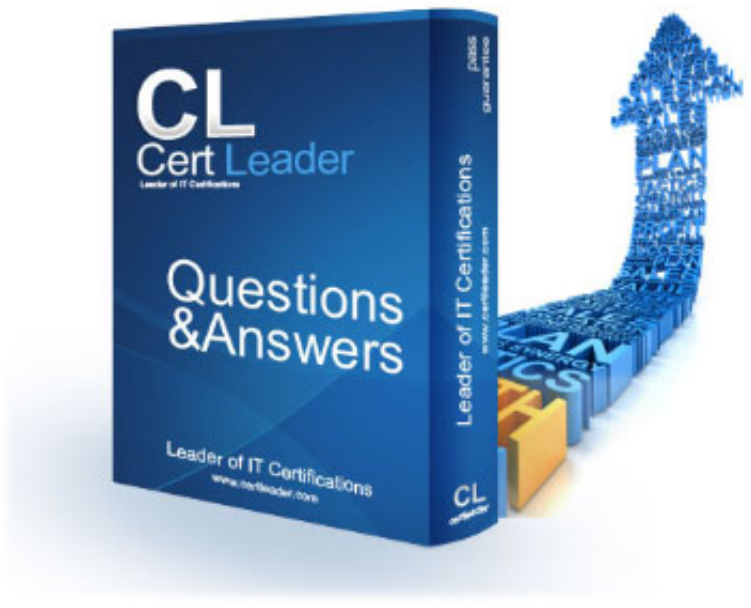
Databricks-Generative-AI-Engineer-Associate Dumps

Databricks Certified Generative AI Engineer Associate

https://www.certleader.com/Databricks-Generative-AI-Engineer-Associate-dumps.html



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
A Generative Al Engineer at an automotive company would like to build a question- answering chatbot for customers to inquire about their vehicles. They have a database containing various documents of different vehicle makes, their hardware parts, and common maintenance information.
Which of the following components will NOT be useful in building such a chatbot?

A. Response-generating LLM
B. Invite users to submit long, rather than concise, questions
C. Vector database
D. Embedding model

Answer: B

NEW QUESTION 3
A Generative AI 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 4
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 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 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 7
A Generative Al Engineer is building an LLM-based application that has an important transcription (speech-to-text) task. Speed is essential for the success of the application
Which open Generative Al models should be used?

A. L!ama-2-70b-chat-hf
B. MPT-30B-Instruct
C. DBRX
D. whisper-large-v3 (1.6B)

Answer: D

NEW QUESTION 8
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 9
A team wants to serve a code generation model as an assistant for their software developers. It should support multiple programming languages. Quality is the primary objective.
Which of the Databricks Foundation Model APIs, or models available in the Marketplace, would be the best fit?

A. Llama2-70b
B. BGE-large
C. MPT-7b
D. CodeLlama-34B

Answer: D

NEW QUESTION 10
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 10
A Generative Al Engineer is setting up a Databricks Vector Search that will lookup news articles by topic within 10 days of the date specified An example query might be "Tell me about monster truck news around January 5th 1992". They want to do this with the least amount of effort.
How can they set up their Vector Search index to support this use case?

A. Split articles by 10 day blocks and return the block closest to the query.
B. Include metadata columns for article date and topic to support metadata filtering.
C. pass the query directly to the vector search index and return the best articles.
D. Create separate indexes by topic and add a classifier model to appropriately pick the best index.

Answer: B

NEW QUESTION 15
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 19
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

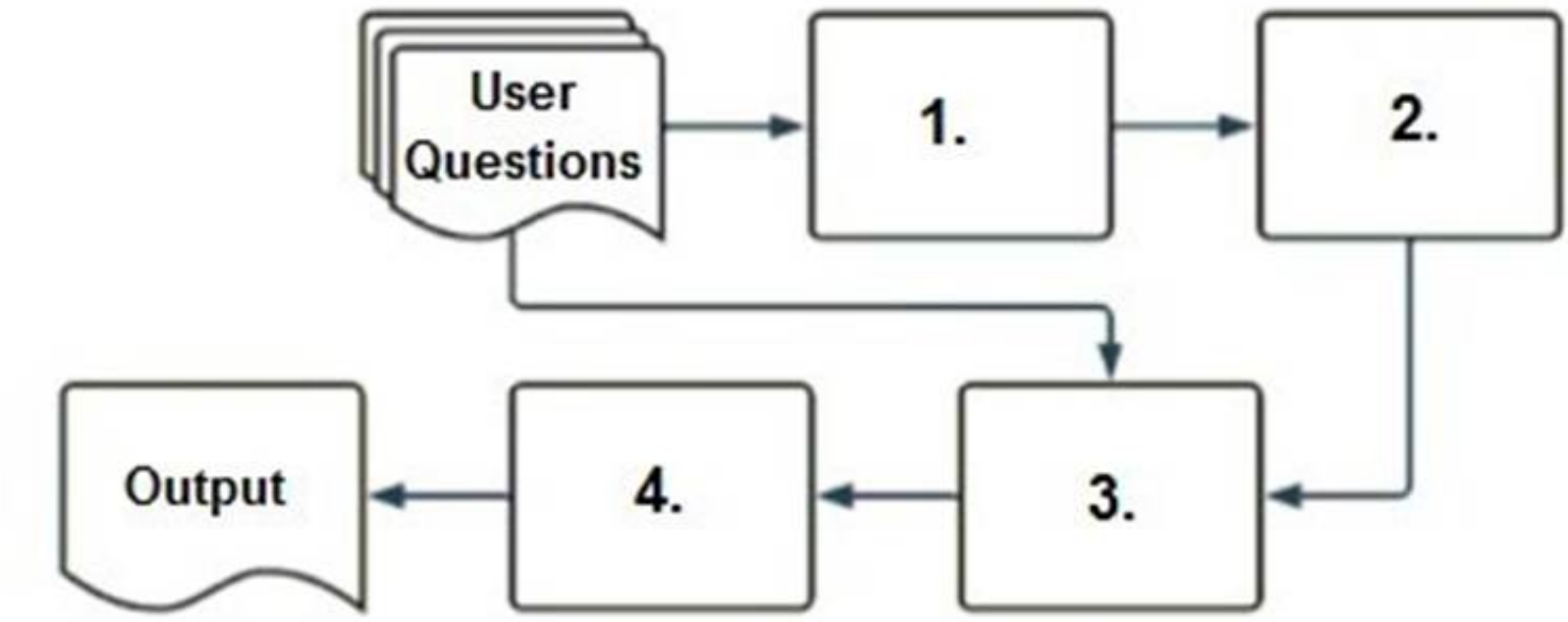
Answer: B

NEW QUESTION 24
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 27
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 30
A Generative Al 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 Al 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 31
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 32
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 35
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 41
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 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 43
A Generative AI 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 48
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 50
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 52
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 57
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
.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your Databricks-Generative-AI-Engineer-Associate Exam with Our Prep Materials Via below:

https://www.certleader.com/Databricks-Generative-AI-Engineer-Associate-dumps.html