Oracle

1Z0-1158-26

Oracle Cloud Infrastructure Enterprise AI Professional

Latest Version: 6.0

Question: 1

In OCI Generative AI on-demand pricing, how is a transaction defined for character-billed models?

A. One transaction equals one full API call, regardless of length.
B. One transaction equals one gigabyte of Object Storage.
C. One character equals one transaction, billed per 10,000 transactions for Cohere and Meta models.
D. One transaction equals one unit-hour of dedicated GPU capacity.

Answer: C

Explanation:
For on-demand inferencing, 1 character = 1 transaction, billed per 10,000 transactions for Cohere and Meta models. Frontier models (xAI Grok, Google Gemini, OpenAI gpt-oss) are billed per 1,000,000 tokens instead.

Question: 2

Which three base models support fine-tuning on OCI?

A. Every model in the catalog can be fine-tuned without restriction
B. Meta Llama 3.3 70B, Meta Llama 3.1 405B, and Cohere Command R 08-2024
C. Cohere Rerank 3.5, Cohere Embed 4, and Command A Vision
D. Google Gemini 2.5 Pro, Gemini 2.5 Flash, and Gemini 2.5 Flash-Lite

Answer: B

Explanation:
Three models support fine-tuning on OCI: Meta Llama 3.3 70B (recommended, most flexible), Meta Llama 3.1 405B (maximum depth), and Cohere Command R 08-2024 (enterprise RAG, most affordable). OCI uses the T-Few method.

Question: 3

Which sequence correctly names the four steps of the agent loop?

A. Perceive, Reason, Act, Observe
B. Compile, Link, Load, Execute
C. Encode, Decode, Rerank, Return
D. Perceive and then Act, with no reasoning or observation steps

Answer: A

Explanation:
Every agent run is the perceive-reason-act-observe loop repeating until a termination condition is met — read the inputs, decide what to do, produce an output, then record the result, which feeds back into the next iteration.

Question: 4

Vector Store operations in the OCI Responses API require two separate clients.
What are they used for?

A. A single client is used for everything; two clients are never required.
B. One client for training and one for billing.
C. One client for Docker and one for Kubernetes.
D. The control plane client (cp_client) for management, and the data plane client (dp_client) for file operations and search.

Answer: D

Explanation:
Vector Store operations use two clients on different endpoints: the control plane client (cp_client) creates, lists, updates, and deletes stores, while the data plane client (dp_client) uploads files, attaches them, and runs search.

Question: 5

In the Responses API, which mechanism chains a follow-up turn to the previous turn without resending the full message history?

A. Re-uploading the whole conversation to a vector store each turn
B. Rebuilding the messages list manually on every call
C. Passing previous_response_id to continue the conversation
D. Storing the prior response inside the messages array

Answer: C

Explanation:
The Responses API stores each turn server-side and threads follow-ups with previous_response_id — no messages array to manage. Input tokens stay flat because you send only the new message, not the full accumulated history.

Question: 6

Which characteristics describe the OCI Responses API Code Interpreter tool?

A. It executes Python on the developer's laptop with full internet access.
B. It runs Python in an isolated OCI container with 420+ preinstalled libraries, no external network access, and a container that expires after 20 minutes of inactivity.
C. It replaces Vector Stores for document retrieval.
D. It gives the model unrestricted shell access to the customer's tenancy.

Answer: B

Explanation:
Code Interpreter runs Python in a sandboxed OCI container with 420+ preinstalled libraries and no external network access. A container expires after 20 minutes of inactivity, and all in-memory state is lost on expiry.

Question: 7

Which statement describes the relationship between a Generative AI Application and a Deployment for hosted agents?

A. Each Application can have only one Deployment ever, so upgrades require a new Application.
B. The Application provides the stable endpoint URL and top-level configuration; only one Deployment can be active at a time, and it serves all requests.
C. The Application is the Dockerfile, and the Deployment is the OCIR repository.
D. The Deployment provides the endpoint URL, and the Application is a temporary container.

Answer: B

Explanation:
The Generative AI Application is the top-level resource that provides the stable endpoint URL. A Deployment represents a specific container image; only one can be active per Application at a time, and switching deployments enables zero-downtime version upgrades.

Question: 8

In a function-calling tool schema, which field is the most critical for correct tool selection?

A. The length of the tool name in characters
B. The color assigned to the tool in the console
C. The description field, which tells the model when to call the tool
D. The tool's file-modification timestamp

Answer: C

Explanation:

The description is the most important field in any tool schema — the model reads it to decide when to call the tool. A vague description leads to poor or random tool selection, so it should state when to use the tool, its inputs, and what it returns.

Question: 9

What role does a Project play when using the OCI Responses API?

A. It is the required organizing container for agent artifacts — responses, conversations, files, and containers — and holds data-retention, memory, and compaction settings.
B. It is the container image that must be rebuilt after every API call.
C. It replaces compartments and IAM policies for all Generative AI resources.
D. It is a scheduled job that only syncs Object Storage documents.

Answer: A

Explanation:
A Project is required to call the OCI Responses API. It organizes all agent artifacts and holds data-retention policies, long-term memory configuration, and conversation-history compaction rules. Projects are fully isolated from each other.

Question: 10

What is the main characteristic of greedy decoding?

A. It requires a high temperature to produce diverse output.
B. It picks the highest-probability word at each step of decoding.
C. It always ignores the vocabulary distribution and emits a fixed token.
D. It selects words at random from the low-probability candidates.

Answer: B

Explanation:
Greedy decoding picks the highest-probability word at each step. The chosen word is appended to the input and decoding continues iteratively, one word at a time.

Thank You for Trying Our Product

For More Information – Visit link below:

https://www.examsboost.com/

15 USD Discount Coupon Code:

G74JA8UF

FEATURES

- ✓ 90 Days Free Updates
- ✓ Money Back Pass Guarantee
- ✓ Instant Download or Email Attachment
- ✓ 24/7 Live Chat Support
- ✓ PDF file could be used at any Platform
- ✓ 50,000 Happy Customer