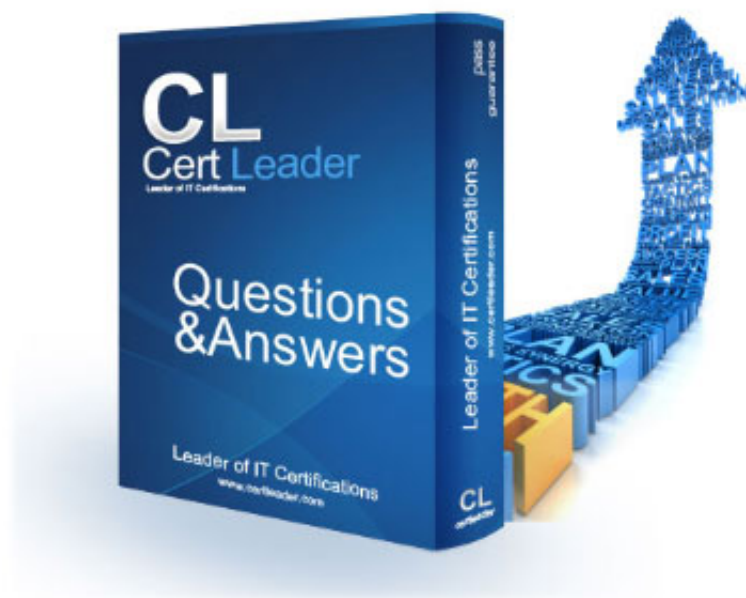


Databricks-Generative-AI-Engineer-Associate Dumps

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

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 AI 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 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 store
- 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 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 AI.
- D. Consolidate all SnoPen AI related documents into a single chunk in the vector database.

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

Generative AI 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 8

A Generative AI 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 9

A Generative AI 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 analysis
- C 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 AI 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 AI Engineer is confident they have the correct list of possible labels, but the output frequently contains additional reasoning in the answer when the Generative AI 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 shot 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 AI 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 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 17

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 21

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 23

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 28

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 32

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 37

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 41

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-Oriented-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 44

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 49

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 53

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 55

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 planning which 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 *.wav files, but also a text transcript as *.txt files.
- call_cust_history: a Delta table with primary keys customer_id, call_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 60

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 64

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 68

.....

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>