DBCLOUD
EXAMS

QUESTIONS & ANSWERS

verified



Technology Provider: Databricks

Exam: Certified Generative AI Engineer Associate

Number of questions in the database: 88

Exam Version: NOV.2026

Question #1Topic 1

A Generative Al 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 answer. Use this to filter retrieval.

C. 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 chapters. Choose the strategy that gives the best performance metric.

D. Pass known questions and best answers to an LLM and instruct the LLM to provide the best token count. Use a summary statistic (mean, median, etc.) of the best token counts to choose chunk size.

E. Create an LLM-as-a-judge metric to evaluate how well previous questions are answered by the most appropriate chunk. Optimize the chunking parameters based upon the values of the metric.

Correct Answer: CE

Question #2Topic 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?

https://dbcloud.academy/

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

Correct Answer: B

Question #3Topic 1

A Generative AI Engineer just deployed an LLM application at a digital marketing company that assists with answering customer service inquiries.

Which metric should they monitor for their customer service LLM application in production?

A. Number of customer inquiries processed per unit of time

B. Energy usage per query

C. Final perplexity scores for the training of the model

D. HuggingFace Leaderboard values for the base LLM

Correct Answer: A

Question #4Topic 1

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?

https://dbcloud.academy/

A. Create a tool for finding available team members given project dates. Embed all project scopes into a vector store, perform a retrieval using team member profiles to find the best team member.

B. Create a tool for finding team member availability given project dates, and another tool that uses an LLM to extract keywords from project scopes. Iterate through available team members' profiles and perform keyword matching to find the best available team member.

C. Create a tool to find available team members given project dates. Create a second tool that can calculate a similarity score for a combination of team member profile and the project scope. Iterate through the team members and rank by best score to select a team member.

D. Create a tool for finding available team members given project dates. 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.

Correct Answer: D

Question #5Topic 1

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

Correct Answer: C

Question #6Topic 1

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?

https://dbcloud.academy/