Oracle Cloud Infrastructure 2025 – Generative AI Professional Practice Exam

**1. What is the role of the inputs parameter in the given code snippet?**

inputs = [

"Learn about the Employee Stock Purchase Plan",

"Reassign timecard approvals during leave",

"View my payslip online",

]

embed\_text\_detail.inputs = inputs

☐ It sets the output format for the embeddings.

☐ It controls the maximum number of embeddings the model can generate.

☐ It provides metadata about the embedding process.

☐ It specifies the text data that will be converted into embeddings.

**2. What must be done before you can delete a knowledge base in Generative AI Agents?**

☐ Disconnect the database tool connection.

☐ Archive the knowledge base for future use.

☐ Reassign the knowledge base to a different agent.

☐ Delete the data sources and agents using that knowledge base.

**3. A data scientist is training a machine learning model to predict customer purchase behavior. After each training epoch, they analyze the loss metric reported by the model to evaluate its performance. They notice that the loss value is decreasing steadily over time. What does the loss metric indicate about the model's predictions in this scenario?**

☐ Loss measures the total number of predictions made by the model during training.

☐ Loss only evaluates the accuracy of correct predictions, ignoring the impact of incorrect predictions.

☐ Loss quantifies how far the model's predictions deviate from the actual values, indicating how wrong the predictions are.

☐ Loss reflects the quality of predictions and should increase as the model improves.

**4. What does accuracy measure in the context of fine-tuning results for a generative model?**

☐ The proportion of incorrect predictions made by the model during an evaluation.

☐ The depth of the neural network layers used in the model.

☐ How many predictions the model made correctly out of all the predictions in an evaluation.

☐ The number of predictions a model makes, regardless of whether they are correct or incorrect.

**5. What is the purpose of the VECTOR field in the Oracle Database 23ai table for Generative AI Agents?**

☐ To store the document TITLE.

☐ To store the URL references for the documents.

☐ To store the embeddings generated from the BODY content.

☐ To assign a unique identifier DOCID to each document.

**6. A researcher is exploring generative models for various tasks. While diffusion models have shown excellent results in generating high-quality images, they encounter significant challenges in adapting these models for text. What is the primary reason why diffusion models are difficult to apply to text generation tasks?**

☐ Because diffusion models can only produce images.

☐ Because text representation is categorical, unlike images.

☐ Because text generation does not require complex models.

☐ Because text is not categorical.

**7. In the simplified workflow for managing and querying vector data, what is the role of indexing?**

☐ Mapping vectors to a data structure for faster searching, enabling efficient retrieval.

☐ Compressing vector data for minimized storage usage.

☐ Categorizing vectors based on their originating data type (text, images, audio).

☐ Converting vectors into a non-indexed format for easier retrieval.

**8. How does a presence penalty function when using OCI Generative AI chat models?**

☐ It applies a penalty only if the token has appeared more than twice.

☐ It only penalizes tokens that have never appeared in the text before.

☐ It penalizes a token each time it appears after the first occurrence.

☐ It penalizes all tokens equally, regardless of how often they have appeared.

**9. A data science team is fine-tuning multiple models using the Oracle Generative AI service. They select the cohere.command-r-08-2024 base model and fine-tune it on three different datasets for three separate tasks. They plan to use the same fine-tuning AI cluster for all models. What is the total number of units provisioned for the cluster?**

☐ 2

☐ 8

☐ 6

☐ 1

**10. Accuracy in vector databases contributes to the effectiveness of LLMs by preserving a specific type of relationship. What is the nature of these relationships, and why are they crucial for language models?**

☐ Semantic relationships, and they are crucial for understanding context and generating precise language.

☐ Linear relationships, and they simplify the modeling process.

☐ Hierarchical relationships, and they are important for structuring database queries.

☐ Temporal relationships, and they are necessary for predicting future linguistic trends.

**11. How does the temperature setting in a decoding algorithm influence the probability distribution over the vocabulary?**

☐ Temperature has no effect on the probability distribution; it only changes the speed of decoding.

☐ Decreasing temperature broadens the distribution, making less likely words more probable.

☐ Increasing temperature flattens the distribution, allowing for more varied word choices.

☐ Increasing temperature removes the impact of the most likely word.

**12. Which component of Retrieval-Augmented Generation (RAG) evaluates and prioritizes the information retrieved by the retrieval system?**

☐ Encoder-decoder

☐ Ranker

☐ Retriever

☐ Generator

**13. What is the destination port range that must be specified in the subnet's ingress rule for an Oracle Database in OCI Generative AI Agents?**

☐ 8080-8081

☐ 1433-1434

☐ 3306-3307

☐ 1521-1522

**14. Which statement is true about the "Top p" parameter of OCI Generative AI chat models?**

☐ "Top p" selects tokens from the "top k" tokens sorted by probability.

☐ "Top p" assigns penalties to frequently occurring tokens.

☐ "Top p" determines the maximum number of tokens per response.

☐ "Top p" limits token selection based on the sum of their probabilities.

**15. You are debugging and testing an OCI Generative AI chat model. What is the model behavior if you don't provide a value for the seed parameter?**

☐ The model generates responses deterministically.

☐ The model restricts the maximum number of tokens that can be generated.

☐ The model assigns a default seed value of 9999.

☐ The model gives diverse responses.

**16. You are hosting a dedicated AI cluster using the OCI Generative AI service. You need to employ the maximum number of endpoints due to high workload. How many dedicated AI clusters will you require to host at least 60 endpoints?**

☐ 1

☐ 5

☐ 2

☐ 3

**17. A data scientist is exploring Retrieval-Augmented Generation (RAG) for a natural language processing project. Which statement is true about RAG?**

☐ It is solely used in QA-based scenarios.

☐ It is non-parametric and can theoretically answer questions about any corpus.

☐ It is primarily parametric and requires a different model for each corpus.

☐ It is not suitable for fact-checking because of high hallucination occurrences.

**18. How many numerical values are generated for each input phrase when using the cohere.embed-english-light-v3.0 embedding model?**

☐ 256

☐ 384

☐ 512

☐ 1024

**19. What advantage does fine-tuning offer in terms of improving model efficiency?**

☐ It eliminates the need for annotated data during training.

☐ It reduces the number of tokens needed for model performance.

☐ It increases the model's context window size.

☐ It improves the model's understanding of human preferences.

**20. In an OCI Generative AI chat model, which of these parameter settings is most likely to induce hallucinations and factually incorrect information?**

☐ `temperature = 0.5`, `top\_p = 0.9`, and `frequency\_penalty = 0.5`

☐ `temperature = 0.0`, `top\_p = 0.7`, and `frequency\_penalty = 1.0`

☐ `temperature = 0.2`, `top\_p = 0.6`, and `frequency\_penalty = 0.8`

☐ `temperature = 0.9`, `top\_p = 0.8`, and `frequency\_penalty = 0.1`

**21. What happens to the status of an endpoint after initiating a move to a different compartment?**

☐ The endpoint becomes Inactive permanently, and you need to create a new endpoint.

☐ The endpoint is deleted and recreated in the new compartment.

☐ The status remains Active throughout the move.

☐ The status changes to Updating during the move and returns to Active after completion.

**22. You are developing an application that displays a house image along with its related details. Assume that you are using Oracle Database 23ai. Which data type should be used to store the embeddings of the images in a database column?**

☐ Float32

☐ INT

☐ Double

☐ VECTOR

**23. A startup is using Oracle Generative AI's on-demand inferencing for a chatbot. The chatbot processes user queries and generates responses dynamically. One user enters a 200-character prompt, and the model generates a 500-character response. How many transactions will be billed for this inference call?**

☐ 1 transaction per API call, regardless of length

☐ 200 transactions

☐ 700 transactions

**24. Which of these does NOT apply when preparing PDF files for OCI Generative AI Agents?**

☐ Charts must be two-dimensional with labeled axes.

☐ PDF files can include images and charts.

☐ Hyperlinks in PDFs are excluded from chat responses.

☐ Reference tables must be formatted with rows and columns.

**25. A marketing team is using Oracle's Generative AI service to create promotional content. They want to generate consistent responses for the same prompt across multiple runs to ensure uniformity in their messaging. They notice that the responses vary each time they run the model, despite keeping the prompt and other parameters the same. Which parameter should they modify to ensure identical outputs for the same input?**

☐ frequency\_penalty

☐ temperature

☐ seed

☐ top\_p

**26. What happens when this line of code is executed?**

```

embed\_text\_response = generative\_ai\_inference\_client.embed\_text(embed\_text\_detail)

```

☐ It sends a request to the OCI Generative AI service to generate an embedding for the input text.

☐ It initiates a connection to OCI and authenticates using the user's credentials.

☐ It processes and configures the OCI profile settings for the inference session.

☐ It initializes a pretrained OCI Generative AI model for use in the session.

**27. Which phase of the RAG pipeline includes loading, splitting, and embedding of documents?**

☐ Retrieval

☐ Ingestion

☐ Generation

☐ Evaluation

**28. In the context of RAG, how might the concept of Groundedness differ from that of Answer Relevance?**

☐ Groundedness pertains to factual correctness, while Answer Relevance concerns query relevance.

☐ Groundedness measures relevance to the user query, while Answer Relevance evaluates data integrity.

☐ Groundedness refers to contextual alignment, while Answer Relevance deals with syntactic accuracy.

☐ Groundedness focuses on data integrity, while Answer Relevance emphasizes lexical diversity.

**29. In the context of generating text with a Large Language Model (LLM), what does the process of greedy decoding entail?**

☐ Picking a word based on its position in a sentence structure.

☐ Using a weighted random selection based on a modulated distribution.

☐ Choosing the word with the highest probability at each step of decoding.

☐ Selecting a random word from the entire vocabulary at each step.

**30. What is the role of the OnDemandServingMode in the following code snippet?**

```

chat\_detail.serving\_mode = oci.generative\_ai\_inference.models.OnDemandServingMode(model\_id="ocidl.generativeaimodel.ocl.eu-frankfurt-1.XXXXXXXXXXXXXXXXXXXXXX")

```

☐ It defines the retry strategy for handling failures during model inference.

☐ It initializes the model with the default configuration profile for inference.

☐ It specifies that the Generative AI model should serve requests only on demand, rather than continuously.

☐ It configures the model to use batch processing for requests.

**31. A machine learning engineer is exploring T-Few fine-tuning to efficiently adapt a Large Language Model (LLM) for a specialized NLP task. They want to understand how T-Few fine-tuning modifies the model compared to standard fine-tuning techniques. Which of these best describes the characteristic of T-Few fine-tuning for LLMs?**

☐ It selectively updates only a fraction of the model's weights.

☐ It increases the training time as compared to Vanilla fine-tuning.

☐ It does not update any weights but restructures the model architecture.

☐ It updates all the weights of the model uniformly.

**32. How can you affect the probability distribution over the vocabulary of a Large Language Model (LLM)?**

☐ By adjusting the token size during the training phase.

☐ By modifying the model's training data.

☐ By using techniques like prompting and training.

☐ By restricting the vocabulary used in the model.

**33. Which of these is NOT a supported knowledge base data type for OCI Generative AI Agents?**

☐ OCI Object Storage files with text and PDFs

☐ OCI Search with OpenSearch

☐ Oracle Database 23ai vector search

☐ Custom-built file systems

**34. What happens to chat data and retrieved context after the session ends in OCI Generative AI Agents?**

☐ They are archived for audit purposes.

☐ They are stored for training the Large Language Models (LLMs).

☐ They are permanently deleted and not retained.

☐ They are stored in isolation for future customer usage, ensuring maximum security but not used for training.

**35. A company is using a Generative AI model to assist customer support agents by answering product-related queries. Customer query: "What are the supported features of your new smart watch?" Generative AI model response: "The smart watch includes ECG monitoring, blood sugar tracking, and solar charging." Upon review of this response, the company notes that blood sugar tracking and solar charging are not actual features of their smart watch.**

**These details were not part of the company's product documentation or database. What is the most likely cause of this model behavior?**

☐ The model is hallucinating, confidently generating responses that are not grounded in factual or provided data.

☐ The model encountered a prompt that was too ambiguous, leading to random outputs.

☐ The model is overfitting to specific details from unrelated training data, causing inaccuracies.

☐ The model was unable to access the company's database, so it defaulted to guessing feature sets based on similar products.

**36. How is the totalTrainingSteps parameter calculated during fine-tuning in OCI Generative AI?**

☐ `totalTrainingSteps = (totalTrainingEpochs \* trainingBatchSize) / size(trainingDataset)`

☐ `totalTrainingSteps = (totalTrainingEpochs + size(trainingDataset)) \* trainingBatchSize`

☐ `totalTrainingSteps = (size(trainingDataset) \* trainingBatchSize) / totalTrainingEpochs`

☐ `totalTrainingSteps = (totalTrainingEpochs \* size(trainingDataset)) / trainingBatchSize`

**37. When specifying a data source, what does enabling multi-modal parsing do?**

☐ Automatically tags files and folders in the bucket.

☐ Merges multiple data sources into a single knowledge base after parsing the files.

☐ Parses and includes information from charts and graphs in the documents.

☐ Parses and converts non-supported file formats into supported ones.

**38. When does a chain typically interact with memory in a run within the LangChain framework?**

☐ Continuously throughout the entire chain execution process.

☐ After user input but before chain execution, and again after core logic but before output.

☐ Before user input and after chain execution.

☐ Only after the output has been generated.

**39. What does a cosine distance of 0 indicate about the relationship between two embeddings?**

☐ They are similar in direction.

☐ They are unrelated.

☐ They have the same magnitude.

☐ They are completely dissimilar.

**40. In which scenario is soft prompting more appropriate compared to other training styles?**

☐ When the model requires continued pretraining on unlabeled data.

☐ When there is a need to add learnable parameters to a LLM without task-specific training.

☐ When there is a significant amount of labeled, task-specific data available.

☐ When the model needs to be adapted to perform well in a domain it was not originally trained on.

**41. A company is using a model in the OCI Generative AI service for text summarization. They receive a notification stating that the model has been deprecated. What action should the company take to ensure continuity in their application?**

☐ The company should ignore the notification as deprecated models remain available indefinitely.

☐ The company can continue using the model but should start planning to migrate to another model before it is retired.

☐ The company can request an extension to continue using the model after it is retired.

☐ The company must immediately stop using the model because it is no longer available and start using the newer model.

**42. Which fine-tuning methods are supported by the cohere.command-r-08-2024 model in OCI Generative AI?**

☐ T-Few and Vanilla

☐ LORA and Vanilla

☐ T-Few and LORA

☐ T-Few, LoRA, and Vanilla

**43. You want to build an LLM application that can connect application components easily and allow for component replacement in a declarative manner. What approach would you take?**

☐ Use agents.

☐ Use LangChain Expression Language (LCEL).

☐ Use Python classes like LLMChain.

☐ Use prompts.

**44. What does the OCI Generative AI service offer to users?**

☐ Only pretrained LLMs with customization options.

☐ A service requiring users to share GPUs for deploying LLMs.

☐ A limited platform that supports chat-based LLMs without hosting capabilities.

☐ Fully managed LLMs along with the ability to create custom fine-tuned models.

**45. When is fine-tuning an appropriate method for customizing an LLM?**

☐ When the LLM does not perform well on a particular task and the data required to adapt the LLM is too large for prompt engineering.

☐ When the LLM already understands the topics necessary for text generation.

☐ When you want to optimize the model without any instructions.

☐ When the LLM requires access to the latest data for generating outputs.

**48. What is the purpose of this endpoint variable in the code?**

```

endpoint = "[https://inference.generativeai.eu-frankfurt-1.oci.oraclecloud.com](https://inference.generativeai.eu-frankfurt-1.oci.oraclecloud.com)"

```

☐ It sets the retry strategy for the inference client.

☐ It specifies the availability domain where the OCI Generative AI model is hosted, ensuring inference happens in the correct region.

☐ It stores the OCI API key required for authentication.

☐ It defines the URL of the OCI Generative AI inference service.

**49. Which statement regarding fine-tuning and Parameter-Efficient Fine-Tuning (PEFT) is correct?**

☐ Fine-tuning requires training the entire model on new data, often leading to substantial computational costs, whereas PEFT involves updating only a small subset of parameters, minimizing computational requirements and data needs.

☐ PEFT requires replacing the entire model architecture with a new one designed specifically for the new task, making it significantly more data-intensive than fine-tuning.

☐ Both fine-tuning and PEFT require the model to be trained from scratch on new data, making them equally data and computationally intensive.

☐ Fine-tuning and PEFT do not involve model modification; they differ only in the type of data used for training, with fine-tuning requiring labeled data and PEFT utilizing unlabeled data.

**50. When activating content moderation in OCI Generative AI Agents, which of these can you specify?**

☐ Whether moderation applies to user prompts, generated responses, or both.

☐ The maximum file size for input data.

☐ The threshold for language complexity in responses.

☐ The type of vector search used for retrieval.