

Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate



NEW QUESTION 1

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 2

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 3

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 SnoPen AI, 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 4

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 Volume
- 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 5

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 6

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 required
- 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 updated 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 required
- 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 7

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 8

A Generative AI 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 9

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 10

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..." }
```

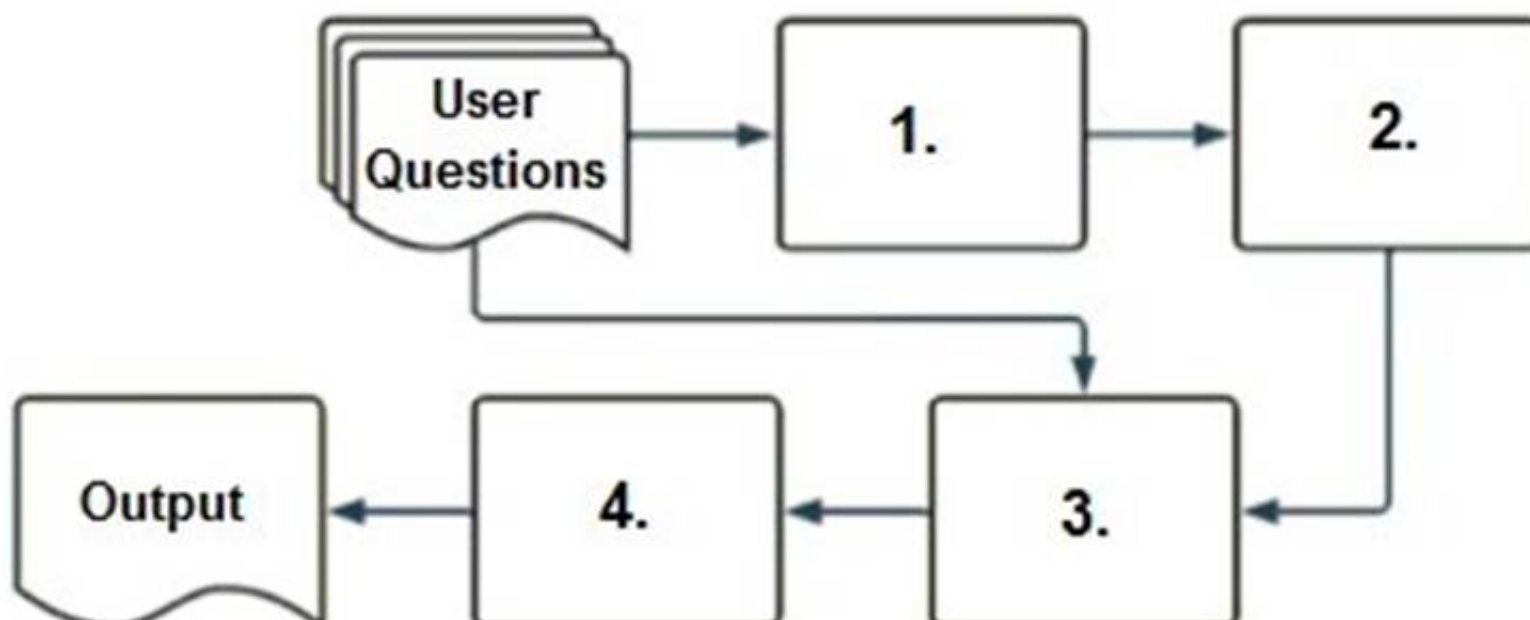
What TWO solutions should the Generative AI Engineer implement without changing the response generating model? (Choose two.)

- 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 10

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 12

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 15

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 17

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 18

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 21

.....

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](#)