Databricks

Exam Questions Databricks-Generative-AI-Engineer-Associate

Databricks Certified Generative AI Engineer Associate

NEW QUESTION 1
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 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 tasked with developing an application that is based on an open source large language model (LLM). They need a foundation LLM with a large context window.
Which model fits this need?

A. DistilBERT
B. MPT-30B
C. Llama2-70B
D. DBRX

Answer: C

NEW QUESTION 4
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 5
A Generative Al 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 6
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 7
A Generative Al 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 8
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 9
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 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 10
A Generative Al Engineer has developed an LLM application to answer questions about internal company policies. The Generative Al 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 13
A small and cost-conscious startup in the cancer research field wants to build a RAG application using Foundation Model APIs.
Which strategy would allow the startup to build a good-quality RAG application while being cost-conscious and able to cater to customer needs?

A. Limit the number of relevant documents available for the RAG application to retrieve from
B. Pick a smaller LLM that is domain-specific
C. Limit the number of queries a customer can send per day
D. Use the largest LLM possible because that gives the best performance for any general queries

Answer: B

NEW QUESTION 17
A Generative Al 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 Al 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 22
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 23
What is an effective method to preprocess prompts using custom code before sending them to an LLM?

A. Directly modify the LLM??s internal architecture to include preprocessing steps
B. It is better not to introduce custom code to preprocess prompts as the LLM has not been trained with examples of the preprocessed prompts
C. Rather than preprocessing prompts, it??s more effective to postprocess the LLM outputs to align the outputs to desired outcomes
D. Write a MLflow PyFunc model that has a separate function to process the prompts

Answer: D

NEW QUESTION 28
A Generative Al 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 Al 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-Onented-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 29
A Generative Al Engineer is helping a cinema extend its website's chat bot 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 tabl
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
D. implementatio
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 33
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 34
.....

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