



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 save 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 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 3

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 4

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 5

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 6

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 7

A Generative AI 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 AI 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 AI 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 separately
- 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 8

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 9

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 availability
- 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 availability
- F. Respond with "In Stock" if the product is available or "Out of Stock" if not.

Answer: B

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

Which indicator should be considered to evaluate the safety of the LLM outputs when qualitatively assessing LLM responses for a translation use case?

- A. The ability to generate responses in code
- B. The similarity to the previous language
- C. The latency of the response and the length of text generated
- D. The accuracy and relevance of the responses

Answer: D

NEW QUESTION 13

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 15

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 scope
- 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 18

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 21

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 browsing
- B. Let me know if you have questions after reading them.
- C. Please call your local emergency services.
- D. Headaches can be tough
- 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 22

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 Gisting 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 26

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 30

A Generative AI Engineer is helping a cinema extend its website's chatbot 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 table
- 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 implementation
- D. Implement a query that reads from the Delta table and uses a text-to-SQL LLM to generate the SQL query
- 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 32

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 34

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions 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](#)