Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
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 2
A Generative Al Engineer wants their (inetuned LLMs in their prod Databncks workspace available for testing in their dev workspace as well. All of their workspaces are Unity Catalog enabled and they are currently logging their models into the Model Registry in MLflow.
What is the most cost-effective and secure option for the Generative Al Engineer to accomplish their gAi?

A. Use an external model registry which can be accessed from all workspaces
B. Setup a script to export the model from prod and import it to dev.
C. Setup a duplicate training pipeline in dev, so that an identical model is available in dev.
D. Use MLflow to log the model directly into Unity Catalog, and enable READ access in the dev workspace to the model.

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 Al 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 Al 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 Al 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 Al Engineer is building a RAG application that answers questions about internal documents for the company SnoPen AI.
The source documents may contain a significant amount of irrelevant content, such as advertisements, sports news, or entertainment news, or content about other companies.
Which approach is advisable when building a RAG application to achieve this goal of filtering irrelevant information?

A. Keep all articles because the RAG application needs to understand non-company content to avoid answering questions about them.
B. Include in the system prompt that any information it sees will be about SnoPenAI, even if no data filtering is performed.
C. Include in the system prompt that the application is not supposed to answer any questions unrelated to SnoPen Al.
D. Consolidate all SnoPen AI related documents into a single chunk in the vector database.

Answer: C

NEW QUESTION 7

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 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.

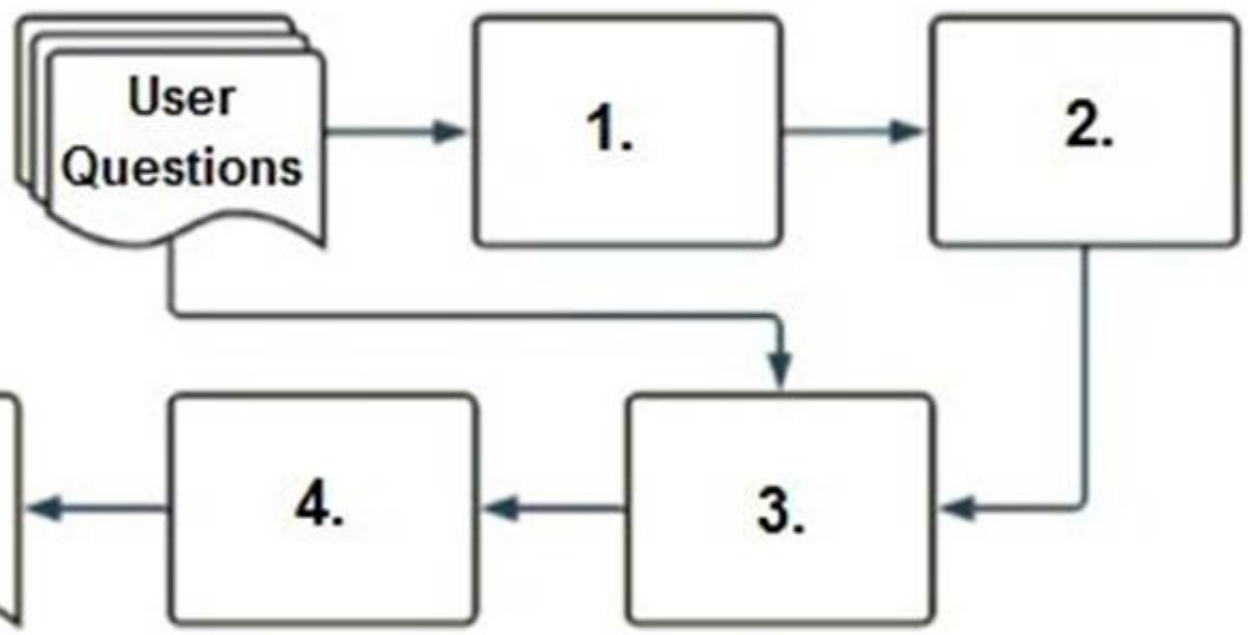
Answer: D

NEW QUESTION 9
A Generative AI Engineer has been asked to build an LLM-based question-answering application. The application should take into account new documents that are frequently published. The engineer wants to build this application with the least cost and least development effort and have it operate at the lowest cost possible.
Which combination of chaining components and configuration meets these requirements?

A. For the application a prompt, a retriever, and an LLM are require
B. The retriever output is inserted into the prompt which is given to the LLM to generate answers.
C. The LLM needs to be frequently with the new documents in order to provide most up-to- date answers.
D. For the question-answering application, prompt engineering and an LLM are required to generate answers.
E. For the application a prompt, an agent and a fine-tuned LLM are require
F. The agent is used by the LLM to retrieve relevant content that is inserted into the prompt which is given to the LLM to generate answers.

Answer: A

NEW QUESTION 10
A company has a typical RAG-enabled, customer-facing chatbot on its website.



Select the correct sequence of components a user's questions will go through before the final output is returned. Use the diagram above for reference.

A. 1.embedding model, 2.vector search, 3.context-augmented prompt, 4.response- generating LLM
B. 1.context-augmented prompt, 2.vector search, 3.embedding model, 4.response- generating LLM
C. 1.response-generating LLM, 2.vector search, 3.context-augmented prompt, 4.embedding model
D. 1.response-generating LLM, 2.context-augmented prompt, 3.vector search, 4.embedding model

Answer: A

NEW QUESTION 10
A Generative AI 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 AI 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 11
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 16
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 21
A Generative AI Engineer is creating an LLM system that will retrieve news articles from the year 1918 and related to a user's query and summarize them. The engineer has noticed that the summaries are generated well but often also include an explanation of how the summary was generated, which is undesirable.
Which change could the Generative AI Engineer perform to mitigate this issue?

A. Split the LLM output by newline characters to truncate away the summarization explanation.
B. Tune the chunk size of news articles or experiment with different embedding models.
C. Revisit their document ingestion logic, ensuring that the news articles are being ingested properly.
D. Provide few shot examples of desired output format to the system and/or user prompt.

Answer: D

NEW QUESTION 25
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 28
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 29
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 31
.....

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