such a chatbot?

A. The chatbot should only look at previous event information
B. There should be two different chatbots handling different types of user queries.
C. The chatbot should be implemented as a multi-step LLM workflo
D. First, identify the type of question asked, then route the question to the appropriate mode
E. If it??s an upcoming event question, send the query to a text-to-SQL mode
F. If it??s about ticket purchasing, the customer should be redirected to a payment platform.
G. The chatbot should only process payments

Answer: C

NEW QUESTION 17
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 22
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 23
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 24
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 28
.....

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