Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
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 2
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 3
A Generative Al 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 4
A Generative Al 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 5
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 6
A Generative Al 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
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 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 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 11
A Generative Al 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 15
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 20
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

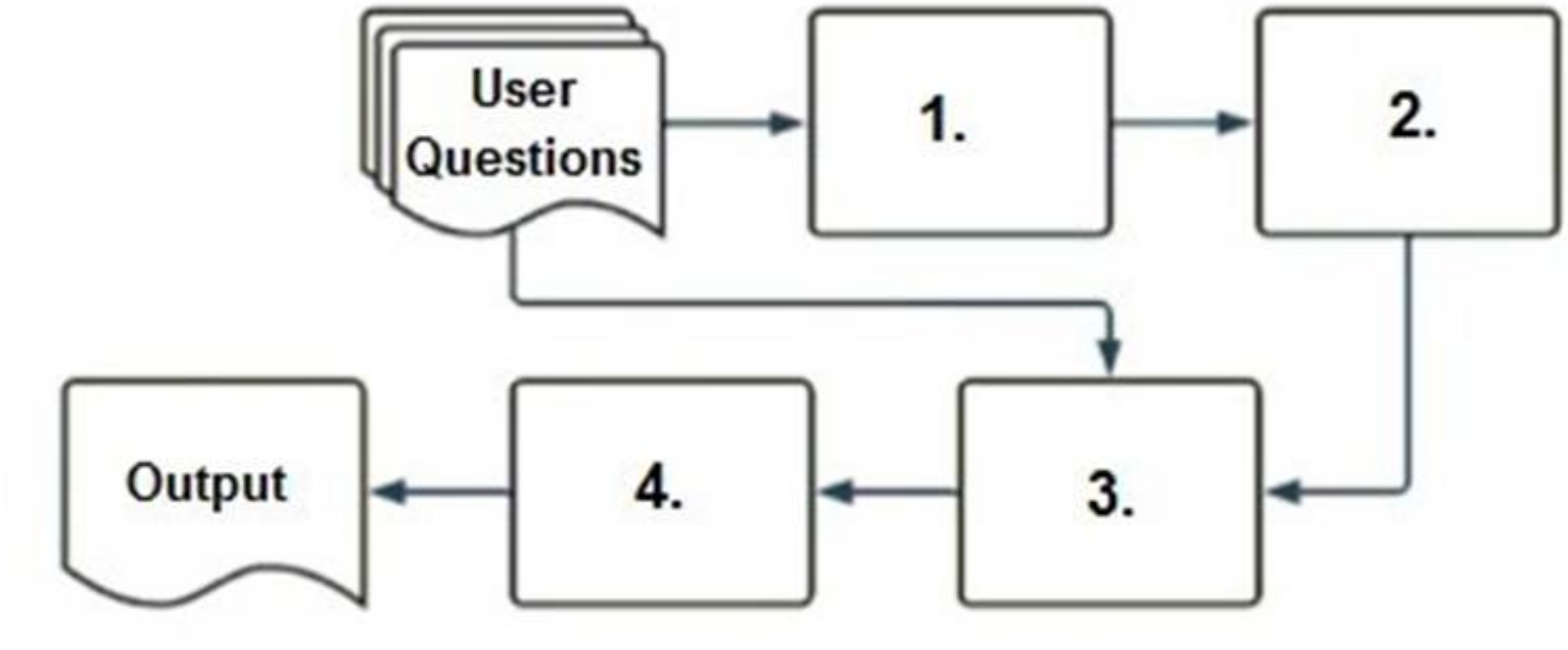
Answer: CD

NEW QUESTION 21
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 23
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 24
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 27
A Generative AI Engineer is tasked with deploying an application that takes advantage of a custom MLflow Pyfunc model to return some interim results.
How should they configure the endpoint to pass the secrets and credentials?

A. Use spark.conf.set ()
B. Pass variables using the Databricks Feature Store API
C. Add credentials using environment variables
D. Pass the secrets in plain text

Answer: C

NEW QUESTION 28
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 33
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 35
A Generative AI Engineer is deciding between using LSH (Locality Sensitive Hashing) and HNSW (Hierarchical Navigable Small World) for indexing their vector database Their top priority is semantic accuracy
Which approach should the Generative AI Engineer use to evaluate these two techniques?

A. Compare the cosine similarities of the embeddings of returned results against those of a representative sample of test inputs
B. Compare the Bilingual Evaluation Understudy (BLEU) scores of returned results for a representative sample of test inputs
C. Compare the Recall-Onented-Understudy for Gistmg Evaluation (ROUGE) scores of returned results for a representative sample of test inputs
D. Compare the Levenshtein distances of returned results against a representative sample of test inputs

Answer: A

NEW QUESTION 36
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 40
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 45
A Generative Al 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 Al Engineer??s needs in this situation?

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

Answer: C

NEW QUESTION 49
A Generative Al Engineer has created a RAG application to look up answers to questions about a series of fantasy novels that are being asked on the author??s web forum. The fantasy novel texts are chunked and embedded into a vector store with metadata (page
number, chapter number, book title), retrieved with the user??s query, and provided to an LLM for response generation. The Generative AI Engineer used their intuition to pick the chunking strategy and associated configurations but now wants to more methodically choose the best values.
Which TWO strategies should the Generative AI Engineer take to optimize their chunking strategy and parameters? (Choose two.)

A. Change embedding models and compare performance.
B. Add a classifier for user queries that predicts which book will best contain the answe
C. Use this to filter retrieval.
D. Choose an appropriate evaluation metric (such as recall or NDCG) and experiment with changes in the chunking strategy, such as splitting chunks by paragraphs or chapter
E. Choose the strategy that gives the best performance metric.
F. Pass known questions and best answers to an LLM and instruct the LLM to provide the best token coun
G. Use a summary statistic (mean, median, etc.) of the best token counts to choose chunk size.
H. Create an LLM-as-a-judge metric to evaluate how well previous questions are answered by the most appropriate chun
I. Optimize the chunking parameters based upon the values of the metric.

Answer: CE

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 is responsible for developing a chatbot to enable their company??s internal HelpDesk Call Center team to more quickly find related tickets and provide resolution. While creating the GenAI application work breakdown tasks for this project, they realize they need to start planningwhich data sources (either Unity Catalog volume or Delta table) they could choose for this application. They have collected several candidate data sources for consideration:
call_rep_history: a Delta table with primary keys representative_id, call_id. This table is maintained to calculate representatives?? call resolution from fields call_duration and call start_time.
transcript Volume: a Unity Catalog Volume of all recordings as a *.wav files, but also a text transcript as *.txt files.
call_cust_history: a Delta table with primary keys customer_id, cal1_id. This table is maintained to calculate how much internal customers use the HelpDesk to make sure that the charge back model is consistent with actual service use.
call_detail: a Delta table that includes a snapshot of all call details updated hourly. It includes root_cause and resolution fields, but those fields may be empty for calls that are still active.
maintenance_schedule – a Delta table that includes a listing of both HelpDesk application outages as well as planned upcoming maintenance downtimes.
They need sources that could add context to best identify ticket root cause and resolution. Which TWO sources do that? (Choose two.)

A. call_cust_history
B. maintenance_schedule
C. call_rep_history
D. call_detail
E. transcript Volume

Answer: DE

NEW QUESTION 54
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 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 59
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 60
.....

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