

Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate



NEW QUESTION 1

A Generative AI Engineer is designing a RAG application for answering user questions on technical regulations as they learn a new sport. What are the steps needed to build this RAG application and deploy it?

- A. Ingest documents from a source → Index the documents and saves to Vector Search → User submits queries against an LLM → LLM retrieves relevant documents → Evaluate model → LLM generates a response → Deploy it using Model Serving
- B. Ingest documents from a source → Index the documents and save to Vector Search → User submits queries against an LLM → LLM retrieves relevant documents → LLM generates a response → Evaluate model → Deploy it using Model Serving
- C. Ingest documents from a source → Index the documents and save to Vector Search → Evaluate model → Deploy it using Model Serving
- D. User submits queries against an LLM → Ingest documents from a source → Index the documents and save to Vector Search → LLM retrieves relevant documents → LLM generates a response → Evaluate model → Deploy it using Model Serving

Answer: B

NEW QUESTION 2

What is the most suitable library for building a multi-step LLM-based workflow?

- A. Pandas
- B. TensorFlow
- C. PySpark
- D. LangChain

Answer: D

NEW QUESTION 3

A Generative AI Engineer is tasked with improving the RAG quality by addressing its inflammatory outputs. Which action would be most effective in mitigating the problem of offensive text outputs?

- A. Increase the frequency of upstream data updates
- B. Inform the user of the expected RAG behavior
- C. Restrict access to the data sources to a limited number of users
- D. Curate upstream data properly that includes manual review before it is fed into the RAG system

Answer: D

NEW QUESTION 4

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 prompt
- E. Use a RAG architecture to lookup generic text questions and otherwise leverage the information in the system prompt.

Answer: B

NEW QUESTION 5

A Generative AI 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 AI 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 6

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 7

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 8

A Generative AI 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 Volumn
- 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 9

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 10

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. Train 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 10

A Generative AI Engineer is creating an LLM-based application. The documents for its retriever have been chunked to a maximum of 512 tokens each. The Generative AI 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 14

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 19

A Generative AI 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 23

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 24

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 28

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 33

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 35

A Generative AI Engineer is designing an LLM-powered live sports commentary platform. The platform provides real-time updates and LLM-generated analyses for any users who would like to have live summaries, rather than reading a series of potentially outdated news articles. Which tool below will give the platform access to real-time data for generating game analyses based on the latest game scores?

- A. DatabricksIQ
- B. Foundation Model APIs
- C. Feature Serving
- D. AutoML

Answer: C

NEW QUESTION 38

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 42

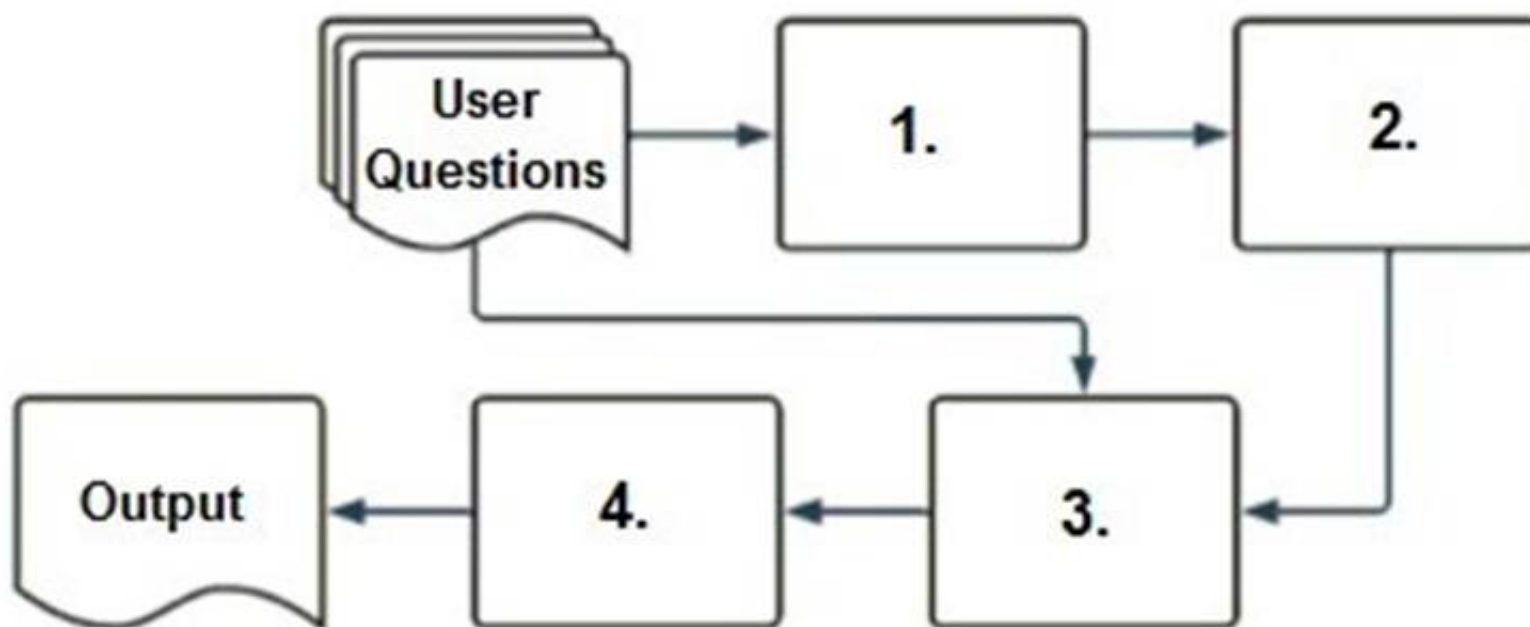
A Generative AI Engineer has developed an LLM application to answer questions about internal company policies. The Generative AI Engineer must ensure that the application doesn't hallucinate or leak confidential data. Which approach should NOT be used to mitigate hallucination or confidential data leakage?

- 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 45

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 50

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 53

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 56

When developing an LLM application, it's crucial to ensure that the data used for training the model complies with licensing requirements to avoid legal risks.

Which action is NOT appropriate to avoid legal risks?

- A. Reach out to the data curators directly before you have started using the trained model to let them know.
- B. Use any available data you personally created which is completely original and you can decide what license to use.
- C. Only use data explicitly labeled with an open license and ensure the license terms are followed.
- D. Reach out to the data curators directly after you have started using the trained model to let them know.

Answer: D

NEW QUESTION 61

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 64

A Generative AI Engineer is creating an LLM system that will retrieve news articles from the year 1918 and related to a user's query and summarize them. The engineer has noticed that the summaries are generated well but often also include an explanation of how the summary was generated, which is undesirable.

Which change could the Generative AI Engineer perform to mitigate this issue?

- A. Split the LLM output by newline characters to truncate away the summarization explanation.
- B. Tune the chunk size of news articles or experiment with different embedding models.
- C. Revisit their document ingestion logic, ensuring that the news articles are being ingested properly.
- D. Provide few shot examples of desired output format to the system and/or user prompt.

Answer: D

NEW QUESTION 69

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 a *.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 74

.....

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

100% Actual & Verified — Instant Download, Please Click
[Order The Databricks-Generative-AI-Engineer-Associate Practice Test Here](#)