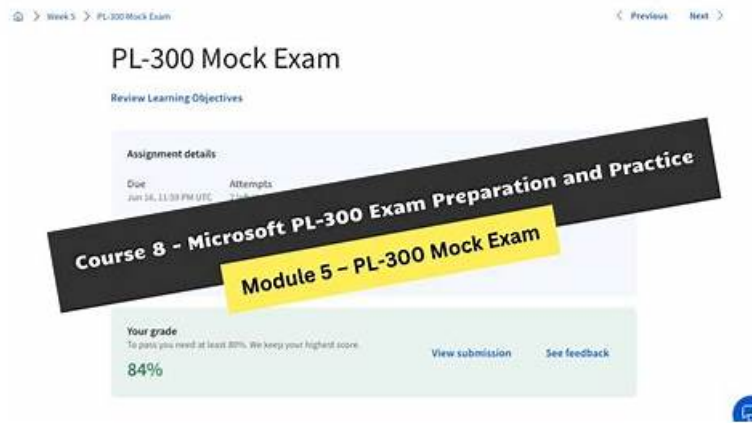


New AI-300 Exam Format - AI-300 Latest Mock Test



With the simulation function, our AI-300 training guide is easier to understand and have more vivid explanations to help you learn more knowledge. You can set time to test your study efficiency, so that you can accomplish your test within the given time when you are in the Real AI-300 Exam. Besides, you can get the real feeling of taking part in the real exam for our AI-300 exam questions have the function of simulating the real exam. So that you can have a better performance when you attend the real exam.

The second format of Operationalizing Machine Learning and Generative AI Solutions (AI-300) is the web-based practice exam that can be taken online through browsers like Firefox, Chrome, Safari, MS Edge, Internet Explorer, and Microsoft Edge. You don't need to install any excessive plugins or Software to attempt the web-based Practice AI-300 Exam. All operating systems also support the web-based practice exam.

>> New AI-300 Exam Format <<

Unparalleled New AI-300 Exam Format & Passing AI-300 Exam is No More a Challenging Task

The simulation of the actual AI-300 test helps you feel the real AI-300 exam scenario, so you don't face anxiety while giving the final examination. You can even access your last test results, which help to realize your mistakes and try to avoid them while taking the Operationalizing Machine Learning and Generative AI Solutions (AI-300) certification test.

Microsoft Operationalizing Machine Learning and Generative AI Solutions Sample Questions (Q46-Q51):

NEW QUESTION # 46

Case Study 1 - Fabrikam Inc.

Background

Fabrikam Inc. is a mid-sized healthcare analytics company that provides population health dashboards and predictive insights to regional hospital systems across the United States.

Fabrikam Inc. customers rely on near real time analytics to monitor patient flow, staffing needs, and readmission risks. They use multiple traditional forecasting machine learning models for predictions.

Fabrikam Inc. has an established Microsoft Azure footprint. The company uses Jupyter Notebooks that run on a local server as the primary development environment. The data science team is experiencing scalability, asset management and code management issues with the current development platform. Fabrikam Inc. plans to migrate to a cloud-based development environment to mitigate the issues.

Additionally, the company plans to implement a Retrieval-Augmented Generation (RAG)-based chat application for client support. Leadership requires the application to be developed and deployed with a low operational risk.

Current Environment

Fabrikam Inc. operates a single Azure subscription that has the following components:

- * Azure Data Lake Storage Gen2 that contains de-identified clinical and operational datasets
- * Azure AI Search indexing curated analytical documents and reference materials
- * A small set of Python-based training scripts maintained by data scientists
- * Azure OpenAI Service with deployed foundational models

* A Microsoft Foundry resource for building a RAG-based solution

Evaluation data has manually defined expected responses.

The current challenges faced by the data science team include the following:

* Model training jobs are run manually from notebooks.

* Experiment tracking is inconsistent

* Model versions are registered without standardized metadata.

* Deployment is performed manually by data scientists, with limited rollback capability.

* The team has no standardized evaluation process for generative AI outputs.

The environment currently allows public network access. Authentication relies on user accounts rather than managed identities.

Compute targets are manually created and shared across experiments. This has led to resource contention during peak usage.

Business Requirements

Fabrikam Inc. has the following business requirements for the modernization initiative:

* Provide a conversational interface that answers analytics questions by using internal documents and datasets.

* Ensure that sensitive healthcare-related data is not exposed outside the Fabrikam Inc. Azure tenant.

* Enable repeatable and auditable model training and deployment processes.

* Support experimentation to compare prompt strategies and fine-tuned models.

* Align the model with the ranked preferences and optimize behavior for the long term.

* Minimize disruption to existing analytics workloads during rollout.

Technical Requirements

To support the business goals, Fabrikam Inc. identifies these technical requirements:

* Use Azure Machine Learning workspaces to centrally manage data assets, models, and environments.

* Implement experiment tracking and model versioning for all training jobs.

* Orchestrate training and evaluation by using pipelines rather than manually running notebooks.

* Deploy traditional machine learning models with support for staged rollout and rollback.

* Improve RAG-based solution output quality.

* Use the existing evaluation datasets that are based on real data with input-output pairs.

* Apply advanced fine-tuning techniques only when prompt engineering is insufficient

Issues and Constraints Fabrikam Inc. must comply with internal security policies that require the company to restrict network access and avoid long-lived secrets. The data science team has limited Azure DevOps experience, so solutions must favor managed services and automation over custom infrastructure.

Cost predictability is important. Leadership prefers serverless or managed compute options where possible but is willing to approve dedicated compute for stable production workloads.

Problem Statement

Fabrikam Inc. must design and implement an Azure-based AI operations solution that enables reliable training, evaluation, deployment, and iteration of generative AI models. The solution must support experimentation and gradual rollout while ensuring governance, security, and operational stability. The data science and platform teams must collaborate to deliver this solution by using Azure Machine Learning and Microsoft Foundry capabilities.

You need to isolate training workloads while remaining cost-aware to address Fabrikam Inc.'s issues, constraints, and technical requirements. What should you implement?

- A. Fixed-size compute cluster
- **B. Managed compute targets with autoscaling**
- C. Dedicated compute clusters per experiment
- D. Training jobs that run on a single shared compute cluster

Answer: B

Explanation:

Scenario: Issues and Constraints: Cost predictability is important. Leadership prefers serverless or managed compute options where possible but is willing to approve dedicated compute for stable production workloads.

Managed compute targets with autoscaling are the best choice for Azure Machine Learning training workloads when serverless or managed options are preferred and cost predictability is critical.

Best Implementation: Managed Compute with Autoscaling

This option, specifically using Azure Machine Learning compute clusters (AmlCompute), aligns with all your requirements:

Managed Infrastructure: Azure handles the creation, patching, and lifecycle of the virtual machines, reducing management overhead.

Cost Predictability & Efficiency: Autoscaling allows you to set a minimum of zero nodes. This ensures you only pay for compute while a job is running, preventing costs from idle resources.

Scalability: It can automatically scale up to a multi-node cluster to handle large datasets or distributed training jobs.

Enterprise Governance: Administrators can enforce cost control by setting quotas at the subscription or workspace level.

Reference:

<https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-serverless-compute>

NEW QUESTION # 47

A Retrieval-Augmented Generation (RAG) solution returns incomplete answers because relevant content is inconsistently retrieved from the knowledge source.

You need to improve RAG accuracy without changing the embedding model currently in use. You need to achieve this goal while minimizing operational costs.

Which two actions should you perform? Each correct answer presents part of the solution.

Choose two.

NOTE: Each correct selection is worth one point.

- A. Optimize the length of embedding vectors.
- **B. Tune chunk size and overlap to match content structure.**
- C. Increase token limits for all requests.
- **D. Implement an optimized re-ranker.**

Answer: B,D

Explanation:

To improve Retrieval-Augmented Generation (RAG) accuracy, address inconsistent retrieval, and eliminate incomplete answers without changing the embedding model or increasing costs significantly, you must move beyond naive fixed-length chunking and implement a two-stage retrieval process.

Here is the targeted, low-cost strategy:

1. Tune Chunk Size and Overlap to Match Content Structure

Inconsistent retrieval often occurs because important information is split across chunk boundaries (breaking context) or chunks are too large, diluting the semantic meaning.

2. Implement an Optimized Re-ranker

The initial vector search often returns "noise"-chunks that are semantically close but not actually relevant. A re-ranker acts as a second, smarter, but more "expensive" step that works on a smaller subset of data, making it low-cost overall.

Reference:

<https://medium.com/@sthanikamsanthosh1994/how-to-improve-rag-retrieval-augmented-generation-performance-2a42303117f8>

NEW QUESTION # 48

A pipeline step fails intermittently due to transient compute issues. You need to improve reliability without modifying core logic or increasing cost significantly. What is the BEST approach?

- **A. Enable retry policy on pipeline step**
- B. Remove failing step
- C. Run pipeline manually
- D. Increase compute size

Answer: A

Explanation:

Retry policies allow pipeline steps to automatically recover from transient failures, such as temporary compute or network issues. This improves reliability without modifying core logic or increasing infrastructure costs. Increasing compute resources does not address transient failure scenarios effectively.

NEW QUESTION # 49

You manage an Azure Machine Learning workspace. You develop a machine learning model.

You must deploy the model to use a low-priority VM with a pricing discount.

You need to deploy the model.

Which compute target should you use?

- **A. Azure Machine Learning compute clusters**
- B. Azure Kubernetes Service (AKS)
- C. Local deployment
- D. Azure Container Instances (ACI)

Answer: A

Explanation:

The best compute target for deploying a model using low-priority VMs (or their modern successor, Spot VMs) is an Azure Machine Learning compute cluster.

Best Compute Target: AML Compute Cluster

For low-priority/Spot pricing, you should use an Azure Machine Learning compute cluster configured with the LowPriority tier.

Primary Use Case: This target is specifically recommended for batch deployments. Batch inference is ideal for low-priority VMs because these jobs are asynchronous and can tolerate the interruptions (preemptions) inherent to discounted capacity.

Pricing Advantage: Low-priority VMs offer significant discounts-often up to 80% off standard rates-by utilizing unused Azure capacity.

Automatic Handling: When a node is preempted during a batch job, Azure Machine Learning automatically attempts to replace the lost capacity and re-queues failed tasks to the cluster.

Reference:

<https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-low-priority-batch>

NEW QUESTION # 50

A team is experimenting with traditional models for a classification workflow in Azure Machine Learning.

The team requires a consistent way to manage assets that are created during experimentation.

You need to ensure that artifacts can be reused and governed across projects.

Which asset should you register?

- A. Component
- B. Environment
- C. Pipeline
- **D. Model**

Answer: D

Explanation:

In an Azure Machine Learning classification workflow, you should register Models.

Registration creates a versioned asset in your workspace or a centralized registry, which is essential for ensuring that artifacts are reusable, governed, and trackable across different projects and environments.

Key Assets for Reuse and Governance

To maintain a consistent and governed workflow, you should focus on registering these specific assets:

Models: The primary artifact. Registering a model allows you to track its lineage (which experiment created it), version it, and deploy it consistently across environments.

Components: These are self-contained pieces of code that perform specific steps in a pipeline (e.g., data cleaning, training).

Registering them allows different teams to reuse the same

"traditional" classification logic without rewriting code.

Environments: Encapsulates the software dependencies (Python packages, Docker images) required for your model to run.

Registering these ensures reproducibility across different compute targets.

Data Assets: Registering your training and testing datasets as versioned assets ensures that you can always audit exactly what data was used to train a specific model version.

Reference:

<https://learn.microsoft.com/en-us/azure/machine-learning/concept-azure-machine-learning-v2>

NEW QUESTION # 51

.....

Microsoft AI-300 can ensure your success. So here comes Microsoft, who provides you with the Microsoft AI-300 exam dumps to get your dream Microsoft AI-300 certification with no hassle. Microsoft AI-300 Certification will add up to your excellence in your field and leave no space for any doubts in the mind of the hiring team.

AI-300 Latest Mock Test: <https://www.vceprep.com/AI-300-latest-vce-prep.html>

Microsoft New AI-300 Exam Format Whether your cellphone is Android system or Apple system, they all can download the App version, After you buy the AI-300 latest training material, you can get a year free updates, Most people regard Microsoft certification as a threshold in this industry, therefore, for your convenience, we are fully equipped with a professional team with

