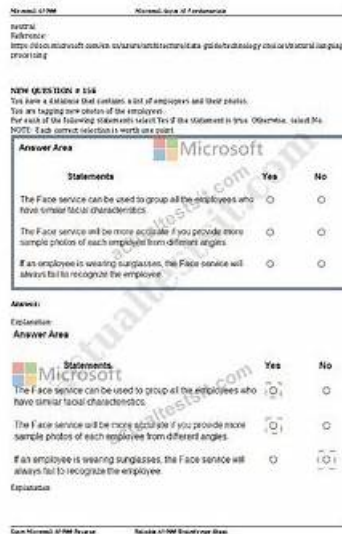


# Quiz 2026 Unparalleled Microsoft New AI-300 Braindumps Ebook



P.S. Free & New AI-300 dumps are available on Google Drive shared by ActualCollection: [https://drive.google.com/open?id=1fQtUs1X4ofTEoTR2UwnNhLCzYoimm0q\\_](https://drive.google.com/open?id=1fQtUs1X4ofTEoTR2UwnNhLCzYoimm0q_)

Our experts generalize the knowledge of the exam into our AI-300 exam materials showing in three versions. PDF version of AI-300 study questions - support customers' printing request, and allow you to have a print and practice in papers. Software version of AI-300 learning guide - supporting simulation test system. App/online version of mock quiz - Being suitable to all kinds of equipment or digital devices, and you can review history and performance better. And you can choose the favorite one.

Since different people have different preferences, we have prepared three kinds of different versions of our AI-300 practice test: PDF, Online App and software. Last but not least, our customers can accumulate exam experience as well as improving their exam skills in the mock exam. And your success is 100 guaranteed for our pass rate of AI-300 Exam Questions is as high as 99% to 100%. And We have put substantial amount of money and effort into upgrading the quality of our AI-300 Exam Preparation materials.

>> New AI-300 Braindumps Ebook <<

**Quiz New AI-300 Braindumps Ebook - Realistic Valid Operationalizing  
Machine Learning and Generative AI Solutions Test Syllabus**

Another great format of our AI-300 exam dumps is the real questions in a PDF file. This is a portable file that contains the most probable AI-300 test questions. The Microsoft AI-300 Pdf Dumps format is a convenient preparation method as these AI-300 questions document is printable and portable.

## Microsoft Operationalizing Machine Learning and Generative AI Solutions Sample Questions (Q70-Q75):

### NEW QUESTION # 70

A team manages an Azure Machine Learning workspace where they deploy models to online endpoints. The team needs to introduce a new version of a model to production without disrupting existing users. The team must validate the new version before full rollout. You need to reduce risk during deployment. What should you do?

- A. Split traffic between deployments.
- B. Route all traffic to the new deployment.
- C. Replace the existing endpoint.
- D. Deploy the model to a batch endpoint.

**Answer: A**

Explanation:

To introduce a new model version in Azure Machine Learning without service interruption, you should use Blue/Green Deployment with Traffic Splitting.

This strategy allows you to run two versions of a model simultaneously under a single Online Endpoint, gradually shifting users to the new version once it is validated.

Key Benefits

Zero Downtime: The endpoint URL stays the same; only the backend routing changes.

Easy Rollback: If the new model fails, you can instantly flip traffic back to 100% on the old version.

Risk Mitigation: Only a small subset of users is exposed to the unproven model initially.

Implementation Steps

1. Create the "Green" Deployment

Deploy the new model version as a second deployment under the existing online endpoint.

Initially, set its traffic allocation to 0%.

2. Canary Testing (Initial Split)

Shift a small percentage of traffic (e.g., 10%) to the new deployment. Monitor performance metrics, error rates, and model accuracy in a real-world environment.

3. Validation & Monitoring

Use Azure Monitor and Application Insights to compare the two deployments. Check for:

Latency: Is the new model slower?

HTTP Status Codes: Are there 4xx or 5xx errors?

Model Drift: Is the prediction quality as expected?

4. Full Rollout

If the new version is stable, increase the traffic split (e.g., 50/50) until the new model handles 100% of the traffic.

5. Cleanup

Once the "Green" deployment is confirmed as the new production standard, you can delete the old ("Blue") deployment to save costs.

Reference:

<https://learn.microsoft.com/en-us/azure/well-architected/ai/operations>

### NEW QUESTION # 71

A company plans to deploy a foundation model in Microsoft Foundry.

The mode must support the following workloads:

A customer support workload used across multiple regions

A marketing workload that must remain within a specific region due to data residency requirements You need to select the deployment type.

Which deployment type should you use for each workload? To answer, move the appropriate deployment types to the correct requirements. You may use each deployment type once, more than once, or not at all. You may need to move the split bar between

panes or scroll to view content . NOTE: Each correct selection is worth one point.

**Deployment types**

Standard

Global Standard

Data Zone Standard

Data Zone Batch

**Workload deployment types**

**Workload**

Customer support

Marketing

**Deployment type**

**Answer:**

Explanation:

**Deployment types**

Standard

Global Standard

Data Zone Standard

Data Zone Batch

**Workload deployment types**

**Workload**

Customer support

Marketing

**Deployment type**

Global Standard

Data Zone Standard

Explanation:

For a customer support workload used across multiple regions, Global Standard deployment is the right choice: it routes each request to the nearest available Azure region automatically, reducing latency globally and providing the highest throughput and availability. For a marketing workload that must remain within a specific region due to data residency requirements, a Data Zone Standard or single-region deployment ensures all compute and data processing occurs within a defined geographic boundary, satisfying GDPR and local data sovereignty rules. Microsoft Foundry 's deployment types are designed around exactly this trade-off:

Global routing for performance-critical multi-region workloads, and Data Zone or Regional isolation for data- residency-constrained workloads. Choosing the wrong deployment type can result in either compliance violations or unnecessary latency.

Microsoft Learn Reference Topic: Model deployment options in Microsoft Foundry - Global, Data Zone, and Regional deployment types

Deployment types	Workload deployment types
Standard	
Global Standard	Customer support
Data Zone Standard	Marketing
Data Zone Batch	

**NEW QUESTION # 72**

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. NOTE: Each correct selection is worth one point. Choose two .

- A. Tune chunk size and overlap to match content structure.

- B. Optimize the length of embedding vectors.
- C. Increase token limits for all requests.
- D. Implement an optimized re-ranker.

**Answer: A,D**

Explanation:

Microsoft's RAG optimization guidance identifies two high-impact, low-cost improvements for retrieval quality that do not require changing the embedding model. First, tuning chunk size and overlap (option A): chunk size determines how much context each retrieved piece contains - too large and irrelevant content dilutes the signal; too small and answers may be split across chunks. Adjusting these parameters requires only re-indexing the documents with zero additional compute cost. Second, implementing a re-ranker (option B): a re-ranker is a cross-encoder model that takes the top-N retrieved chunks and re-scores them based on their specific relevance to the query, significantly improving precision by filtering out contextually irrelevant chunks. Re-rankers add modest compute cost but are far cheaper than changing the embedding model, which would require re-embedding the entire knowledge base. Increasing token limits (option C) and optimizing embedding vector length (option D) do not address retrieval accuracy without an embedding model change. Microsoft Learn Reference Topic: Optimize RAG pipelines - Chunk size tuning and re-ranking in Azure AI Search and Azure Machine Learning

**NEW QUESTION # 73**

Drag and Drop Question

A team deploys a generative AI application that uses a model deployed in Microsoft Foundry. The application must support latency monitoring under production load.

You need to enable performance observability.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Enable performance observability	Answer Area
Set up the model endpoint in Foundry.	
Enable logging.	
Apply a term blocklist.	
Generate token embeddings.	
Configure monitoring metrics.	

**Answer:**

Explanation:

**Enable performance observability**

Set up the model endpoint in Foundry.

Enable logging.

Apply a term blacklist.

Generate token embeddings.

Configure monitoring metrics.

**Answer Area**

Set up the model endpoint in Foundry.

Configure monitoring metrics.

Enable logging.

#### NEW QUESTION # 74

An Azure Machine Learning workspace processes sensitive training data.

The workspace must NOT be accessible from the public internet.

You need to restrict network access.

Which configuration should you implement?

- A. Azure Firewall rules
- B. Service endpoints
- C. Private endpoints
- D. Network security groups

**Answer: C**

Explanation:

Azure Private Endpoints are network interfaces that connect your Azure ML workspace to your Virtual Network using a private IP address from the VNet 's address space. Once a private endpoint is created and DNS is configured, all traffic to the workspace - including the Studio UI, REST API, and SDK calls - travels entirely over Microsoft 's private backbone rather than the public internet. The workspace 's public endpoint can then be completely disabled. Azure Firewall (option A) filters traffic at the network layer but still requires the workspace to have a public IP. Network Security Groups (option C) control traffic within VNets but cannot block the public endpoint of a PaaS service. Service endpoints (option D) keep traffic on the Azure backbone but the workspace still has a public-facing address. Private endpoints are the only option that fully removes the public network presence. Microsoft Learn Reference Topic: Configure a private endpoint for Azure Machine Learning workspace

#### NEW QUESTION # 75

.....

The certificate is of significance in our daily life. At present we will provide all candidates who want to pass the AI-300 exam with three different versions for your choice. Any of the three versions can work in an offline state, and the version makes it possible that the websites is available offline. If you use the quiz prep, you can use our latest AI-300 Exam Torrent in anywhere and anytime. How can you have the chance to enjoy the study in an offline state? You just need to download the version that can work in an offline state, and the first time you need to use the version of our AI-300 quiz torrent online.

**Valid AI-300 Test Syllabus:** <https://www.actualcollection.com/AI-300-exam-questions.html>

Complex designs do not exist in our AI-300 exam guide, Our 24/7 customer support provides assistance to help AI-300 dumps users solve their technical hitches during their test preparation, You will get our valid AI-300 dumps torrent and instantly download the exam pdf after payment, Our Valid AI-300 Test Syllabus team prepares the mind of the client as per the Valid AI-300 Test Syllabus exam, Fully SSL Secure System On The Purchase of Microsoft AI-300 Brandumps.

Joseph's current role gives him visibility AI-300 into the latest trends in cyber security, from both leading vendors and customers, Discover hundreds of tips and tricks you can Valid AI-300 Test Syllabus use right away with your iPad, iPad mini, or iPhone to maximize its functionality.

