# Oracle

# 1Z0-1127-24

## Oracle Cloud Infrastructure 2024 Generative AI Professional

**For More Information – Visit link below:**

**https://www.examsempire.com/**

**Product Version**

1. **Up to Date products, reliable and verified.**
2. **Questions and Answers in PDF Format.**

# Latest Version: 7.0

## Question: 1

In LangChain, which retriever search type is used to balance between relevancy and diversity?

A. top k
B. mmr
C. similarity_score_threshold
D. similarity

**Answer: B**

Explanation:
In LangChain, the "mmr" (Maximal Marginal Relevance) search type is used to balance between relevancy and diversity when retrieving documents. This technique aims to select documents that are not only relevant to the query but also diverse from each other. This helps in avoiding redundancy and ensures that the retrieved set of documents covers a broader aspect of the topic.
Maximal Marginal Relevance (MMR) works by iteratively selecting documents that have high relevance to the query but low similarity to the documents already selected. This ensures that each new document adds new information and perspectives, rather than repeating what is already included.
Reference
LangChain documentation on retrievers and search types
Research papers and articles on Maximal Marginal Relevance (MMR)

## Question: 2

What does a dedicated RDMA cluster network do during model fine-tuning and inference?

A. It leads to higher latency in model inference.
B. It enables the deployment of multiple fine-tuned models.
C. It limits the number of fine-tuned model deployable on the same GPU cluster.
D. It increases G PU memory requirements for model deployment.

**Answer: B**

Explanation:
A dedicated RDMA (Remote Direct Memory Access) cluster network is crucial during model fine-tuning and inference because it facilitates high-speed, low-latency communication between GPUs. This capability is essential for scaling up the deployment of multiple fine-tuned models across a GPU cluster. RDMA allows data to be transferred directly between the memory of different computers without involving the CPU, leading to significantly reduced latency and higher throughput. This efficiency is particularly important in the context of fine-tuning and deploying large language models, where the

speed and efficiency of data transfer can impact overall performance and scalability.
By enabling fast and efficient communication, a dedicated RDMA cluster network supports the deployment of multiple fine-tuned models on the same GPU cluster, enhancing both flexibility and scalability in handling various AI workloads.
Reference
Oracle Cloud Infrastructure (OCI) documentation on RDMA cluster networks
Technical resources on the benefits of RDMA in high-performance computing environments

## Question: 3

Which role docs a "model end point" serve in the inference workflow of the OCI Generative AI service?

A. Hosts the training data for fine-tuning custom model
B. Evaluates the performance metrics of the custom model
C. Serves as a designated point for user requests and model responses
D. Updates the weights of the base model during the fine-tuning process

**Answer: C**

Explanation:
In the inference workflow of the OCI Generative AI service, a "model endpoint" is a critical component. It serves as a designated point for handling user requests and providing model responses. When users or applications send requests to the model endpoint, the endpoint processes these requests by passing them to the deployed model. The model then generates responses based on the input data, and these responses are returned to the user through the same endpoint. This setup facilitates efficient and scalable interaction with the AI model, ensuring that inference can be performed seamlessly and reliably.
Reference
Oracle Cloud Infrastructure (OCI) Generative AI service documentation
General principles of model deployment and inference in cloud services

## Question: 4

Which is a distinguishing feature of "Parameter-Efficient Fine-tuning (PEFT)" as opposed to classic Tinetuning" in Large Language Model training?

A. PEFT involves only a few or new parameters and uses labeled, task-specific data.
B. PEFT modifies all parameters and uses unlabeled, task-agnostic data.
C. PEFT does not modify any parameters but uses soft prompting with unlabeled data. PEFT modifies
D. PEFT parameters and b typically used when no training data exists.

**Answer: A**

Explanation:

Parameter-Efficient Fine-Tuning (PEFT) is a technique used in large language model training that focuses on adjusting only a subset of the model's parameters rather than all of them. This approach involves using labeled, task-specific data to fine-tune new or a limited number of parameters. PEFT is designed to be more efficient than classic fine-tuning, which typically adjusts all the parameters of the model. By only updating a small fraction of the model's parameters, PEFT reduces the computational resources and time required for fine-tuning while still achieving significant performance improvements on specific tasks.
Reference
Research papers on Parameter-Efficient Fine-Tuning (PEFT)
Technical documentation on fine-tuning techniques for large language models

## Question: 5

How does the Retrieval-Augmented Generation (RAG) Token technique differ from RAG Sequence when generating a model's response?

A. Unlike RAG Sequence, RAG Token generates the entire response at once without considering individual parts.
B. RAG Token does not use document retrieval but generates responses based on pre-existing knowledge only.
C. RAG Token retrieves documents oar/at the beginning of the response generation and uses those for the entire content
D. RAG Token retrieves relevant documents for each part of the response and constructs the answer incrementally.

**Answer: D**

Explanation:
The Retrieval-Augmented Generation (RAG) technique enhances the response generation process of language models by incorporating relevant external documents. RAG Token and RAG Sequence are two variations of this technique.
RAG Token retrieves relevant documents for each part of the response and constructs the answer incrementally. This means that during the response generation process, the model continuously retrieves and incorporates information from external documents as it generates each token (or part) of the response. This allows for more dynamic and contextually relevant answers, as the model can adjust its retrieval based on the evolving context of the response.
In contrast, RAG Sequence typically retrieves documents once at the beginning of the response generation and uses those documents to generate the entire response. This approach is less dynamic compared to RAG Token, as it does not adjust the retrieval process during the generation of the response.
Reference
Research articles on Retrieval-Augmented Generation (RAG) techniques
Documentation on advanced language model inference methods

# Thank You for Trying Our Product

**Special 16 USD Discount Coupon: NSZUBG3X**

**Email:** support@examsempire.com

**Check our Customer Testimonials and ratings available on every product page.**

**Visit our website.**

**https://examsempire.com/**