1 0.4 / 0.4 points

Which AWS generative AI offering focuses on providing multiple foundation models (e.g., from Anthropic, AI21) in a single managed service without requiring users to manage their own model infrastructure?

Hint: Check the newly introduced AWS service that hosts partner foundation models behind a unified API.

- Amazon S3 Glacier.
- ✓ Amazon Bedrock.
- AWS CloudFormation Pipeline for AI.
- Amazon Redshift ML.

2 0.4 / 0.4 points

In Azure's generative AI ecosystem, which statement is most accurate regarding the Azure OpenAI Service?

- It only supports open-source language models from Hugging Face repositories.
- It can be deployed on-premises with no connection to Azure's cloud.
- ✓ It provides fully managed access to GPT-4, ChatGPT, and DALL·E via enterprise security features.
- It automatically fine-tunes models without user-provided data or prompts.

3 0.4 / 0.4 points

When applying SHAP (SHapley Additive exPlanations) to a random forest, what best describes how feature attributions are computed?

Hint: SHAP draws on game theory's concept of fairly distributing contribution across all feature subsets.

- By applying anchor conditions that guarantee identical model decisions.
- By counting the number of leaf nodes each feature appears in.
- By assigning each feature a weight from a single global linear regression.
- ✓ By averaging local permutations that show how each feature changes the prediction across all possible subsets.

4 0.4 / 0.4 points

A RAG (Retrieval-Augmented Generation) system is being designed for real-time legal document summaries. Which approach is least aligned with RAG principles?

- Incorporating retrieved facts into the generated text to reduce hallucinations.
- ✓ Storing the entire text corpus as part of the model's fine-tuned parameters.
- Using embeddings to semantically match queries with relevant passages.
- Dynamically retrieving relevant paragraphs from an external knowledge base before generation.

5 0.4 / 0.4 points

A diffusion model trained on high-resolution artwork slowly denoises random noise into a coherent image. Which statement is incorrect about this approach?

Hint: Recall that diffusion-based generative models use many small steps in the denoising phase.

- It can produce diverse outputs by beginning with different noise initializations.
- It requires learning how data is progressively corrupted, then reversing that corruption.
- It learns a reverse noising process to systematically remove noise from random samples.
- ✓ It typically relies on short discrete sampling steps rather than iterative refinement.

6 0.4 / 0.4 points

Which of the following statements about LIME (Local Interpretable Model-agnostic Explanations) is most accurate?

Hint: LIME's hallmark is generating synthetic neighbors around the instance and training a small surrogate.

- It creates multiple local tree-based surrogates to approximate cluster-level model behavior.
- ✓ It perturbs features in the neighborhood of a prediction and fits a simpler local model.
- It focuses on ranking feature importances by integrating gradients across all training samples.
- It constructs a single global linear model to replicate the entire black-box model.

7 0.4 / 0.4 points

Why might a few-shot approach outperform a zero-shot approach when requesting a specialized data classification from a large language model?

Hint: Short, well-crafted examples can guide domain or format-specific tasks effectively.

- Few-shot triggers catastrophic forgetting in the LLM, making it more creative.
- Zero-shot forces the LLM to rely only on knowledge from its system prompt.
- ✓ Few-shot examples prime the LLM with format and context, reducing confusion in the specific domain.
- Zero-shot typically allocates a higher token context window, thus overshadowing the final output.

8 1 / 1 point

An enterprise wants a domain-specific LLM for specialized biotech patents. Which rationale best supports a parameter-efficient finetuning approach (e.g., LoRA) over training a foundation LLM from scratch?

Hint: Large pre-trained models can be adapted with minimal overhead using advanced fine-tuning strategies.

- Foundation models rarely contain any knowledge about biotech topics.
- ✓ Parameter-efficient fine-tuning maintains large-scale knowledge while cheaply specializing on new domain data.
- Fine-tuning discards the model's original general knowledge in favor of new tasks.
- Training from scratch is typically faster if you have large GPU clusters.

9 0.4 / 0.4 points

In chain-of-thought prompting, an LLM is guided to reveal intermediate reasoning steps. Which result is most commonly observed if the chain-of-thought is systematically hidden vs. exposed?

Hint: Long-form reasoning can be boosted when the model "shows its work."

- Exposing chain-of-thought ensures fewer computations are performed by the model.
- Hiding the chain-of-thought always reduces token usage costs.
- Hiding the chain-of-thought typically increases transparency.
- ✓ Exposing chain-of-thought can improve the model's final accuracy on multi step tasks.

10 0.4 / 0.4 points

When building a generative chatbot on Google Cloud, which approach is typically recommended for production usage in Vertex AI?

Hint: Google emphasizes a pipeline that fetches relevant data, checks for safe outputs, then finalizes generation.

- Deploying BigQuery ML for storing embedding vectors and ignoring the pre-trained LLMs.
- Hard-coding all possible user queries as if-else statements in a Cloud Function.
- Using Vertex AI Model Garden with a Llama 2 foundation model fully memorizing all user data.
- ✓ Using a retrieval-augmented pipeline and a Vertex AI-hosted foundation model, with content filtering.