Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
A Generative AI Engineer is creating an agent-based LLM system for their favorite monster truck team. The system can answer text based questions about the monster truck team, lookup event dates via an API call, or query tables on the team??s latest standings.
How could the Generative AI Engineer best design these capabilities into their system?

A. Ingest PDF documents about the monster truck team into a vector store and query it in a RAG architecture.
B. Write a system prompt for the agent listing available tools and bundle it into an agent system that runs a number of calls to solve a query.
C. Instruct the LLM to respond with ??RAG??, ??API??, or ??TABLE?? depending on the query, then use text parsing and conditional statements to resolve the query.
D. Build a system prompt with all possible event dates and table information in the system promp
E. Use a RAG architecture to lookup generic text questions and otherwise leverage the information in the system prompt.

Answer: B

NEW QUESTION 2
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 3
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 4
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 5
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 6
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 7
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 8
Generative Al 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 9
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 10
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 10
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 12
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 15

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 18
A Generative AI 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 23
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 24
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 27
A Generative AI 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 28
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 29
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 34
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 36
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 38
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 39
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 40
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 45
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 50
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 53
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 56
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 61
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 64
.....

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