

PRACTICE QUESTIONS

Oracle

1Z0-1158-26

Oracle Cloud Infrastructure Enterprise AI Professional

SHARED BY [u/crispzebra1981](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	Oracle 1Z0-1158-26, Oracle Cloud Infrastructure Enterprise AI Professional

Shared by u/crispzebra1981 on aiexamcerts

10 practice questions for the Oracle 1Z0-1158-26 exam, Oracle Cloud Infrastructure Enterprise AI Professional, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/oracle/1z0-1158-26. Spotted a mistake? Email support@aiexamcerts.com with the exam code 1Z0-1158-26 and the question number.

1. Which technique supplies a few worked input and output pairs inside the prompt to steer behavior without any training?
 - A Gradient checkpointing
 - B Few-shot prompting
 - C Quantization of the model weights
 - D Knowledge distillation into a smaller student

Answer: B

Few-shot prompting places examples in the prompt so the model imitates the pattern with no weight updates. Gradient checkpointing saves memory during training, quantization shrinks weights, and distillation trains a smaller model from a larger one.

2. On-demand inference is billed for reserved hosting units whether or not any requests are sent.
 - A True
 - B False

Answer: B (False)

On-demand pricing charges for the requests or tokens actually processed, while a dedicated AI cluster charges for reserved capacity regardless of traffic. The statement describes the dedicated cluster model, not on-demand.

3. A team indexes millions of chunks and needs fast approximate similarity search. Which index family fits best?
 - A HNSW graph index
 - B Unique B-tree on the chunk text column
 - C Bitmap index on the chunk identifier
 - D Materialized view refreshed nightly

Answer: A

HNSW is an approximate nearest-neighbor graph index that trades a small loss of recall for much faster search at scale. A unique B-tree supports exact key lookups rather than similarity ranking, bitmap indexes suit low-cardinality columns, and a materialized view precomputes results without indexing vectors.

4. A model trained on broad data that can be adapted to many downstream tasks is called a _____ model.

Answer: foundation

Foundation models are large pretrained models that serve as a base for many tasks through prompting, retrieval, or fine-tuning. The term describes how the model is built and reused, not its size alone.

5. Which two conditions favor a dedicated AI cluster over on-demand inference? (Choose two.)

- A Steady high traffic that needs predictable latency
- B A custom fine-tuned model that must be hosted behind its own endpoint
- C Occasional prototype calls from a single developer
- D A requirement to pay only for tokens generated, with no reserved capacity
- E A nightly batch of fewer than ten requests

Answer: A, B

Steady high traffic and custom hosted models both justify reserved capacity for predictable latency and endpoint isolation. Occasional prototype calls and tiny batches are cheaper on demand, and the token-only pricing option describes on-demand usage.

6. What is the main risk of fine-tuning on a narrow dataset?

- A The endpoint can no longer be called over HTTPS
- B Degraded performance on tasks outside the narrow training distribution
- C Automatic deletion of the base model from the catalog
- D Mandatory migration of every prompt and template into a new language for all markets

Answer: B

Narrow fine-tuning can erode general capability, which is why teams re-test behavior beyond the target task. The other options describe transport, catalog, or language issues that fine-tuning does not cause.

7. An agent loops between two tools without making progress. Which control best limits this?

- A A maximum number of reasoning or tool-call steps with an explicit stop condition
- B Unlimited retries that repeat the same arguments
- C Raising the temperature to its highest setting so the agent explores more options at each step
- D Removing the tool descriptions from the prompt

Answer: A

A step cap with an explicit stop condition ends runaway loops and surfaces the failure for review. Unlimited retries repeat the same problem, a very high temperature adds randomness, and removing tool descriptions makes the agent less able to use its tools.

8. An agent should always be given write access to production systems so that it can fix problems without waiting for human approval.
- A True
 - B False

Answer: B (False)

Least privilege and human approval for risky actions are the safer design, because a faulty model output can then cause real damage. Broad write access turns every mistake into a production incident.

9. Which step comes first in a responsible AI review for a customer-facing assistant?
- A Launch to all customers and then collect complaints
 - B Choose the model with the lowest price and nothing else as the deciding factor
 - C Define the intended use, the foreseeable risks, and the affected groups
 - D Disable logging to protect privacy before any testing

Answer: C

Defining the intended use and foreseeable harms sets the controls and evaluations that follow. Launching first makes customer complaints the test, choosing purely on price ignores risk, and disabling logs removes the evidence needed for oversight.

10. Which evaluation approach best tests a model for bias across user groups?
- A Measure only overall accuracy on one random sample
 - B Compare error rates and output quality across defined user segments
 - C Test only with English prompts written by one region's staff and their colleagues
 - D Remove demographic columns from the dataset and stop there

Answer: B

Comparing error rates and output quality across defined segments reveals disparities that overall accuracy hides. A single random sample, one language, or simply dropping columns does not test for unequal outcomes.