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
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 Al Engineer is developing a RAG system for their company to perform internal document Q&A for structured HR policies, but the answers returned are frequently incomplete and unstructured It seems that the retriever is not returning all relevant context The Generative Al Engineer has experimented with different embedding and response generating LLMs but that did not improve results.
Which TWO options could be used to improve the response quality? Choose 2 answers

A. Add the section header as a prefix to chunks
B. Increase the document chunk size
C. Split the document by sentence
D. Use a larger embedding model
E. Fine tune the response generation model

Answer: AB

NEW QUESTION 4
A Generative Al 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 5
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 stor
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 6
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 7
A Generative Al 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 Al Engineer is creating an LLM-based application. The documents for its retriever have been chunked to a maximum of 512 tokens each. The Generative Al 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 9
A Generative Al 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 analysisC 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 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 10
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 Al 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 separatel
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 14
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 16
A Generative Al 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 17
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 18
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:

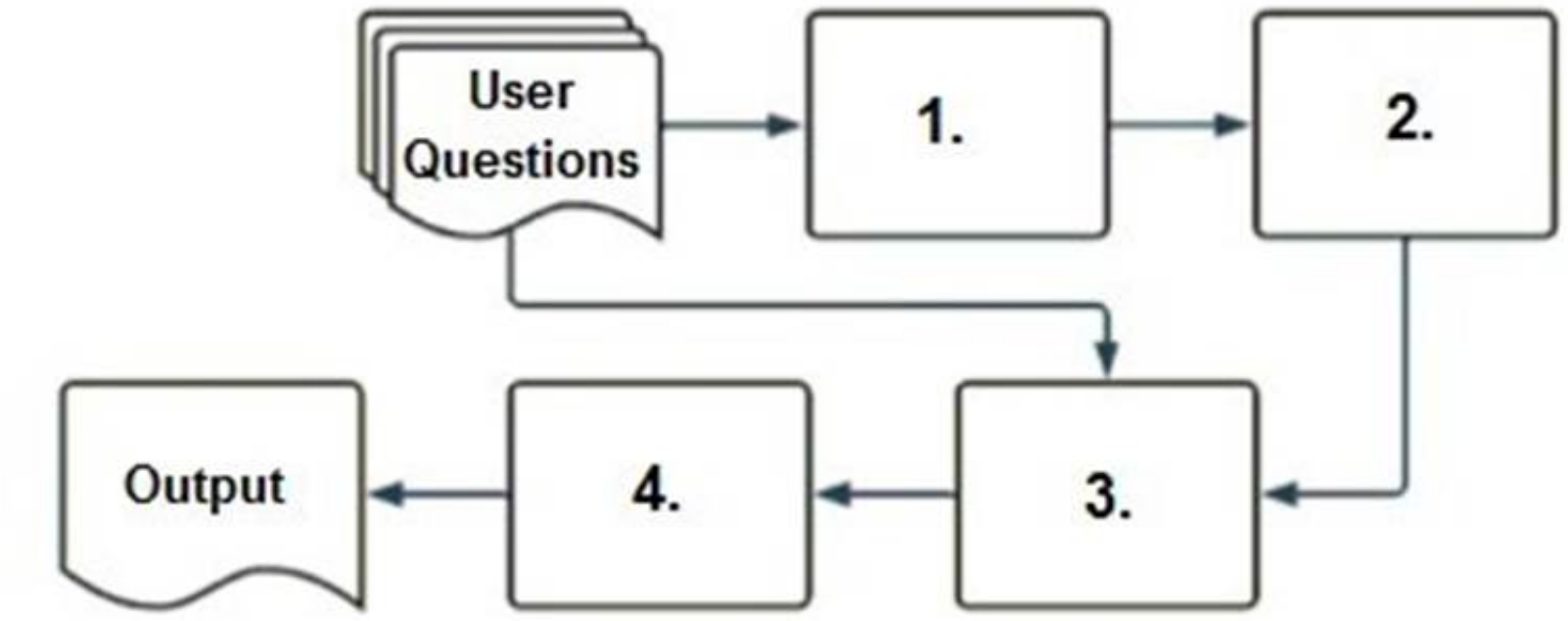
```
{"error_code": "BAD_REQUEST", "message": "Bad request: rpc error:
code = InvalidArgument desc = prompt token count (4595) cannot
exceed 4096…"}
```

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 21
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 24
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 28
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 29
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 33
.....

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