Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

https://www.2passeasy.com/dumps/Databricks-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
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 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 5
A Generative Al Engineer has successfully ingested unstructured documents and chunked them by document sections. They would like to store the chunks in a Vector Search index. The current format of the dataframe has two columns: (i) original document file name (ii) an array of text chunks for each document.
What is the most performant way to store this dataframe?

A. Split the data into train and test set, create a unique identifier for each document, then save to a Delta table
B. Flatten the dataframe to one chunk per row, create a unique identifier for each row, and save to a Delta table
C. First create a unique identifier for each document, then save to a Delta table
D. Store each chunk as an independent JSON file in Unity Catalog Volum
E. For each JSON file, the key is the document section name and the value is the array of text chunks for that section

Answer: B

NEW QUESTION 6
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 Al 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 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 8
A Generative Al Engineer has created a RAG application which can help employees retrieve answers from an internal knowledge base, such as Confluence pages or Google Drive. The prototype application is now working with some positive feedback from internal company testers. Now the Generative Al Engineer wants to formally evaluate the system??s performance and understand where to focus their efforts to further improve the system.
How should the Generative Al Engineer evaluate the system?

A. Use cosine similarity score to comprehensively evaluate the quality of the final generated answers.
B. Curate a dataset that can test the retrieval and generation components of the system separatel
C. Use MLflow??s built in evaluation metrics to perform the evaluation on the retrieval and generation components.
D. Benchmark multiple LLMs with the same data and pick the best LLM for the job.
E. Use an LLM-as-a-judge to evaluate the quality of the final answers generated.

Answer: B

NEW QUESTION 9
A Generative Al Engineer is ready to deploy an LLM application written using Foundation Model APIs. They want to follow security best practices for production scenarios
Which authentication method should they choose?

A. Use an access token belonging to service principals
B. Use a frequently rotated access token belonging to either a workspace user or a service principal
C. Use OAuth machine-to-machine authentication
D. Use an access token belonging to any workspace user

Answer: A

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 Al 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 12
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 15
A Generative Al 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 18
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 20
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 24
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 29
A Generative AI Engineer is tasked with developing a RAG application that will help a small internal group of experts at their company answer specific questions, augmented by an internal knowledge base. They want the best possible quality in the answers, and neither latency nor throughput is a huge concern given that the user group is small and they??re willing to wait for the best answer. The topics are sensitive in nature and the data is highly confidential and so, due to regulatory requirements, none of the information is allowed to be transmitted to third parties.
Which model meets all the Generative AI Engineer??s needs in this situation?

A. Dolly 1.5B
B. OpenAI GPT-4
C. BGE-large
D. Llama2-70B

Answer: C

NEW QUESTION 30
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 32
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 34
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 35
.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Databricks-Generative-AI-Engineer-Associate Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Databricks-Generative-AI-Engineer-Associate Product From:

https://www.2passeasy.com/dumps/Databricks-Generative-AI-Engineer-Associate/

Money Back Guarantee

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