ATMIC NETWORKS
YOUR GUIDE TO GLOBAL CERTIFIATION

"Please note that these files may not be up to date. However, the questions will help you understand the exam format and typical question patterns."

www.atmicnetworks .com

Warning: Keep connected with our support team for latest updates

Question: 1

In the context of IBM Watsonx and generative AI models, you are tasked with designing a model that needs to classify customer support tickets into different categories. You decide to experiment with both zero-shot and few-shot prompting techniques. Which of the following best explains the key difference between zero-shot and few-shot prompting?

A. Zero-shot prompting does not use any examples in the input prompt, while few-shot prompting includes a few examples to guide the model.
B. Zero-shot prompting provides the model with a few example tasks to help it understand the problem, while few-shot prompting provides no examples at all.
C. In zero-shot prompting, the model learns from a large number of examples during the inference stage, while in few-shot prompting, only a single example is used.
D. Few-shot prompting is used only for training the model, while zero-shot prompting is used only for inference tasks.

Answer: A

Question: 2

In prompt engineering, prompt variables are used to make your prompts more dynamic and reusable. Which of the following statements best describes a key benefit of using prompt variables in IBM Watsonx Generative AI?

A. Prompt variables eliminate the need to change model parameters every time you generate a new response.
B. Prompt variables automatically improve the accuracy of responses by reducing model variance.
C. Prompt variables ensure that the AI's response format will always be consistent, regardless of the input data.
D. Prompt variables allow a single prompt template to handle multiple data points or scenarios by inserting different values.

Answer: D

Question: 3

You are working on a project where the AI model needs to generate personalized customer support responses based on various input fields like customer name, issue type, and product details. To make the system scalable and flexible, you decide to use prompt variables in your implementation. Which of the following statements accurately describe the benefits of using prompt variables in this scenario? (Select two)

A. Prompt variables improve the model's performance by optimizing its internal architecture, reducing computation time for each request.
B. Prompt variables reduce redundancy by allowing dynamic inputs to be injected into a single prompt template, improving scalability.

C. Using prompt variables allows the model to dynamically adjust its output based on context, without requiring multiple task-specific prompts.
D. Prompt variables eliminate the need for fine-tuning the model on specific tasks since they allow on-the-fly customization of responses.
E. Prompt variables require a complete re-training of the model whenever a new variable is introduced, which can be time-consuming.

Answer: B,C

Question: 4

You are tasked with designing an AI prompt to extract specific data from unstructured text. You decide to use either a zero-shot or a few-shot prompting technique with an IBM Watsonx model. Which of the following statements best describes the key difference between zero-shot and few-shot prompting?

A. Zero-shot prompting provides the model with examples, while few-shot prompting does not.
B. Zero-shot prompting requires no examples in the prompt, while few-shot prompting provides the model with one or more examples to clarify the task.
C. Few-shot prompting is used when the model is trained on supervised learning, while zero-shot prompting works only with unsupervised models.
D. Zero-shot prompting requires retraining the model with additional data, while few-shot prompting uses a pre-trained model without retraining.

Answer: B

Question: 5

You are building a chatbot using a generative AI model for a medical advice platform. During testing, you notice that the model occasionally generates medical information that contradicts established guidelines. This is an example of a model hallucination. Which prompt engineering technique would best mitigate the risk of hallucination in this scenario?

A. Implementing zero-shot learning techniques
B. Providing a list of credible sources in the prompt
C. Using more open-ended prompts
D. Increasing the model's temperature parameter

Answer: B

Question: 6

Your team has developed an AI model that generates automated legal documents based on user inputs. The client, a large law firm, wants to deploy this model but has stringent security, compliance, and auditability requirements due to the sensitive nature of the data. What is the most appropriate deployment strategy to meet these specific requirements?

A. Deploy the model on a hybrid cloud, with inference done on the client's on-premise servers and training done in the public cloud.
B. Deploy the model on a public cloud with built-in encryption and use APIs to connect to the client's private data.
C. Deploy the model using a serverless architecture to minimize operational overhead and maintain compliance.
D. Use a private cloud with role-based access controls (RBAC) and ensure model activity is logged for auditing purposes.

Answer: D

Question: 7

Your team is responsible for deploying a generative AI system that will interact with customers through automated chatbots. To improve the quality and consistency of responses across different queries and customer profiles, the team has developed several prompt templates. These templates aim to standardize input to the model, ensuring that outputs are aligned with business objectives. However, the team is debating whether using these prompt templates will provide tangible benefits in the deployment. What is the primary benefit of deploying prompt templates in this AI system?

A. Reducing the overall inference time by streamlining the input-output process for the model, ensuring faster responses.
B. Improving the scalability of the system by allowing the model to handle more diverse inputs without requiring additional fine-tuning.
C. Enhancing the model's ability to generalize across unseen data by training it specifically on the variations included in the prompt template.
D. Enabling more predictable and consistent outputs across different inputs, aligning the model's responses more closely with the business goals.

Answer: D

Question: 8

You have applied a set of prompt tuning parameters to a language model and collected the following statistics: ROUGE-L score, BLEU score, and memory utilization. Based on these metrics, how would you prioritize further optimizations to balance the model's performance in terms of output relevance and resource efficiency?

A. Maximize BLEU score and reduce memory utilization
B. Reduce memory utilization and maintain BLEU and ROUGE-L scores
C. Focus on improving the ROUGE-L score while increasing memory utilization
D. Increase memory utilization to reduce BLEU and ROUGE-L scores

Answer: B

Question: 9

You are working on a Retrieval-Augmented Generation (RAG) system to enhance the performance of a generative model. The RAG model needs to leverage a document corpus to generate answers to

complex questions. Which of the following steps is critical in the RAG pipeline to ensure accurate and relevant answer generation?

A. Fine-tuning the generative model on the entire document corpus without retrieval components.
B. Retrieving only the longest document in the corpus as the generative model can synthesize information more effectively from detailed content.
C. Indexing the document corpus using embeddings, retrieving relevant documents, and feeding them as context into the generative model.
D. Using keyword-based search to retrieve documents and then allowing the generative model to synthesize answers from those documents.

Answer: C

Question: 10

You are tasked with designing a prompt to fine-tune an IBM Watsonx model to summarize legal documents. The summaries must include only factual information, highlight key legal terms, and exclude any personal interpretations or subjective analysis. Which of the following is the best prompt to achieve this goal?

A. "Generate a detailed and engaging summary of this legal document, adding your insights to clarify complex legal points for the reader."
B. "Provide a summary of this legal document, focusing on factual information, including key legal terms and avoiding personal interpretation or subjective analysis."
C. "Create a brief summary of this legal document, ensuring to exclude any legal jargon and simplifying the content for a layperson audience."
D. "Summarize this legal document, focusing on key arguments and providing an analysis of the potential outcomes of the case."

Answer: B

Question: 11

When deploying AI assets in a deployment space, what is the most critical benefit of using deployment spaces in a large-scale enterprise environment?

A. Faster training times due to streamlined compute resources
B. Better data labeling quality through automated labeling tools
C. Improved model accuracy through hyperparameter tuning
D. Isolated environments to manage and monitor multiple model versions

Answer: D

Question: 12

When generating data for prompt tuning in IBM watsonx, which of the following is the most effective method for ensuring that the model can generalize well to a variety of tasks?

A. Use a diverse set of prompts covering multiple task domains with varying levels of complexity.

B. Prioritize prompts with repetitive patterns to help the model memorize key responses.
C. Focus on generating prompts specific to a single domain to train the model on specialized tasks.
D. Generate a single highly-detailed prompt that covers all potential use cases to maximize generalization.

Answer: A

Question: 13

You are using IBM watsonx Prompt Lab to experiment with different versions of a prompt to generate accurate and creative responses for a customer support chatbot. Which of the following best describes a key benefit of using Prompt Lab in the process of prompt engineering?

A. It provides a real-time environment for testing and refining prompts, helping to improve response quality.
B. It limits the number of iterations a user can test to prevent overfitting the prompt to specific outputs.
C. It allows users to generate AI models without the need for training data.
D. It automatically generates prompts based on industry-specific data without any user input.

Answer: A

Question: 14

You are working on enhancing the search functionality in a customer service chatbot by implementing the Retrieval-Augmented Generation (RAG) pattern. The chatbot needs to answer customer queries about various technical issues by retrieving relevant information from a knowledge base. Your team is discussing different ways to structure the RAG system and how to implement the pattern efficiently using existing tools. Which of the following statements best describes the RAG pattern, and how it should be implemented in the context of this chatbot?

A. The RAG pattern combines dense retrieval with a vector store, where retrieved documents are directly presented as final answers.
B. The RAG pattern integrates sparse retrieval with a rule-based system for generating responses based on exact document matches.
C. The RAG pattern prioritizes generating answers based on the frequency of document appearances in the retrieval phase, improving precision.
D. The RAG pattern enhances a generative model by retrieving relevant documents, which are then used as context for generating a final response.

Answer: D

Question: 15

A team is using IBM InstructLab to customize a large language model (LLM) to automate responses in a healthcare chatbot application. The team wants to ensure the chatbot can handle user queries accurately, based on domain-specific instructions. Which of the following correctly describes the role of the instruction optimization phase within the InstructLab workflow?

A. Instruction optimization involves retraining the model on a larger dataset for better accuracy.

B. Instruction optimization focuses on improving the dataset's quality by removing outliers and noise.
C. Instruction optimization refines prompts to improve the model's ability to follow task-specific instructions.

Answer: C

Question: 16

You are working on optimizing a generative AI model that will handle large-scale text generation tasks. The current model is slow during inference, and you need to improve its performance without increasing operational costs. You decide to use IBM Tuning Studio for optimization. Which of the following is the most significant benefit of using Tuning Studio in this scenario?

A. It pre-loads commonly used datasets, reducing the need for data handling during the training process.
B. It provides guidance on reducing the number of parameters in the model to improve inference speed.
C. It optimizes hyperparameters such as learning rate and batch size to reduce computational overhead during inference.
D. It automatically scales the model up or down depending on the input data size.

Answer: C

Question: 17

Your company is working on deploying a Watsonx Generative AI model for a client, and you have been asked to define the roles involved in the deployment process. Which of the following roles is responsible for ensuring that the model is properly integrated into the client's existing systems and that data pipelines are established for continuous model improvement?

A. DevOps Engineer
B. Data Scientist
C. Data Engineer
D. Solution Architect

Answer: D

Question: 18

You are working on a Retrieval-Augmented Generation (RAG) system where large-scale document retrieval is a critical component. To improve the efficiency and accuracy of retrieval, you need to store and query vector embeddings. Given that the system needs to handle billions of high-dimensional embeddings while maintaining low latency for search queries, you are evaluating the use of a vector database. Which of the following databases would be the most appropriate choice for this purpose, and why?

A. A document-based NoSQL database like MongoDB, utilizing full-text search capabilities.
B. A graph database like Neo4j, which is designed for traversing relationships between data points. C.

A vector database like Pinecone or Weaviate that supports approximate nearest neighbor (ANN) search.
D. Relational databases with B-tree indexes.

Answer: C

Question: 19

You are working on a project that involves deploying a series of prompt templates for a large language model on the IBM Watsonx platform. The team has requested a system that supports prompt versioning so that updates to the prompts can be tracked and tested over time. Which of the following is the most important consideration when planning prompt versioning for deployment?

A. Prompts should be stored in a proprietary IBM format, as other formats are not compatible with the Watsonx platform when using versioning.
B. The versioning system should automatically downgrade to the previous prompt version if the model returns a confidence score below a certain threshold during inference.
C. Version control should focus exclusively on the syntactical structure of the prompts, as changes to prompt content rarely impact the model's performance.
D. Each version of the prompt must have a unique identifier that can be referenced during model inference, to avoid conflicting results from different prompt versions.

Answer: D

Question: 20

You are designing a generative AI model to generate customer support responses. During testing, you notice that the model frequently outputs gendered language when referring to certain professions, reinforcing stereotypes. Which of the following strategies would most effectively reduce bias in the model's responses?

A. Increase the diversity of the dataset used to train the model, ensuring that all professions are equally represented.
B. Reduce the maximum token limit so that the model generates shorter responses, minimizing the chance for bias.
C. Train the model with a lower learning rate to make it less sensitive to biased patterns in the data. D. Apply a post-processing filter that removes any gendered language after the model generates the response.

Answer: A

Question: 21

While customizing an LLM in InstructLab to generate more human-like responses for a customer service chatbot, you notice that the responses are too formal and lack empathy. Which of the following techniques will best address this problem and help tailor the model to generate more empathetic responses?

A. Use prompt engineering to guide the model towards empathetic responses
B. Change the decoder strategy from greedy decoding to beam search to increase response quality

C. Apply transfer learning with a dataset containing casual language
D. Adjust the model's max sequence length to encourage longer responses

Answer: A

Question: 22

You are tasked with designing a prompt to translate a sentence from English to French using an AI model. Which of the following prompt would best guide the AI to achieve accurate translation, while maintaining cultural nuance and avoiding literal word-for-word translation?

A. "Translate 'The weather is nice today' to French but ensure that the translation reflects word-for-word accuracy and no cultural considerations."
B. "Explain the meaning of 'The weather is nice today' in French."
C. "Translate the sentence 'The weather is nice today' into French and make sure to avoid literal translation, focusing on cultural nuances."
D. "Translate the following sentence from English to French: 'The weather is nice today.'"

Answer: C

Question: 23

You are tasked with generating synthetic data for a fine-tuning task on an IBM watsonx model. The goal is to mimic the distribution of existing training data while ensuring the synthetic data maintains its statistical similarity to the original. You are provided with two algorithms, Algorithm A (Kolmogorov-Smirnov Test) and Algorithm B, to assess the similarity between the original and synthetic data distributions. Which of the following best describes how you should implement synthetic data generation using the User Interface and choose the correct algorithm?

A. Use Algorithm A (Kolmogorov-Smirnov Test) to compare the original and synthetic data distributions, checking for deviations across the entire data range.
B. Use Algorithm A (Kolmogorov-Smirnov Test) to match the covariance matrix of the original and synthetic data distributions, ensuring high correlation between data points.
C. Use the User Interface to generate synthetic data and validate it using Algorithm A, which compares the distributions' mean values to ensure close alignment.
D. Use the User Interface to generate synthetic data and validate it using Algorithm B, which assesses the overall shape of the distributions but does not provide a significance test for statistical similarity.

Answer: A

Question: 24

You are building a generative AI model to assist with customer service responses. During evaluation, you notice that the responses generated tend to favor one specific demographic group, showing bias toward certain dialects and cultural references. How should you adjust the prompt and model parameters to reduce this bias?

A. Use a prompt that explicitly asks for neutrality across demographic groups.
B. Incorporate additional training data from underrepresented demographic groups.

C. Switch to using deterministic (greedy) decoding to ensure more consistent outputs
D. Lower the temperature to reduce randomness in the model's response.

Answer: B

Question: 25

In IBM Watsonx's Prompt Lab, you are refining a prompt to improve the clarity and relevance of the AI's responses. You need to understand which prompt editing options are available to optimize your results. Which of the following is NOT an available prompt editing option?

A. Adjusting the context window to include or exclude specific sections of input text.
B. Setting conditions within the prompt to handle different scenarios based on detected input patterns.
C. Using tone adjustments to modify the emotional tone or style of the AI's responses.
D. Adding dynamic variables to the prompt, allowing for flexible and context-specific responses.

Answer: C

Question: 26

You are developing a generative AI application using LangChain, and you want the system to perform actions like searching a database or retrieving live web content based on a user's request. How can you best incorporate tools in LangChain to enable the AI to perform such tasks autonomously?

A. Rely on LangChain's memory module to remember previous user queries and provide real-time data access.
B. Build a LangChain chain that uses user inputs to sequentially call all the available tools and pick the one with the most relevant output.
C. Use a LangChain agent with a predefined set of tools to dynamically select and invoke the appropriate tool (e.g., database access, API call) based on the user's request.
D. Configure LangChain to automatically load data from static sources based on historical query patterns, avoiding the need for dynamic tool selection.

Answer: C

Question: 27

You are designing a workflow using watsonx.ai to generate complex text summaries from multiple sources. To achieve this, you plan to implement a LangChain-based chain that orchestrates different generative AI tasks: document retrieval, natural language processing (NLP) analysis, and summarization. What is the best way to structure the LangChain-based chain to ensure that each task is effectively handled and results in an accurate summary?

A. Start with NLP analysis, pass the data to watsonx.ai for summarization, and then perform document retrieval to verify the accuracy of the summary.
B. Break the LangChain-based chain into individual steps that allow for manual intervention at each stage, ensuring control over the process at every step.
C. Use watsonx.ai to generate a summary immediately, and then perform NLP analysis and document

retrieval in parallel to verify the accuracy of the output.
D. Perform document retrieval first, followed by NLP analysis to extract relevant information, and then pass the processed data to watsonx.ai for summarization.

Answer: D

Question: 28

You are building a generative AI system that uses synthetic data to mimic an existing dataset. You have learned about two primary algorithms: one that focuses on ensuring the synthetic data passes statistical normality tests and another designed to generate realistic-looking data without focusing on distribution conformity. Which algorithm should you choose if your primary concern is statistical accuracy and passing the Anderson-Darling test?

A. Anderson-Darling Based Synthetic Data Generation (ADS-DG)
B. Gaussian Mixture Models (GMMs)
C. K-Nearest Neighbors (KNN)
D. Bootstrapping Algorithm

Answer: A

Question: 29

In which of the following scenarios would zero-shot prompting be more effective than few-shot prompting when interacting with a generative AI model?

A. When the goal is to adjust the model's response based on few labeled examples that help refine its predictions.
B. When the model is expected to perform a novel task it has never seen, but the prompt can include several examples for guidance.
C. When the prompt is designed for a general task like summarizing a text, which the model is pre-trained on.
D. When the task requires highly domain-specific knowledge that the model has not been exposed to before.

Answer: C

Question: 30

You are working with IBM Watsonx and need to generate synthetic data to improve your model's performance on a custom domain-specific task. After importing a dataset, you want to use the User Interface to generate this synthetic data. What is the primary benefit of using synthetic data generation in fine-tuning your model?

A. It automatically anonymizes sensitive data points to comply with data privacy regulations during the synthetic data generation process.

B. It improves the model’s generalization by exposing it to a wider variety of data points and scenarios.
C. It creates a larger training dataset by duplicating and randomizing the existing data, which enhances

model accuracy.
D. It eliminates the need for any human intervention in the fine-tuning process.

Answer: B

Question: 31

Which of the following decoding strategies would most likely result in generating creative and diverse text outputs while minimizing repetition, when using a generative AI model?

A. Greedy Decoding
B. Beam Search Decoding with a small beam size (e.g., 2)
C. Nucleus Sampling (Top-p) with p = 0.9
D. Temperature Sampling with temperature = 0.0

Answer: C

Question: 32

You are building a customer support chatbot using IBM watsonx.ai and Watson Assistant. The chatbot must use watsonx.ai's large language model (LLM) to generate dynamic responses and Watson Assistant to manage dialog and interaction flow. What is the most efficient way to integrate these two services to deliver an optimal solution?

A. Deploy watsonx.ai's LLM within Watson Assistant by embedding the LLM directly into the Watson Assistant environment.
B. Use Watson Assistant as the primary interface and call watsonx.ai's LLM through an API for generating dynamic responses in specific intents.
C. Use Watson Assistant to directly generate all responses, bypassing watsonx.ai's LLM.
D. Build a separate microservice for each service, allowing Watson Assistant and watsonx.ai's LLM to operate independently, with no communication between them.

Answer: B

Question: 33

You are optimizing a prompt-tuned LLM for a financial institution's automated assistant. The assistant's main tasks include responding to customer inquiries about account balances, providing detailed transaction histories, and explaining complex financial products. Which task should be prioritized for prompt-tuning to improve the model's performance in this domain?

A. Focus on improving the model's ability to generate financial market forecasts.
B. Train the model to generate creative financial advice tailored to each customer.
C. Fine-tune the model for information retrieval tasks related to customer accounts.
D. Optimize for natural language generation to improve customer engagement.

Answer: C

Question: 34

In the context of prompt engineering for IBM Watsonx Generative AI, which of the following is the most accurate description of a prompt variable?

A. A prompt variable is a fixed string that the AI uses to refine its generative process for more context-aware responses.
B. A prompt variable is a function that allows real-time feedback from the model to modify the prompt after generation.
C. A prompt variable is a predefined input that alters the architecture of the AI model during runtime.
D. A prompt variable is a placeholder within a prompt template that can be replaced with specific input values during execution.

Answer: D

Question: 35

When designing a generative AI system to minimize the risk of producing hate speech or abusive content, which of the following strategies in prompt engineering is the most effective?

A. Use offensive words in prompts to test the model's robustness in handling them.
B. Implement real-time content moderation at the output level rather than relying on input prompts alone.
C. Use strict prompt filtering to block any input that contains offensive words.
D. Ensure that the model is trained on a larger dataset, even if the dataset contains diverse viewpoints, including some controversial opinions.

Answer: B

Question: 36

You are tasked with optimizing a generative AI model's usage in a chatbot that provides troubleshooting instructions for software issues. The current prompt template is:
"Please provide step-by-step troubleshooting instructions for the following issue: [Issue Description]. Be detailed, include specific commands or settings the user should check, and provide potential reasons for failure."
To reduce the token count and ensure cost efficiency, which of the following prompt template modifications would best manage token usage while preserving essential information?

A. "Provide detailed troubleshooting instructions for the issue: [Issue Description], with steps, commands, and potential reasons for failure."
B. "Give troubleshooting instructions for [Issue Description], including steps, commands, and reasons for failure."
C. "Troubleshoot the following issue: [Issue Description]. Offer step-by-step commands and reasons for failure."
D. "Provide step-by-step instructions for troubleshooting the issue: [Issue Description]. Include

commands and reasons for failure."

Answer: D

Question: 37

You are working on generating synthetic training data using IBM InstructLab to supplement a small dataset for a question-answering system. Which strategy would most effectively enhance the dataset without introducing biases or artifacts?

A. Use prompts that closely mimic the structure and semantics of the real dataset's questions to maintain consistency.
B. Automatically generate synthetic data using a different model architecture than the one being fine-tuned.
C. Manually tweak each generated response to ensure it's free of errors and aligns with the intended task.
D. Generate a large amount of synthetic data by directly feeding the model with random prompts, ensuring data diversity.

Answer: A

Question: 38

When optimizing a prompt-tuned model, which parameter adjustment would most likely help prevent overfitting without negatively impacting the model’s ability to generalize to unseen prompts?

A. Removing weight decay entirely
B. Increasing the model’s dropout rate
C. Reducing the model’s learning rate
D. Increasing the model’s learning rate

Answer: B

Question: 39

You are experimenting with a generative AI model to write a personalized email response template. You want to ensure that the output maintains a formal tone but occasionally produces creative phrasing without making nonsensical sentences. You are advised to adjust the top-p (nucleus sampling) parameter. Which of the following settings would most effectively balance between formal coherence and occasional creativity in the generated output?

A. Set top-p to 0.5
B. Set top-pto 0.95
C. Set top-p to 1.0
D. Set top-p to 0.0

Answer: B

Question: 40

You are developing a Retrieval-Augmented Generation (RAG) system using IBM WatsonX LLM and a

vector database. Your dataset consists of long legal documents, and you want to ensure the system retrieves the most relevant sections of these documents efficiently. Which of the following best describes the appropriate approach to text chunking for this RAG implementation?

A. Chunking the documents at arbitrary points, ignoring sentence or paragraph boundaries to enhance retrieval speed.
B. Splitting the documents into smaller chunks based on logical or semantic breaks such as paragraphs, while maintaining a token count that matches the LLM's context window.
C. Splitting the legal documents into fixed-size chunks of 10,000 tokens each to maximize retrieval accuracy.
D. Chunking the documents based solely on page numbers, as legal documents typically follow consistent formatting.

Answer: B

Question: 41

You are deploying a large language model in a financial advisory platform to assist users in making investment decisions. Which of the following represent significant risks that should be mitigated before full deployment? (Select two)

A. The model occasionally generates offensive or inappropriate content when responding to user queries.
B. The model generates recommendations that align with historical financial trends but fail to account for recent economic disruptions.
C. The model is trained on open-source financial data, which results in slower response times during inference.
D. The model provides longer-than-expected responses, potentially causing user frustration and increasing abandonment rates on the platform.
E. The model offers speculative advice without indicating the associated level of uncertainty, which may mislead inexperienced investors.

Answer: B,E

Question: 42

You are developing a machine learning pipeline using IBM watsonx that includes fine-tuning an LLM with a dataset containing sensitive personal information. To ensure privacy, you decide to apply differential privacy. Which of the following actions is most critical to configure in the user interface to meet the differential privacy requirements during model fine-tuning?

A. Increase the learning rate and batch size to maximize the noise added by differential privacy algorithms.
B. Remove differential privacy settings for fine-tuning, but apply them in the final inference model to reduce performance degradation.
C. Apply a differential privacy mechanism that adds calibrated noise to both the model updates and synthetic data generation process.
D. Use synthetic data only, which eliminates the need for differential privacy as it does not contain real

user information.

Answer: C

Question: 43

IBM Watsonx Tuning Studio allows users to fine-tune pre-trained models for their specific use cases. Which of the following correctly describes the primary benefits of using Tuning Studio for optimizing a generative AI model?

A. It fully retrains the base model from scratch, ensuring the highest possible accuracy for each new task, regardless of prior training.
B. It allows users to add new architectural layers to the model to improve accuracy without retraining the entire model.
C. It significantly reduces the computational costs associated with model fine-tuning by only updating the model's parameters relevant to the specific task, preserving the general knowledge of the base model.
D. It enables on-the-fly model optimization during inference, adjusting model weights dynamically based on real-time data input.

Answer: C

Question: 44

Which of the following best describes the process of large-scale iterative alignment tuning in the context of customizing LLMs with InstructLab?

A. Repeated fine-tuning of a model using reinforcement learning, focusing on aligning its outputs with human preferences across a diverse set of tasks
B. Fine-tuning the model exclusively on binary classification tasks to improve its generalization on all other tasks
C. Direct training of the model on an expanded version of the dataset, without adjusting prompts or training tasks
D. A single training run of the model on a dataset to generate better predictions for a fixed number of prompts

Answer: A

Question: 45

You are tasked with designing prompts for an IBM Watsonx Generative AI model to minimize hallucinations in responses. One of the ways to reduce hallucinations is by improving the quality of the prompt to guide the model more effectively. Which of the following prompt engineering strategies would be most effective in reducing the likelihood of hallucinations?

A. Use highly abstract and open-ended prompts to allow the model more freedom in generating responses.
B. Include explicit instructions and specific constraints within the prompt to limit the scope of the model's generation.

C. Increase the temperature parameter to introduce more diversity and creativity into the model's output.
D. Set the minimum token length high to ensure the model has enough time to fully develop its response.

Answer: B

Question: 46

You are tasked with generating high-quality responses from a large language model for a customer support application. You want to minimize the amount of provided examples while ensuring that the model generates relevant and specific answers. Which of the following statements best differentiates between zero-shot and few-shot prompting in this context? (Select two)

A. In zero-shot prompting, the model's response is generated purely based on pre-trained knowledge and the structure of the task, while in few-shot prompting, the examples provided offer the model additional context.
B. Zero-shot prompting is better suited for tasks requiring domain-specific knowledge, while few-shot prompting is better for general knowledge tasks.
C. Few-shot prompting involves fine-tuning the model on a specific dataset before generating output, whereas zero-shot prompting uses pre-trained knowledge without additional fine-tuning.
D. ct selection
E. Zero-shot prompting does not require any examples in the input prompt, while few-shot prompting uses a limited number of examples to guide the model's response.
F. Few-shot prompting improves model performance for unfamiliar tasks by fine-tuning weights based on examples, while zero-shot prompting leaves the model weights unchanged.

Answer: A

Question: 47

When analyzing the results of a prompt tuning experiment, which two of the following actions are most appropriate if you observe a consistently high variance in model predictions across different prompt templates? (Select two)

A. Enable regularization techniques like dropout
B. Increase the batch size during training
C. Tune the prompt templates further by standardizing the structure
D. Increase the number of training samples used for tuning
E. Add more layers to the model to increase complexity

Answer: C,D

Question: 48

You are generating a list of items using IBM watsonx's generative AI, but you notice that the model sometimes cuts off mid-sentence when using a stop sequence. What could be the best approach to ensure that the model finishes generating complete sentences while also stopping after a specific sequence is reached?

A. Increase the token limit to avoid premature cut-off
B. Set the stop sequence to a punctuation mark like ";"
C. Use a more distinct and unlikely stop sequence, such as "<END>"
D. Use multiple stop sequences, including a period "."

Answer: C

Question: 49

In the context of the decoding process for generative AI models in IBM Watsonx, what is the main characteristic of greedy decoding?

A. Greedy decoding selects the highest probability token at each step, leading to deterministic and often coherent outputs.
B. Greedy decoding alternates between high and low probability tokens, ensuring a balance between creativity and correctness.
C. Greedy decoding generates multiple possible sequences and selects the most grammatically correct one based on predefined rules.
D. Greedy decoding always selects the token with the lowest probability to encourage diversity in the generated response.

Answer: A

Question: 50

When debating the drawbacks of soft prompts in a generative AI application, which of the following is the most significant challenge compared to hard prompts?

A. Soft prompts introduce more complexity during the training phase, as the model must learn embeddings that are not inherently interpretable by humans.
B. Soft prompts significantly limit the flexibility of the model because they are tied to specific tasks, unlike hard prompts which can generalize to various scenarios.
C. Soft prompts require more human intervention during generation because they depend on predefined patterns and rules for guidance.
D. Soft prompts offer simpler debugging processes because the learned embeddings are directly linked to specific model behaviors and outputs.

Answer: A

Question: 51

You are tasked with fine-tuning a language model using a prompt-tuning approach on a dataset consisting of customer service chat logs. The goal is to optimize the model's ability to generate polite and

contextually appropriate responses. Which of the following steps are essential when preparing the dataset for prompt-tuning in this context? (Select two)

A. Remove any conversations that contain excessive user slang or misspellings.
B. Separate the dataset into training, validation, and test subsets.

C. Convert all user queries into lowercase to reduce noise in the dataset.
D. Ensure each conversation includes both customer input and agent response as context for the model.
E. Ensure all examples in the dataset follow the exact same input-output format.

Answer: B,D

Question: 52

After tuning a generative AI model to produce more concise legal document summaries, you notice that while the summaries are accurate, they tend to be overly verbose. The tuning report shows that the model's perplexity is relatively high, suggesting that it is struggling with token prediction uncertainty, possibly due to an overly complex output format. Which of the following tuning parameters would you most likely adjust to address the verbosity issue without reducing accuracy?

A. Increase the number of epochs
B. Decrease the learning rate
C. Increase the maximum token length
D. Decrease the temperature

Answer: D

Question: 53

While optimizing the cost of running a Generative AI model, you are instructed to adjust the prompt structure. Which of the following changes to a prompt would most reduce computational costs while still maintaining effective results?

A. Including multiple tasks in a single prompt to maximize efficiency.
B. Switching from a narrative-style prompt to a bulleted list format.
C. Breaking complex prompts into simpler, sequential prompts.
D. Using stop tokens early in the prompt to minimize generation length.

Answer: D

Question: 54

While developing a Retrieval-Augmented Generation (RAG) system using the transformers library, you want to improve the retrieval quality by ensuring that your queries and documents are represented in the same latent space for effective similarity matching. Which of the following techniques would be the most appropriate to ensure this alignment between queries and documents?

A. Use a randomly initialized transformer model to encode both documents and queries for unbiased similarity calculation.
B. Use different transformer models for documents and queries, and normalize their embeddings to align them in the same latent space.
C. Fine-tune a transformer model on a document-query similarity task, so that both queries and documents are encoded into the same vector space for retrieval.

D. Use a pre-trained BERT model to encode the documents and a pre-trained GPT model to encode the queries, ensuring diversity in embeddings.

Answer: C

Question: 55

You are designing a Retrieval-Augmented Generation (RAG) system that will handle real-time queries from users, using a combination of a retriever and a transformer-based generator. Which of the following implementation details is the most critical to ensure that the system delivers responses in a timely manner while maintaining accuracy?

A. The retriever should use exact match algorithms to minimize retrieval time and complexity.
B. The retriever and generator should operate independently of each other to avoid communication overhead.
C. The system should cache previous queries and responses to avoid invoking the retriever and generator multiple times.
D. The retriever should utilize an efficient vector search algorithm with approximate nearest neighbor (ANN) techniques to balance speed and accuracy.

Answer: D

Question: 56

You are developing an AI-driven application using IBM watsonx and LangChain to automate legal document summarization for a law firm. The application needs to extract key legal points, summarize them, and generate insights from various sources, including external APIs, court databases, and private document repositories. You are tasked with creating a LangChain chain that integrates these sources, customizes prompt templates, and uses Large Language Models (LLMs) to provide legal summaries. The prompt template must allow for dynamic insertion of text from external sources and adapt based on the type of legal document. Which LangChain chain design would best meet the needs of this application?

A. Use a SequentialChain that first extracts text from external APIs and databases, processes it through custom prompt templates, and then sends the final processed text to an LLM.
B. Employ a Retrieval-Augmented Generation (RAG) Chain, where the LLM queries external knowledge sources in real-time while applying a fixed prompt template.
C. Design a ParallelChain where the text from different sources is processed in parallel by multiple LLMs, combining the results at the end.
D. Implement a SimpleChain that retrieves the required data from external APIs and directly sends the text to the LLM without prompt templates.

Answer: A

Question: 57

You are optimizing a large language model (LLM) for deployment on edge devices with limited computational resources. To reduce the model size and improve efficiency without significantly compromising performance, which of the following quantization techniques is most appropriate for this

scenario?

A. Post-training 8-bit integer quantization
B. 32-bit floating point quantization with fine-tuning
C. Binary quantization (1-bit)
D. Post-training 16-bit floating point quantization

Answer: A

Question: 58

You are using IBM's Tuning Studio to fine-tune a generative AI model for a custom text classification task. The model was pre-trained on a large corpus but shows suboptimal performance when applied to your domain-specific data. You aim to improve both accuracy and computational efficiency. Which of the following is a primary benefit of using Tuning Studio to optimize this model?

A. Tuning Studio allows for the customization of training data at runtime without needing preprocessing.
B. Tuning Studio helps reduce overfitting by applying regularization techniques during the fine-tuning process.
C. Tuning Studio provides detailed performance analytics that allow you to adjust hyperparameters in real-time.
D. Tuning Studio automatically generates prompt templates that can be used for different tasks without further configuration.

Answer: C

Question: 59

Which of the following describes a key benefit of using Prompt Lab in IBM Watsonx for developing generative AI applications?

A. Prompt Lab provides automatic optimization of model hyperparameters, ensuring the best performance without user intervention.
B. Prompt Lab enables real-time collaboration between developers, allowing them to modify prompts simultaneously for faster development.
C. Prompt Lab ensures that all generated outputs are free from bias by automatically filtering inappropriate content.
D. Prompt Lab allows users to iteratively test and refine prompts in a controlled environment, improving prompt effectiveness and reducing guesswork.

Answer: D

Question: 60

Which of the following stopping criteria can help in generating coherent and well-structured text without cutting off mid-sentence or continuing unnecessarily?

A. Stopping when the model generates special end-of-sequence tokens, such as <EOS>
B. Stopping the model only when it reaches the end of a predefined phrase from the input prompt

C. Monitoring the likelihood of the next token and stopping when the likelihood drops below a threshold
D. Stopping the model after a predetermined number of tokens, regardless of context

Answer: A

Question: 61

You are fine-tuning a generative model to generate text-based responses in a customer service chatbot. You want to ensure the responses are concise and relevant, without causing the model to produce overly long or irrelevant output. Which of the following parameters and stopping criteria would be most effective for achieving this goal?

A. Increase the temperature to 1.5 and set a high maximum token limit.
B. Use beam search decoding with a low beam width and a high repetition penalty.
C. Use greedy decoding with no repetition penalty and a high stopping probability threshold.
D. Set a low top-k value and implement a repetition penalty with a low maximum token limit.

Answer: D

Question: 62

Which of the following statements accurately describes a drawback of using soft prompts in generative AI model optimization?

A. Soft prompts can increase the model's interpretability by providing clear, user-defined input instructions.
B. Soft prompts require additional computational resources during training, which can limit their scalability in real-time applications.
C. Soft prompts offer improved performance for specific tasks but are harder to implement when fine-tuning models across multiple domains.
D. Soft prompts make it easier to control the model's behavior as the prompts are flexible and can be adjusted by the user during inference.

Answer: B

Question: 63

When setting up a tuning experiment in IBM watsonx's Tuning Studio, which of the following best describes the process for optimizing a model's hyperparameters?

A. Set the learning rate to its maximum value to speed up the tuning process and reduce experimentation time.
B. Manually adjust one hyperparameter at a time while keeping all other parameters constant to precisely identify its impact on performance.
C. All hyperparameters should be fixed at the default settings to ensure consistency across different experiments.
D. Use automated hyperparameter search techniques such as grid search or random search to explore multiple configurations efficiently.

Answer: D

Question: 64

You are tasked with creating a prompt template for IBM Watsonx to generate customer support responses based on user queries. The response needs to be polite, concise, and address the issue directly. Which of the following is the most appropriate structure for a reusable prompt template to ensure consistency across multiple queries?

A. "Generate a detailed and formal response to the customer, focusing on providing as much information as possible, even if it's unrelated to the query."
B. "Please write a polite and professional response to the customer's query, including any relevant context or background information and focusing on the core issue."
C. "Generate a professional response to the customer's query, avoiding repetition and unnecessary details, while focusing on addressing the issue succinctly."
D. "Write a short and casual response to the customer, focusing on being friendly and engaging, regardless of the content of the query."

Answer: C

Question: 65

You are developing a generative AI model using the IBM Watsonx platform to assist in customer service. While the model's responses are highly accurate, there is concern that the model may inadvertently expose personal information (PII) or sensitive data during interactions. As a responsible AI engineer, it is crucial to mitigate this risk. Which of the following is the most critical risk associated with the exposure of personal information in generative AI models?

A. The model can unintentionally memorize and regurgitate personal information from the training data, leading to privacy violations.
B. The model can generate outputs that are too general, failing to meet the specific needs of the user.
C. The model might produce content that doesn't align with the cultural preferences of the user.
D. The model can generate overly creative or non-factual responses, leading to brand reputation damage.

Answer: A

Question: 66

In the context of Tuning Studio in IBM watsonx, what is one of the key benefits of using Compute Unit Hours (CUHs) during the fine-tuning process?

A. It provides real-time feedback on model deployment success rates.
B. It allows for the precise allocation of computational resources to manage budget constraints.
C. It limits the number of model versions stored, improving system performance.
D. It reduces the time required to train models by lowering the accuracy threshold.

Answer: B

Question: 67

While working with IBM Watsonx to generate synthetic data, you import a sensitive dataset containing personally identifiable information (PII). You are tasked with anonymizing the imported data before proceeding with any fine-tuning or data augmentation. Which of the following steps is the most appropriate to ensure proper anonymization?

A. Generate a synthetic version of the dataset that removes all PII automatically.
B. Apply a differential privacy algorithm that ensures no individual data point can be traced back to a specific user.
C. Randomly shuffle the sensitive data fields to prevent direct re-identification.
D. Use a hashing algorithm to replace PII fields while retaining the ability to reverse the process if needed.

Answer: B

Question: 68

Condition-based prompts, where specific actions are taken depending on input patterns, are part of advanced prompt design, allowing developers to create more context-aware interactions.

A. The temperature parameter controls the length of the generated output by increasing or decreasing the model's word count limit.
B. The learning rate parameter adjusts the creativity of the model's outputs by encouraging the model to explore more diverse topics.
C. The greedy decoding parameter improves output diversity by ensuring that the most likely token is always chosen at each step in the generation process.
D. The top-k sampling parameter controls how many potential next words are considered during each generation step, limiting the randomness of the output.

Answer: D

Question: 69

You are preparing a dataset for fine-tuning a model to classify customer complaints by category. The dataset is imbalanced, with 70% of the data representing complaints about billing, 20% representing complaints about technical issues, and 10% representing complaints about product quality. Which of the following actions would help address the imbalance while preparing the dataset for fine-tuning? (Select two)

A. Leave the dataset as-is, trusting the model to handle the imbalance using its internal mechanisms.
B. Apply class weighting during model training instead of modifying the dataset itself.
C. Add more synthetic data for the minority classes by using generative techniques like GPT-3 to create realistic complaints.
D. Oversample the under-represented classes to ensure a balanced distribution across categories.
E. Undersample the majority class (billing complaints) to balance the dataset.

Answer: B,D

Question: 70

In the context of sampling decoding for IBM Watsonx Generative AI, which of the following statements best describes how top-k sampling works?

A. Top-k sampling ensures that the next token is chosen only if it matches one of the predefined input variables.
B. Top-k sampling selects the token with the highest probability, ignoring all other token options.
C. Top-k sampling automatically filters out low-probability tokens that were not part of the model's training set.
D. Top-k sampling selects the next token only from the top k most probable tokens based on their probabilities.

Answer: D

Question: 71

In a Retrieval-Augmented Generation (RAG) system designed for technical document retrieval, you are tasked with implementing text chunking techniques using the LangChain library. The technical documents are large and contain numerous tables, figures, and bullet points. What is the most effective way to handle text splitting to ensure high-quality retrieval?

A. Split the text into equal-sized chunks of 512 characters, regardless of the content structure, to improve consistency in retrieval.
B. Split the text only at paragraph breaks, ignoring tables and figures, as they can be processed separately.
C. Convert tables and figures into plain text and split the document by character count to maintain even chunk sizes.
D. Use a hybrid approach, splitting the text by both semantic boundaries (like paragraphs) and content-specific markers (like bullet points and tables), while keeping chunks within the model's token limit.

Answer: D

Question: 72

In the context of Retrieval-Augmented Generation (RAG), embeddings play a crucial role in ensuring relevant information is retrieved to augment the generative AI's response. Which of the following best describes the role of embeddings in the RAG process?

A. Embeddings represent the search space for the retriever model, allowing the system to retrieve semantically relevant information based on input queries.
B. Embeddings are pre-trained generative models that augment the retrieval step by generating new query variations.
C. Embeddings are only used in fine-tuning generative models and play no role in the retrieval process.
D. Embeddings are used to directly generate the textual responses in the output.

Answer: A

Question: 73

You are optimizing a generative AI model that writes product descriptions. The cost of using the model is directly related to the number of tokens generated. To minimize token usage, you decide to introduce a stop sequence in your prompt that signals the model to end its generation early when the description reaches a certain length. Given the following prompt:
"Write a product description for [Product Name]. The description should include the main features and benefits of the product in no more than 50 words."
Which of the following stop sequences would be most effective in ensuring the generation is concise and does not exceed the desired word limit?

A.
B.
C.
D. ---"
E. "
F. End of description."
G. "###END###"

Answer: G

Question: 74

You are building a customer support chatbot for an e-commerce company using IBM watsonx and LangChain. The chatbot will interact with an external database that holds customer order history, shipping details, and product catalog data. You need to create a LangChain chain that dynamically generates responses using prompt templates tailored to customer queries, retrieves data from the external database, and incorporates LLMs to refine the answers. The goal is to provide accurate, context-aware responses to questions about order status and product details. Which LangChain strategy will best ensure that the chatbot provides accurate, dynamic responses based on real-time customer data?

A. Use a RetrievalChain to query the external database and combine the retrieved data with a dynamic prompt template before sending it to an LLM.
B. Implement a SimpleChain that directly connects the chatbot to the external database and generates responses from pre-defined LLM outputs.

C. Apply a MemoryChain that remembers past customer queries and uses this memory to answer future questions more accurately.
D. Design a ParallelChain where multiple LLMs process different aspects of the customer query, such as order history and product details, combining them in the final answer.

Answer: A

Question: 75

You are developing a document understanding system that integrates IBM watsonx.ai and Watson Discovery to extract insights from large sets of documents. The system needs to leverage watsonx.ai's

large language model to summarize documents and Watson Discovery to search and extract relevant data from those documents. What is the best approach to achieve this integration?

A. Use watsonx.ai's LLM to both retrieve and summarize the documents, bypassing Watson Discovery.
B. Use Watson Discovery for summarizing documents and watsonx.ai's LLM for only retrieving relevant content from the documents.
C. Use watsonx.ai's LLM to create a summary for each document in advance, and Watson Discovery only for searching pre-generated summaries.
D. Use Watson Discovery to index and search documents, and then send the retrieved documents to watsonx.ai's LLM for summarization through API calls.

Answer: D

Question: 76

You are implementing a few-shot prompting strategy with IBM Watsonx to improve the model's performance in generating customer service responses. The goal is to ensure the model understands the tone and format required for polite and concise replies. Which of the following strategies best illustrates the correct way to use few-shot prompting?

A. Provide example prompts with multiple different output styles to give the model a range of responses to choose from.
B. Provide one or two well-structured examples that demonstrate the expected tone and format of the customer service responses within the prompt.
C. Include a large number of examples, typically over 10, in the input prompt to ensure the model learns from diverse cases.
D. Use only negative examples in the prompt to show the model what not to generate in terms of tone and format.

Answer: B

Question: 77

In a Retrieval-Augmented Generation (RAG) setup, you notice that the model is generating responses that are not always relevant to the query, despite the knowledge base containing useful information. What could be the most likely cause of this issue, and how should you address it?

A. The model is over-relying on the retrieval system and ignoring the language model's ability to generate coherent responses, so you should disable the retrieval component for general questions.
B. The knowledge base might contain outdated or irrelevant documents, so removing all non-recent documents would ensure the model generates more relevant responses.
C. The problem likely lies with the input format, so changing all queries to a pre-structured format (like templates) will ensure the retrieval and generation stages perform optimally.
D. The retrieval mechanism might be failing to fetch the most relevant documents from the knowledge base, so you should improve the search algorithm or use a better ranking system.

Answer: D

Question: 78

You are tasked with creating a prompt-tuned model using IBM watsonx.ai to enhance the quality of text generation for customer support. The goal is to fine-tune the model for improved context understanding based on specific customer queries. Which of the following approaches would be the best method to initialize the prompt for tuning?

A. Use a pre-trained general-purpose prompt with no domain-specific customization
B. Use a manually crafted prompt tailored to the specific context of customer support queries
C. Construct a prompt using a large set of random tokens from the training corpus
D. Use a prompt with pre-defined output patterns to restrict the model's possible responses

Answer: B

Question: 79

Which of the following is a key component of IBM's InstructLab framework for customizing large language models (LLMs)?

A. Prompt engineering module designed to automatically generate synthetic training data for prompt-tuned models
B. Tools to iteratively optimize the model's alignment with human preferences, such as reinforcement learning from human feedback (RLHF)
C. A tokenization algorithm designed to reduce model size by removing unused tokens
D. A fine-tuning mechanism based on few-shot learning that only updates the model's output layer

Answer: B

Question: 80

After prompt-tuning a language model, you notice that certain outputs are semantically correct but syntactically flawed. Which of the following actions is most appropriate to resolve this issue and optimize the tuned model's performance?

A. Fine-tune the prompt template to emphasize grammar
B. Lower the learning rate during the tuning phase
C. Increase the model's training dataset size
D. Use a higher temperature during the generation process

Answer: A

Question: 81

You have completed a prompt-tuning experiment for a large language model (LLM) using IBM Watsonx, aimed at improving its ability to generate accurate responses to customer support queries. After the tuning process, you are analyzing the performance statistics of the model. Which statistical metric is the most appropriate to prioritize when evaluating the success of the prompt-tuning experiment?

A. Log-likelihood of generated responses
B. BLEU score
C. Perplexity score
D. Token generation speed

Answer: C

Question: 82

Your organization is deploying a generative AI model to assist in legal document generation. During testing, you discover that the model generates biased legal advice that could disproportionately affect certain social groups. Additionally, a team member raises concerns about potential data poisoning attacks on your training set. What steps should you take to mitigate both the risks of data bias and poisoning?

A. Apply prompt engineering techniques to avoid triggering known biases in the model.
B. Use adversarial training techniques to make the model more robust against bias and poisoning.
C. Implement a data validation pipeline to detect anomalies and potential poisoning in the training data.
D. Fine-tune the model using only the training data from trusted sources, without expanding the dataset.

Answer: C

Question: 83

You are fine-tuning a pre-trained language model on a dataset of financial news articles to improve its ability to generate summaries of financial reports. After several epochs of training, you observe that the model performs well on the training data, achieving near-perfect accuracy. However, the model's performance on the validation set is much lower, indicating potential overfitting. What is the most effective adjustment to reduce overfitting while continuing to fine-tune the model?

A. Reduce the size of the training dataset
B. Increase the learning rate
C. Apply dropout during training
D. Increase the number of epochs

Answer: C

Question: 84

You are working with a Watsonx Generative AI model to create marketing content that balances creativity with efficiency. The goal is to generate engaging content within a predefined time limit without compromising on quality. Given this context, which two optimization strategies will most effectively help you achieve both speed and content quality? (Select two)

A. Implement early stopping criteria based on token repetition to avoid lengthy generation.
B. Use beam search with a beam width of 1 to minimize computation time.
C. Increase the model's learning rate to accelerate content generation.
D. Reduce the batch size to decrease training time and speed up generation.
E. Apply top-k sampling with k = 50 to ensure diverse and creative outputs.

Answer: A,E

Question: 85

You are tasked with fine-tuning a large language model (LLM) using IBM's InstructLab to improve performance for a specific customer service task. The goal is to enhance the model's ability to answer questions related to account management and customer complaints. Which of the following actions is NOT a component of the fine-tuning process in InstructLab?

A. Selecting and preprocessing a representative dataset of customer interactions for training
B. Defining specific task instructions that the model will follow during inference
C. Tuning the learning rate to prevent overfitting during the fine-tuning process
D. Directly adjusting the model's architecture to increase the number of attention heads in the transformer

Answer: D

Question: 86

Prompt Lab in IBM Watsonx Generative AI offers several advantages for AI prompt engineering. Which of the following best describes a primary benefit of using the Prompt Lab feature?

A. It guarantees that all generated responses adhere to industry-specific regulatory standards.
B. It provides a collaborative environment where multiple users can co-author prompts in real time.
C. It allows users to design custom AI models from scratch to handle specific tasks.
D. It enables users to test different versions of prompts and receive immediate feedback on their effectiveness.

Answer: D

Question: 87

You are tasked with deploying a versioned prompt for a customer-facing generative AI application. The prompts are iteratively improved based on feedback, and you need to ensure that each version of the

prompt is tracked and accessible for rollback in case a newer version produces worse results. Which strategy would best ensure that all prompt versions are stored and easily retrievable, while minimizing disruption to the current deployment?

A. Deploy prompts directly to production without versioning and manually track changes in a spreadsheet.
B. Leverage a cloud-based deployment pipeline with integrated versioning that automates prompt rollback and audit tracking.
C. Store each version of the prompt as a separate file in a local folder and manually deploy the desired version when needed.
D. Use a version control system like Git to track prompt changes and synchronize the latest version with the deployment environment.

Answer: D

Question: 88

A large language model you are fine-tuning occasionally generates completely fabricated references and citations when responding to user queries. This behavior exemplifies a specific model risk. Which of the following techniques would most effectively reduce this risk in a production environment?

A. Using human-in-the-loop (HITL) methods for real-time validation
B. Increasing the model's response diversity by adjusting top-p sampling
C. Switching to greedy decoding for more deterministic responses
D. Deploying rule-based post-processing filters to validate the output

Answer: D

Question: 89

You are tasked with fine-tuning a pre-trained large language model (LLM) on a custom dataset containing customer support interactions for a company. The dataset contains text with specific categories related to issues such as billing, product returns, technical support, and feature requests. Before training, you need to prepare the dataset for optimal fine-tuning. Which of the following steps is the most crucial to ensure the dataset is prepared effectively for fine-tuning the model?

A. Manually categorize each interaction and organize them into a taxonomy tree structure.
B. Perform a spelling correction on the entire dataset to remove any language inconsistencies.
C. Tokenize the dataset before curating it and mapping it to the taxonomy tree.
D. Convert all text to lowercase to ensure uniformity in the dataset.

Answer: A

Question: 90

You are tasked with optimizing a generative AI model's output for a natural language generation task. Which of the following combinations of model parameters is most appropriate for encouraging creative and varied responses without sacrificing too much coherence?

A. Temperature = 0.7, Top-p = 0.4, Max tokens = 100, Frequency penalty = 0.9, Presence penalty = 0.3
B. Temperature = 1.2, Top-p = 1.0, Max tokens = 250, Frequency penalty = 0.3, Presence penalty = 0.2
C. Temperature = 0.5, Top-p = 0.9, Max tokens = 300, Frequency penalty = 0.8, Presence penalty = 0.7
D. Temperature = 1.5, Top-p = 0.8, Max tokens = 150, Frequency penalty = 0.5, Presence penalty = 0.6

Answer: D

Question: 91

You are fine-tuning a large language model (LLM) for a sentiment analysis task using customer reviews. The dataset is relatively small, so you decide to augment it using IBM InstructLab. Which approach would

be the most effective in generating high-quality synthetic data for this fine-tuning process?

A. Fine-tune IBM InstructLab itself to generate data that closely resembles the training data format, ensuring consistent sentiment distribution.
B. Use IBM InstructLab to generate synthetic data, but only for neutral sentiment, as the model already handles positive and negative sentiment well.
C. Increase the diversity of synthetic data by focusing on outliers and rare sentiment cases that are underrepresented in the original dataset.
D. Use a generic prompt to generate a wide variety of data from IBM InstructLab, regardless of sentiment polarity.

Answer: A

Question: 92

You are implementing a RAG system and have chosen LlamaIndex to handle the document indexing process. Your system needs to retrieve relevant documents quickly and efficiently for large datasets. What is the most important function of LlamaIndex in managing document retrieval?

A. LlamaIndex transforms documents into high-dimensional embeddings and stores them in a vector database to enable fast semantic search.
B. LlamaIndex creates keyword-based indexes of documents, optimizing for exact word matches rather than semantic search.
C. LlamaIndex generates summaries of documents and uses these summaries for quick retrieval rather than the full document.
D. LlamaIndex compresses the documents and stores them in a traditional SQL database to improve retrieval speed.

Answer: A

Question: 93

In the context of analyzing prompt-tuning results, which statistical measure is most important to assess

how well the tuned model generalizes to unseen data?

A. Validation loss
B. Training loss
C. Number of epochs completed
D. Accuracy on the training dataset

Answer: A

Question: 94

Which of the following techniques can be most effectively used to mitigate the generation of hate speech, abuse, and profanity in generative AI models when applying prompt engineering?

A. Applying Token Regularization to limit the diversity of generated responses
B. Fine-tuning the model with specific datasets curated to exclude offensive content

C. Restricting the model's ability to generate certain words or phrases using stop-word lists
D. Using Greedy Decoding to ensure that the model outputs the most likely sequence of tokens

Answer: B

Question: 95

In the context of model quantization for generative AI, which of the following statements correctly describes the impact of quantization techniques on model performance and resource efficiency? (Select two)

A. Quantizing a model to 8-bit precision always results in a significant loss in performance, especially when working with language models or large generative AI architectures.
B. Quantization-aware training (QAT) can help mitigate the accuracy degradation that occurs during quantization by simulating lower precision during the training process.
C. Quantization reduces the precision of model weights and activations, allowing for lower memory usage and faster computation with minimal impact on model accuracy.
D. Post-training quantization is more resource-efficient than quantization-aware training, as it applies quantization after the model has been fully trained, eliminating the need for additional fine-tuning.
E. Quantization can increase the inference time of a model since it adds computational complexity when converting from higher to lower precision formats during runtime.

Answer: B,C

Question: 96

You are tasked with deploying a generative AI solution for a client who operates in the healthcare sector. Due to the sensitive nature of the data, the client requires a highly secure deployment with continuous monitoring for regulatory compliance. Which role is primarily responsible for ensuring the AI solution is compliant with these security and regulatory requirements?

A. Security Engineer

B. Data Privacy Officer
C. AI Model Developer
D. Chief Technology Officer (CTO)

Answer: A

Question: 97

As a Generative AI engineer, you're tasked with optimizing the performance and cost-efficiency of a model by adjusting the model parameters. Given that your objective is to reduce the cost of generation while maintaining acceptable quality, which of the following parameter changes is most likely to result in cost savings?

A. Set the temperature parameter to a higher value.
B. Increase the top-k sampling value.
C. Increase the max tokens parameter to allow for more complex output.

D. Decrease the max tokens parameter.

Answer: D

Question: 98

You are working with a foundation model pre-trained on a large general-purpose dataset, and you plan to deploy it for a specialized task in healthcare-related text generation. However, before tuning the model, you want to assess whether tuning is necessary for your use case. Which of the following is the best indicator that it is time to tune the foundation model for your task?

A. You are noticing that the model occasionally makes grammar mistakes in the generated text.
B. The model performs well on general datasets but fails to capture specific domain-related terminology and context.
C. The model's accuracy is already above 90%, but you want to achieve 95% accuracy for your task. D. The model's inference time is longer than expected, and you need to reduce latency for real-time applications.

Answer: B

Question: 99

When tuning model parameters for a generative AI prompt, which of the following adjustments would most likely increase the model's tendency to generate coherent but less creative responses?

A. Increasing the temperature parameter to 1.5
B. Decreasing the value of the temperature parameter to 0.2
C. Reducing the beam size in beam search from 5 to 1
D. Using Top-k Sampling with a k value of 100

Answer: B

Question: 100

You are building a question-answering system using a Retrieval-Augmented Generation (RAG) architecture. You are deciding whether to incorporate a vector database into the system to handle the document embeddings. Under which of the following circumstances is the use of a vector database most appropriate?

A. When the corpus consists mainly of short, structured text like JSON records and traditional SQL indexing will suffice
B. When the data consists primarily of binary files such as images and videos, and full-text search is required
C. When real-time similarity search over high-dimensional embeddings is needed for large-scale unstructured text data
D. When the text corpus consists entirely of predefined categories that can be handled by simple keyword matching algorithms

Answer: C

Question: 101

You are building a RAG system that integrates LlamaIndex for efficient document retrieval from a knowledge base containing millions of documents. Which of the following is the most critical consideration for ensuring the system performs well at scale?

A. Ensuring LlamaIndex generates unique embedding vectors for each document and storing them in an optimized vector database for quick lookups.
B. Adjusting LlamaIndex to prioritize documents that contain certain predefined keywords, optimizing retrieval for specific business terms.
C. Configuring LlamaIndex to perform real-time indexing of incoming data streams to ensure the most up-to-date documents are always retrieved.
D. Setting LlamaIndex to index only a small subset of the documents to minimize the time it takes to retrieve any relevant document.

Answer: A

Question: 102

You are testing a new version of a prompt template designed to improve the accuracy of responses from a generative model deployed on IBM Watsonx. After deploying the new prompt version, you need to ensure that it performs better or at least as well as the previous version. Which of the following approaches provides the most reliable method for testing the performance of the new prompt template version?

A. Replace the old prompt with the new one in the live system immediately to avoid confusion between prompt versions.
B. Run a series of A/B tests comparing the new prompt template to the old one, using a set of predetermined metrics, such as response accuracy and completion time.
C. Test the new prompt in production without monitoring and observe user feedback to gauge performance.
D. Use a random subset of production data and test both versions in a local environment, as local tests always replicate the conditions of production.

Answer: B

Question: 103

You are tasked with developing a customer support system for an e-commerce platform using the Retrieval-Augmented Generation (RAG) pattern. The system needs to retrieve relevant information from a large database of product specifications, user manuals, and FAQs. You decide to use LangChain for constructing the pipeline and SingleStore as the backend for storing and querying the document embeddings. The objective is to efficiently retrieve semantically similar documents and use them as input for a generative model that crafts human-like responses. Which of the following steps best describes the correct implementation of the RAG pattern using LangChain and SingleStore for this customer support system?

A. Use LangChain to fine-tune the generative model, then store the embeddings in SingleStore, and use SQL queries to retrieve documents based on exact keyword matches.
B. Use LangChain to generate embeddings and directly store the generated responses in SingleStore without any retrieval mechanism.
C. Use LangChain to create an LLM chain that integrates with SingleStore for embedding retrieval, where the retrieved documents are used as context for generating responses.
D. Store the pre-trained generative model's parameters in SingleStore and use LangChain to retrieve embeddings from the database for training the model.

Answer: C

Question: 104

You are deploying a Generative AI solution for a client who needs to generate customer service emails in multiple languages. The client has provided a dataset of historical customer service emails, and they want to ensure that their generative model consistently produces accurate, contextually appropriate responses across different languages. The client also has concerns about the latency of the model's responses. Based on these requirements, you are tasked with planning the deployment of the generative AI solution. Which deployment strategy would be most appropriate for this client's needs, considering latency, language handling, and response quality?

A. Deploy a single multilingual model using beam search decoding, and run the model on a single GPU cluster to ensure consistent responses.
B. Use a fine-tuned model per language, with nucleus sampling and run them on a single GPU instance to reduce the infrastructure cost.
C. Deploy a single multilingual model with tuned prompts for each language, using top-k sampling and leveraging a multi-GPU distributed environment to balance response time and quality.
D. Deploy a tuned prompt for each language on separate models, using greedy decoding to minimize latency, and scale across multiple GPUs for each language.

Answer: C

Question: 105

You are tasked with designing a LangChain-based AI workflow using watsonx.ai that incorporates multiple models for different tasks: document classification, entity extraction, and text generation. The final output should consist of a well-structured report that combines these processes. What is the best strategy to orchestrate this workflow to ensure seamless integration of all tasks and a coherent final output?

A. Build a modular workflow where each task operates independently without passing data between them, ensuring each task can be validated separately before generating the final report.
B. Use document classification to filter the input data, followed by entity extraction to identify key terms, and then pass the refined information to watsonx.ai for text generation.
C. Run the document classification, entity extraction, and text generation tasks in parallel and merge the results at the end to produce the final report.
D. Start with text generation using watsonx.ai and then refine the output through document classification and entity extraction.

Answer: B

Question: 106

You are working on generating creative text responses using IBM watsonx's generative AI model. You need to adjust the output so that it is more diverse and creative without losing coherence. Which of the following model parameter settings would best achieve this objective?

A. Set temperature to 0.2
B. Set temperature to 0
C. Set temperature to 2
D. Set temperature to 1.5

Answer: D

Question: 107

You are designing a prompt to converse with a model for multilingual translation. How would you frame the prompt to translate an English business email to Japanese, ensuring that the translated email is formal and appropriate for a business setting in Japan?

A. "Translate this business email into Japanese but do not consider the tone or formalities."
B. "Translate the following business email into Japanese, focusing on word-for-word accuracy."
C. "Translate this business email into Japanese, making sure the tone is formal and culturally appropriate for a business context."
D. "Translate this business email into Japanese using informal language."

Answer: C

Question: 108

In a system that generates product recommendations using a generative AI model, you are tasked with creating prompts that incorporate dynamic variables to ensure more relevant and personalized responses.
Prompt Template:
"Recommend products for [customer_name] based on their interest in [product_category]."
Which portion of the prompt is the best candidate to be replaced with variables to achieve greater personalization and flexibility? (Select two)

A. "based on their interest"
B. "[product_category]"
C. "[customer_name]"
D. "for"
E. "Recommend products"

Answer: B,C

Question: 109

You are tasked with deploying a watsonx Generative AI solution for a client who requires real-time

inference with minimal latency. The solution will be used for large-scale text generation tasks, where throughput is critical. The client also requires the system to handle fluctuating loads efficiently. What would be the most appropriate deployment strategy?

A. Deploy the model on a single GPU-based cloud instance with auto-scaling enabled.
B. Deploy the model on a multi-node GPU cluster with Kubernetes to handle auto-scaling and fault tolerance.
C. Deploy the model in a hybrid architecture, splitting inference tasks between on-premise servers and cloud resources.
D. Use a CPU-based deployment for cost efficiency, and manually adjust the number of instances as needed.

Answer: B

Question: 110

While working on a fine-tuning project in IBM watsonx, you need to generate synthetic data that mimics the properties of your existing dataset for training purposes. You have two algorithms available: Algorithm A (Kolmogorov-Smirnov Test) and Algorithm B, which uses a different methodology for assessing similarity between original and synthetic data. After generating the synthetic data using the User Interface, what would be the primary consideration in choosing the correct algorithm to validate that the generated data sufficiently mimics the original data?

A. Choose Algorithm A (Kolmogorov-Smirnov Test) if you want to ensure that the original and synthetic data distributions are identical across their entire range, not just at specific points.
B. Choose Algorithm A (Kolmogorov-Smirnov Test) if the synthetic data has categorical variables, as this test is specifically designed for discrete distributions.
C. Use Algorithm B to compare the entropy of the original and synthetic data distributions, ensuring both data sets have the same level of uncertainty.
D. Use Algorithm B if you want to focus on minimizing the difference in mean squared error (MSE) between original and synthetic data distributions.

Answer: A

Question: 111

A generative AI model designed for healthcare content generation is being evaluated for ethical risks. The model tends to give preference to certain demographic groups when recommending treatments. What is the most effective method to identify and mitigate this bias during the prompt engineering phase?

A. Limit the model’s context window to prevent it from over-relying on demographic information.
B. Use adversarial debiasing techniques to adjust the model's internal representations during training.
C. Adjust the temperature to 1.0 to ensure the model generates more balanced and less biased outputs.
D. Train the model on a smaller dataset that excludes demographic information, to remove bias from its learned patterns.

Answer: B

Question: 112

In the context of avoiding abusive or profane content generation, which of the following prompt engineering techniques is most likely to reduce the risk of model misuse?

A. Enforce the use of strictly neutral or factual language in all prompts.
B. Ask users to label their own prompts as "safe" or "unsafe" before passing them through the model.
C. Add disclaimers to the prompt input, asking the model to avoid producing harmful or offensive content.
D. Utilize sentiment analysis on generated outputs to detect harmful language and re-prompt the model if necessary.

Answer: D

Question: 113

In the context of generative AI and large language models, text embeddings are a key component. What is the primary purpose of text embeddings in a retrieval-augmented generation (RAG) system, and how are they used?

A. Text embeddings generate a one-to-one mapping of words, allowing for the exact reconstruction of the original input text.
B. Text embeddings are used to randomly assign values to words in the corpus, providing variability in how the model generates text.
C. Text embeddings are used to reduce the dimensionality of the input text for faster training of the generative model.
D. Text embeddings map words and phrases to high-dimensional vectors that capture their semantic meaning, allowing for efficient document retrieval based on contextual similarity.

Answer: D

Question: 114

You are developing a Retrieval-Augmented Generation (RAG) system to enhance the responses of a legal chatbot by integrating it with a vast legal document repository. You are using LangChain to build the pipeline, Watson ML for model hosting, and Elasticsearch as your document store. What would be the most appropriate approach for combining these components into a RAG pipeline?

A. Use Elasticsearch for document retrieval -> Use LangChain to encode the documents -> Generate the response using Watson ML.
B. Use LangChain to chain together query encoding, document retrieval from Elasticsearch, and Watson ML for response generation.
C. Use LangChain to pre-process documents -> Use Elasticsearch for model storage -> Use Watson ML to retrieve documents and generate responses.
D. Use Watson ML for document retrieval and response generation -> Use Elasticsearch to store model responses -> Use LangChain for chaining the responses together.

Answer: B

Question: 115

You are tasked with configuring a generative model to assist legal professionals by generating contract clauses. The model should produce complete, contextually relevant clauses but should avoid overly long and redundant text. You must ensure that the generation process stops appropriately once a clause is completed. Which of the following strategies is the most appropriate for configuring stopping criteria?

A. Set a maximum token limit of 500 and implement a fixed length stopping criterion.
B. Use a beam search algorithm and impose a strict length restriction of 100 tokens.
C. Use a stop sequence based on legal clause delimiters and impose a minimum token limit of 50.
D. Set the top-k value to 1 and use the default maximum token limit, relying on the model's internal stopping criteria.

Answer: C

Question: 116

What is one primary advantage of using Prompt Lab in IBM Watsonx when evaluating prompt variations for a generative AI model?

A. Prompt Lab enables side-by-side comparisons of multiple prompt variations, allowing developers to evaluate different formulations in a systematic manner.
B. Prompt Lab automatically generates prompts based on the dataset, requiring minimal manual input from developers.
C. Prompt Lab automatically selects the best performing prompt based on predefined metrics without any user input.
D. Prompt Lab allows users to lock the output for each prompt, ensuring deterministic results across different runs.

Answer: A

Question: 117

You are fine-tuning a large language model (LLM) with InstructLab to generate high-quality summaries for long-form text documents. After conducting an initial experiment, the performance seems suboptimal, particularly on technical documents with specialized vocabulary. What steps could you take to improve the fine-tuning process to achieve better performance? (Select two)

A. Perform prompt engineering by providing specific task instructions for summarization.
B. Increase the learning rate for better adaptation to specialized data.
C. Experiment with a smaller model to reduce complexity and increase inference speed.
D. Use greedy decoding during inference to generate concise summaries.
E. Add more domain-specific training data to improve the model's understanding of technical vocabulary.

Answer: A,E

Question: 118

You are tasked with developing a system that uses a vector database to store embeddings generated from a large corpus of documents. The system should be able to perform fast and efficient nearest neighbor search while balancing accuracy and speed. Given the large volume of data and the need for scalability, you are considering different indexing strategies offered by vector databases. Which of the following indexing techniques is the most appropriate for balancing search accuracy and speed in high-dimensional vector space, and why?

A. Exact nearest neighbor (ENN) search using KD-trees.
B. Rely on full-text indexing of documents and avoid vector search altogether.
C. Approximate nearest neighbor (ANN) search using HNSW (Hierarchical Navigable Small World) graphs.
D. Linear search over the entire vector dataset.

Answer: C

Question: 119

When using IBM Watsonx Tuning Studio, what is the recommended approach to determining the number of training data examples required for effective model fine-tuning?

A. Use at least 50% of the original training data to ensure the fine-tuned model generalizes well across both new and existing tasks.
B. Use a minimum of 1,000 to 5,000 examples for each task, but focus on the quality and relevance of examples rather than quantity.
C. Use at least 10,000 examples for each unique task to ensure the model retains its general knowledge and effectively adapts to the new task.
D. Use no more than 100 examples per task to avoid overwhelming the model's general capabilities with task-specific data.

Answer: B

Question: 120

A company is using a Retrieval-Augmented Generation (RAG) system to enhance its questionanswering model by incorporating relevant documents from a knowledge base. They need to generate vector embeddings for these documents and the queries using a pretrained transformer-based model. What is the most crucial aspect of generating vector embeddings for effective retrieval in this scenario?

A. Use a different transformer model for document embeddings and query embeddings to ensure diversity in representations.
B. Ensure that both documents and queries are encoded using the same embedding model.
C. Generate vector embeddings with a static embedding model that does not account for fine-tuning on the specific dataset.

Answer: B

Question: 121

You are configuring a chatbot using IBM Watsonx, and you want the chatbot to respond appropriately based on the conversation's context. Which of the following best represents an appropriate stopping criterion for a task where the chatbot generates step-by-step instructions?

A. The model will stop generating once it encounters a user query that requires it to reevaluate the entire prompt context and restart from the beginning.
B. The model will stop generating as soon as the probability of the next token falls below the median probability of previously generated tokens.
C. The model will stop generating once the instruction count reaches five steps, regardless of whether the task has been fully described.
D. The model will stop generating once it identifies a natural completion of the task description, such as reaching a final instruction or a conclusion marker (e.g., "All done").

Answer: D

Question: 122

You are generating product descriptions for an online marketplace using a generative AI model. The output is coherent but tends to repeat the same phrases and words excessively. You decide to apply a repetition penalty to reduce this repetition while keeping the temperature set to a value that maintains creativity in the text generation. Which of the following adjustments would best achieve this goal?

A. Set repetitionpenalty to1.5 and maintain temperature at 0.8
B. Set repetitionpenalty to 0.0
C. Set repetitionpenalty to2.0 andincrease temperature from 0.7 to 1.2
D. Set repetitionpenalty to1.0 anddecrease temperature from 0.8 to 0.3

Answer: A

Question: 123

In the context of zero-shot prompting, you are developing a Watsonx AI model to generate a summary of financial reports. You input the following prompt: "Summarize the financial report in one sentence." Based on your understanding of zero-shot prompting, what outcome should you expect from the model?

A. The model will output a random sentence because zero-shot prompting is unreliable without examples.
B. The model may generate a reasonable summary based solely on the prompt, even if it hasn't been explicitly trained on financial report summaries.
C. he model will refuse to generate a response due to lack of examples provided in the prompt.
D. The model will ask for additional data before generating the summary.

Answer: B

Question: 124

You are tasked with designing a prompt template to assist a chatbot in generating professional email responses for customer service inquiries. The system should prioritize politeness, clarity, and conciseness. What elements should be included in the prompt template to achieve the best results, considering optimal behavior of a large language model (LLM)? (Select two)

A. Provide the customer's emotional context for better alignment with the tone
B. Specify the output tone as polite and professional
C. Include examples of informal customer service responses for variability
D. Ask the model to generate multiple versions of the response and rank them
E. Instruct the model to limit responses to a specific character count

Answer: B,E

Question: 125

What is the key difference between zero-shot and few-shot prompting when used in generative AI models like IBM Watsonx?

A. Zero-shot prompting does not provide any examples in the prompt, while few-shot prompting includes multiple task examples.
B. Zero-shot prompting provides feedback to the model during inference, while few-shot does not allow model feedback.
C. Few-shot prompting requires a model to have pre-trained examples of the task, while zero-shot does not.
D. In zero-shot prompting, the model is fine-tuned before answering, but in few-shot prompting, no fine-tuning occurs.

Answer: A

Question: 126

In the context of quantizing large language models (LLMs), which of the following statements best describes the key trade-offs between model size, performance, and accuracy when using quantization techniques?

A. Quantization always improves model performance but significantly increases model size.
B. Quantization maintains model accuracy but doubles the computation required for inference.
C. Quantization eliminates the need for fine-tuning after deployment, ensuring zero accuracy loss.
D. Quantization reduces model size but may lead to a loss of accuracy, especially with aggressive quantization methods.

Answer: D

Question: 127

You are tasked with fine-tuning a pre-trained generative AI model for customer support automation. The goal is to enhance the model's performance in generating concise, relevant answers to frequently asked questions (FAQs). To do this, you need to optimize the prompt-tuning process. Which two of the following techniques would be most effective for creating a prompt-tuned model for this purpose? (Select two)

A. Introduce randomness in prompts by using variations in the wording for similar FAQs to improve the model's adaptability to different styles.
B. Shorten the prompts to the minimum number of words needed to address the FAQ directly, focusing on the key terms that drive the correct output.
C. Use a large set of domain-specific FAQs and fine-tune the model using those examples, ensuring that prompts are tailored to each type of question.
D. Limit the training data to 100 samples of FAQs to prevent overfitting and keep the prompt-tuning process computationally efficient.
E. Utilize reinforcement learning to penalize long or irrelevant responses during the tuning phase, optimizing the model for concise output.

Answer: B,C

Question: 128

A client is planning to deploy a Watsonx Generative AI model and has raised concerns about ethical usage, bias, and accountability in decision-making. Which of the following is the most critical step to ensure AI governance during the deployment phase of the model?

A. Monitoring and auditing AI decisions for bias and fairness
B. Training the model on additional data to improve accuracy
C. Testing the model's accuracy on a large set of random data
D. Implementing a feedback loop for continuous model improvement

Answer: A

Question: 129

You are tasked with generating creative text outputs using an AI language model for a marketing campaign. You want to ensure that the responses are diverse and unexpected but still somewhat relevant to the prompt. Which combination of temperature and random seed should you use to achieve this?

A. Temperature: 0.7, Random Seed: None
B. Temperature: 1.5, Random Seed: 123
C. Temperature: 1.0, Random Seed: 0
D. Temperature: 0.1, Random Seed: 42

Answer: A

Question: 130

You are integrating watsonx.ai into an external system to handle text generation for a content creation application. The external system requires real-time processing and needs to interact with watsonx.ai frequently. Given this requirement, which integration method is most appropriate for ensuring reliable and scalable communication between the external system and watsonx.ai?

A. Leverage asynchronous REST API calls with callbacks to enable the external system to send requests and continue processing while waiting for the response.
B. Implement Webhooks to receive updates from watsonx.ai when new data is generated and push it to the external system.
C. Integrate watsonx.ai through the SDK to directly embed AI capabilities into the external system, eliminating the need for API calls.
D. Use a REST API with synchronous requests, where the external system waits for watsonx.ai to respond before proceeding.

Answer: A

Question: 131

When conducting prompt engineering to reduce model risks related to hate speech and abusive content, which of the following strategies is least likely to be effective?

A. Applying ethical guidelines during the fine-tuning of the model to prioritize inclusive language
B. Modifying the prompt to include explicit instructions for civil language
C. Increasing the temperature during text generation to encourage more creative outputs
D. Creating a reward model during Reinforcement Learning to penalize hate speech outputs

Answer: C

Question: 132

You are optimizing a Generative AI model for a business application where cost savings are a priority. Which of the following modifications to the model's parameters will most effectively reduce the overall generation cost while minimizing the loss of output quality?

A. Decrease the model's context window size.
B. Set a lower value for the frequency penalty parameter.
C. Set a lower value for the top-p (nucleus sampling) parameter.
D. Reduce the number of layers in the neural network during inference.

Answer: C

Question: 133

You are tasked with creating a prompt template for generating environment descriptions in a generative AI model, which will be used for creating immersive virtual spaces. Which of the following prompt best serves as a flexible template to generate diverse environment descriptions?

A. "Describe a futuristic city with towering skyscrapers and flying cars."
B. "Write about a dense jungle where wild animals roam freely, and the atmosphere is tense, full of suspense."
C. "Describe an environment where {mood} dominates, with {surroundings} contributing to the overall {atmosphere}. Include {time_of_day} and any other important details."
D. "Create a detailed description of a quiet forest during sunrise, focusing on the natural beauty of the trees, birds, and atmosphere."

Answer: C

Question: 134

You have been using a pre-trained foundation model for a financial text summarization application. While the model is generating summaries that are generally accurate, it sometimes fails to handle domain-specific financial jargon. You are considering whether it's time to tune the model to optimize its performance for this task. Which of the following conditions would most strongly justify tuning the foundation model for your specific use case?

A. The model's accuracy fluctuates based on the length of the input text.
B. The model generates output that is highly relevant to general topics but often misinterprets industry-specific terms like "leverage" or "derivative."
C. The model exhibits acceptable performance but occasionally generates off-topic responses unrelated to financial data.
D. The model produces a high number of tokens in each output, which increases the cost of usage.

Answer: B

Question: 135

You are optimizing a large language model (LLM) by prompt-tuning it for specific enterprise-level tasks. The goal is to initialize the prompt in such a way that it helps the model generalize well across various enterprise domains, such as finance, healthcare, and retail. What is the most effective method to initialize the prompt for such a use case?

A. Use a single, highly specific prompt tailored to only one domain, such as finance
B. Initialize the prompt with an ensemble of prompts covering multiple domains
C. Start with a general prompt and gradually specialize it during fine-tuning
D. Use a short prompt that provides no guidance and allow the model to self-optimize

Answer: B

Question: 136

You are tasked with optimizing the cost of text generation using a generative AI model by adjusting model parameters. One of the key parameters you consider is the temperature, which controls the randomness of the output. The client has requested outputs that are more predictable and closer to the intended meaning, without unnecessary creativity, in order to reduce unnecessary token usage and ensure the quality of generated responses. Given the following task:
"Generate a brief summary of a technical article that focuses on key innovations without including unrelated or creative content."
Which of the following temperature settings would be most appropriate to optimize cost while maintaining focus and minimizing unnecessary token usage?

A. 1.0
B. 0.0
C. 1.2
D. 0.1

Answer: D

Question: 137

You are tasked with generating synthetic data for fine-tuning a large language model (LLM) using the IBM Watsonx User Interface. You want to generate relevant training samples to improve the model's accuracy in a text classification task. Which action should you prioritize to generate high-quality synthetic data using the interface?

A. Avoid using domain-specific prompts to keep the synthetic data unbiased and more generalizable.
B. Select only a few generic prompts to generate the largest volume of data possible, ensuring variety.
C. Choose a diverse set of prompts that span different domains unrelated to the original dataset.
D. Customize prompt templates to closely mimic the structure and format of the original training data.

Answer: D

Question: 138

You have fine-tuned a model and notice several issues in the output, including repeated phrases,

incomplete sentences, and factual inaccuracies. Which of the following methods would best help you detect and resolve these data quality problems?

A. Increase the number of training epochs to further refine the model
B. Use a higher learning rate to improve model convergence
C. Apply model quantization to reduce the complexity of the output
D. Analyze token distribution patterns in the generated outputs

Answer: D

Question: 139

A client is deploying a watsonx Generative AI solution to analyze customer feedback in real time. They require a cost-effective solution that can handle occasional traffic spikes but do not expect constant heavy loads. What would be the most appropriate approach to minimize costs while ensuring adequate performance during traffic spikes?

A. Deploy a single powerful GPU instance, relying on its ability to handle both regular and peak traffic without scaling.
B. Deploy a cloud-based instance with manual scaling to adjust resources when traffic spikes occur. C. Deploy the model using a serverless architecture with auto-scaling to handle sudden spikes in traffic.
D. Use a fixed number of on-premise servers that can handle peak traffic, but run them at a lower capacity during low-demand periods.

Answer: C

Question: 140

In IBM Watsonx Generative AI, top-p sampling (nucleus sampling) is a parameter that influences the decoding process. Which of the following best describes how the top-p parameter works during response generation?

A. Top-p sampling selects the next token only from a fixed number of top tokens, ensuring the most predictable outcome.
B. Top-p sampling ensures that only the top 1% most probable tokens are chosen, regardless of context.
C. Top-p sampling adjusts the probability threshold dynamically, allowing the model to sample tokens until the cumulative probability reaches p.
D. Top-p sampling prioritizes tokens based on their frequency in the training dataset, ignoring their conditional probabilities.

Answer: C

Question: 141

You are tasked with deploying a custom prompt template in an enterprise environment. What is the most critical first step in defining the deployment lifecycle to meet client needs?

A. Identify the operational requirements and business constraints
B. Define the monitoring and feedback mechanisms for the prompt's performance
C. Establish a model selection strategy for each prompt template
D. Deploy the prompt template directly to production to get rapid feedback

Answer: A

Question: 142

You are tasked with deploying a custom Watsonx Generative AI model for a client who requires low-latency responses and scalability to handle unpredictable traffic. Which deployment architecture would best meet the client's requirements?

A. Deploying the model in a local on-premise data center
B. Using a serverless architecture with auto-scaling
C. Using a containerized architecture without orchestration
D. Deploying the model on a single dedicated server

Answer: B

Question: 143

You are working on a project that requires generating a large volume of product descriptions for an e-commerce website. The descriptions must be unique, creative, and optimized for SEO. The client has specified that the descriptions must also include certain technical product specifications, but they should not be overly mechanical or robotic in tone. Your team has access to an LLM pre-trained on large-scale, general-purpose corpora. Based on this scenario, what would be your first step to design the most effective Generative AI solution for this task?

A. Optimize the generation process using greedy decoding to ensure concise and accurate descriptions.
B. Use prompt engineering to add technical specifications dynamically to generated text without fine-tuning the model.
C. Fine-tune the pre-trained LLM on a domain-specific dataset of e-commerce product descriptions with an emphasis on SEO-friendly language.
D. Use the pre-trained LLM directly without any modification and feed it the product specifications, prompting it to generate descriptions.

Answer: C

Question: 144

When optimizing the tuning process in IBM Watsonx Tuning Studio for a Generative AI model, which approach would best reduce training time and computational cost while maintaining model performance?

A. Perform full-scale retraining of the model for each new task to ensure maximum adaptability and accuracy.
B. Use all available training data, including unrelated examples, to ensure the model has a broad

understanding of multiple tasks before tuning.
C. Increase the batch size and reduce the learning rate simultaneously to speed up the tuning process and minimize training iterations.
D. Focus the tuning on adjusting only the model's last few layers, which are responsible for taskspecific outputs, while leaving the majority of the model unchanged.

Answer: D

Question: 145

You are designing an AI application that must handle multiple language tasks, such as translation, summarization, and text classification. During testing, you find that for certain specialized tasks, the model performs poorly without examples. Which of the following statements best explains the differences in generalization between zero-shot and few-shot prompting, and how you might improve the model's performance? (Select two)

A. Few-shot prompting typically degrades generalization as it encourages the model to overfit to the specific examples provided, whereas zero-shot prompting forces the model to maintain its general-purpose capabilities.
B. Few-shot prompting enhances the model's generalization by providing the model with a variety of task-specific examples, allowing it to infer the pattern for unfamiliar tasks.
C. Zero-shot prompting is ideal for tasks the model has been explicitly trained for, while few-shot prompting is best for tasks that the model has never encountered before.
D. Few-shot prompting is more effective than zero-shot prompting when the task requires more nuanced, context-dependent outputs, as it allows the model to learn from examples in real-time.
E. Zero-shot prompting leads to better generalization because the model doesn't rely on examples, forcing it to generate answers purely based on the pre-trained knowledge.

Answer: B,D

Question: 146

You are designing a prompt template for generating personalized marketing emails using IBM Watsonx. The emails need to be engaging, personalized based on customer data, and must include a clear call to action. Which of the following is the best structure for a prompt template that can be reused to generate such emails?

A. "Write an email promoting the following product, ensuring to highlight customer-specific preferences and recommend personalized products or offers. Include a clear call to action."
B. "Create a marketing email that lists the features of our latest products. No need to include any personalized information."
C. "Generate a formal email to promote the latest offers, focusing on the technical details of the products and avoiding any emotional appeal."
D. "Generate a generic marketing email promoting our products. Make it professional but avoid using customer-specific details."

Answer: A

Question: 147

You are reviewing the results of a prompt-tuning experiment where the goal was to improve an LLM's ability to summarize technical documentation. Upon inspecting the experiment results, you notice that the model has a high recall but relatively low precision. What does this likely indicate about the model's performance, and how should you approach further tuning?

A. The model is overly conservative, missing relevant details; focus on improving recall.
B. The model's summaries are incomplete, indicating poor understanding of the source material; consider fine-tuning the pre-trained embeddings.
C. The model's length of generated summaries is too short, indicating underfitting.
D. The model is generating too many irrelevant details; focus on improving precision.

Answer: D

Question: 148

You are tasked with designing a prompt for an IBM Watsonx model that will automate customer support responses for a company that sells technical products. The use case requires the model to respond accurately to specific customer inquiries about product troubleshooting. What is the most effective prompt to use for this scenario?

A. "Write a creative explanation of how to fix our product when it fails to function properly."
B. "Based on the following error description, provide a step-by-step solution: 'The device won't power on even after charging for 3 hours.' Be specific and concise in your response."
C. "Write a generic response to help customers with any issue they may have."
D. "Help a customer resolve an issue with our product."

Answer: B

Question: 149

You are tasked with preparing a dataset for training a machine learning model using IBM Watsonx. The dataset contains over 1 million rows, and you notice a significant imbalance in the distribution of class labels. To optimize model performance and minimize bias, what would be the best next step in addressing this imbalance?

A. Apply SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic data for the minority class.
B. Randomly shuffle the data to improve the model's exposure to different instances during training. C. Remove the majority class instances to balance the dataset.
D. Increase the learning rate of the model to improve its ability to learn from the imbalanced data.

Answer: A

Question: 150

You are implementing techniques to ensure that an IBM Watsonx Generative AI model does not expose

any personal or sensitive information (PII) in its outputs. What is the most effective technique for excluding personal information during the inference stage of the generative AI process?

A. Use temperature tuning to control the diversity of outputs, reducing the likelihood of personal information being revealed.
B. Train the model on synthetic data that does not include any personal information.
C. Implement real-time filtering of model outputs using regular expressions to detect and mask personal information.
D. Use a greedy decoding strategy to limit the creativity of the model and prevent unexpected outputs.

Answer: C

Question: 151

When optimizing the tuning process in IBM watsonx Tuning Studio, what is the main purpose of fine-tuning metering options?

A. To monitor the effectiveness of early stopping techniques applied during training.
B. To limit the number of iterations allowed for model training.
C. To track the computational cost and time for fine-tuning models based on resource consumption.
D. To enforce stricter thresholds for model accuracy in order to ensure high precision.

Answer: C

Question: 152

In a RAG system, you need to select an appropriate retriever to fetch relevant documents from a large corpus before generating an answer. You are considering different types of retrievers, including embedding-based and keyword-based retrievers. Which of the following describes a scenario where an embedding-based retriever using a vector database is the best choice?

A. When documents are labeled with metadata, and only metadata needs to be searched
B. When exact keyword matching is required, and synonyms or contextual understanding are irrelevant
C. When most of the queries consist of structured queries with precise Boolean operators and relational database-style searches
D. When retrieval must rely on semantic similarity between a query and documents, even if the exact words in the query don't appear in the document

Answer: D

Question: 153

You are working with IBM Watsonx and have developed a custom machine learning model using Watson Studio. Now, you need to deploy the model so that it can be accessed via an application. Which step is NOT required when deploying the custom model for application access?

A. Assigning appropriate IAM roles for users who need access to the model

B. Verifying model accuracy before deployment.
C. Registering the model in Watson Machine Learning (WML)
D. Defining a scoring endpoint for the model in Watson Machine Learning

Answer: B

Question: 154

IBM Watsonx's Prompt Lab offers various options to refine prompts for generating more effective AI outputs. Which of the following is an accurate description of an editing option available in Prompt Lab?

A. Users can disable the model's access to certain pre-trained knowledge domains within Prompt Lab to focus its output on specific areas.
B. Prompt Lab allows users to experiment with prompt structures, such as adjusting token limits or adding contextual instructions, to improve responses.
C. Users can use Prompt Lab to train the AI model on new datasets and retrain it based on prompt performance.
D. Users can apply real-time machine learning to modify the underlying model parameters within Prompt Lab.

Answer: B

Question: 155

In a generative AI-based customer service chatbot, you notice that the model sometimes generates user responses that inadvertently reveal sensitive personal information, such as names, addresses, or social security numbers. What is the most effective prompt engineering technique to reduce this risk while preserving the chatbot’s functionality?

A. Introduce explicit instructions in the prompt to avoid generating personal information
B. Increase the model's randomness by adjusting the temperature parameter
C. Reduce the model’s token limit to restrict the amount of text generated
D. Use generic prompts that include placeholders for sensitive information (e.g., [USER_NAME])

Answer: A

Question: 156

A client has deployed a generative AI model to generate technical support responses for various issues in their product line. To improve the accuracy of the responses, they have been using tuned prompts for each type of technical issue (e.g., network issues, hardware failures). However, the client is concerned about occasional misclassifications when the prompts are reused across similar but distinct issues. They want to optimize the prompts to handle edge cases better without sacrificing the response speed. What approach would be most appropriate for managing and optimizing these tuned prompts?

A. Leverage prompt chaining, where an initial general prompt narrows down the issue type, followed by a more specific tuned prompt for the final response.
B. Use a single general prompt, apply greedy decoding, and manually edit outputs for any

misclassifications.
C. Create highly specific prompts for each possible issue, fine-tune the model on each prompt, and prioritize correctness over speed.
D. Use a single tuned prompt for each product category, apply top-p sampling, and rely on post-processing to correct any misclassifications.

Answer: A

Question: 157

In the context of generative AI, you are tasked with optimizing a model's performance for a variety of use cases by tuning the prompts. One of your colleagues mentions using a "soft prompt" to improve the model's adaptability. What best describes the difference between a hard prompt and a soft prompt?

A. A hard prompt explicitly specifies all constraints, while a soft prompt relies on implicit learning from continuous inputs during training.
B. Soft prompts are more readable and natural, whereas hard prompts consist of short, technical instructions.
C. A soft prompt is a fixed string of text used in fine-tuning, while a hard prompt adjusts dynamically based on input data.
D. Hard prompts are less efficient because they need to be re-trained with each task, while soft prompts are more versatile and adaptive across multiple tasks.

Answer: A

Question: 158

You are fine-tuning a general-purpose language model on a medical dataset to generate summaries of patient consultations. After fine-tuning, you notice that the model sometimes generates hallucinations—statements that are factually incorrect or irrelevant to the specific domain. You suspect that the fine-tuning process did not sufficiently align the model with the medical domain. Which of the following is the most effective technique to reduce hallucinations during fine-tuning?

A. Use domain-specific tokenization during fine-tuning
B. Add more general-purpose data to the fine-tuning dataset
C. Increase the model's batch size during training
D. Increase the number of layers in the model

Answer: A

Question: 159

You are working on deploying a generative AI model into production. The goal is to ensure that different versions of prompts can be tracked and rolled back in case of degradation in the model's performance. Which of the following strategies would best address versioning for deployment?

A. Integrate prompt versioning into the model deployment pipeline using automated versioning tools like Git.

B. Use endpoint monitoring tools only, without any versioning approach, to track and assess prompt changes in production.
C. Use manual version control through logging and local storage.
D. Maintain different versions of the model and prompts by duplicating them across multiple endpoints without a centralized repository.

Answer: A

Question: 160

While working on generating a concise response to a user prompt, you notice that the generative AI model in IBM watsonx is producing excessively long outputs. You want to ensure that the response is informative but doesn't exceed a specific length. Which of the following parameters should you adjust, and what is the most appropriate value to achieve a concise output without cutting off essential information?

A. Set max_tokens to 100
B. Set max_tokens to 10
C. Set max_tokens to 5
D. Set max_tokens to 1000

Answer: A

Question: 161

You are tasked with designing a Retrieval-Augmented Generation (RAG) system using embeddings to improve the response quality of a generative AI model. In this context, what are embeddings used for, and how do they contribute to enhancing the generative AI's performance?

A. Embeddings serve as a form of knowledge storage within the generative model, allowing it to answer questions without retrieving external information.
B. Embeddings provide a summary of the input data, which the model then uses to generate its final output without retrieving external content.
C. Embeddings compress the input data to reduce computational load, improving the efficiency of the retrieval and generation process.
D. Embeddings transform the input data into high-dimensional vectors, capturing semantic similarities between the input query and potential retrieval candidates to provide contextually relevant information for the generative model.

Answer: D

Question: 162

You are developing a Retrieval-Augmented Generation (RAG) system for a question-answering application. The system relies on generating vector embeddings to retrieve relevant documents based on the input query. What is the key advantage of using vector embeddings for document retrieval in a RAG pipeline compared to traditional keyword-based search methods?

A. Vector embeddings do not provide any meaningful improvement over keyword-based methods unless combined with reinforcement learning algorithms.
B. Vector embeddings capture the semantic meaning of text, allowing for more accurate retrieval of contextually similar documents, even if they do not share exact keywords with the query.
C. Vector embeddings represent text as fixed-length vectors, allowing for faster indexing but no improvements in retrieval accuracy.
D. Vector embeddings increase the memory requirements of the system, making retrieval slower but improving the generation quality of the model.

Answer: B

Question: 163

A team is fine-tuning a large language model (LLM) for a healthcare application. They have decided to implement a taxonomy tree-based curation to prepare their dataset of medical records and patient interactions. What is the primary benefit of using a taxonomy tree in the curation process for such a model?

A. It reduces the overall size of the dataset by filtering out irrelevant data.
B. It provides a hierarchical structure that helps the model understand the relationships between different medical concepts.
C. It ensures that the model only focuses on specific medical specialties by eliminating other categories.
D. It eliminates the need for data cleaning and preprocessing since the taxonomy provides the structure.

Answer: B

Question: 164

You are tasked with generating reproducible and consistent results for a particular GenAI model prompt during development and testing. Which of the following is the primary model parameter to adjust in order to ensure that identical inputs produce identical outputs every time the model is run?

A. Random Seed
B. Max Tokens
C. Top-p (Nucleus Sampling)
D. Temperature

Answer: A

Question: 165

You are tuning a generative AI model to reduce repetitive outputs, which often occur when generating long texts. Which of the following parameter adjustments would most likely reduce the model's tendency to repeat words or phrases without compromising the quality of the generated text?

A. Increasing the Top-k value to 200

B. Lowering the temperature to 0.1
C. Increasing the repetition penalty to 1.2
D. Reducing the beam size to 1

Answer: C

Question: 166

IBM Watsonx Tuning Studio provides multiple benefits for optimizing pre-trained models for specific tasks. Which of the following correctly describes one of the primary benefits of using Tuning Studio for model optimization?

A. It provides automated data augmentation to increase the size of the training dataset without additional labeling, improving the model's generalization.
B. It enables low-cost model fine-tuning by allowing users to update only a subset of the model's layers, reducing the need to retrain the entire model.
C. It performs a full retraining of the model every time new data is added, ensuring the highest level of accuracy for the specific task.
D. It allows for the dynamic adjustment of the model's architecture during inference to optimize performance in real time.

Answer: B

Question: 167

You are tasked with building a Retrieval-Augmented Generation (RAG) system using Elasticsearch for document storage, Watson ML for model hosting, and LangChain for orchestration. The chatbot is supposed to query a large database of medical records and generate responses based on the retrieved information. During testing, you notice that irrelevant documents are often retrieved, leading to low-quality responses. What would be the best approach to improve document relevance in this RAG setup?

A. Update the Elasticsearch index by re-indexing documents using embeddings from the Watson ML model.
B. Fine-tune the Watson ML model on a dataset containing both queries and relevant documents, then update Elasticsearch with the model outputs.
C. Implement a BM25 similarity algorithm in Elasticsearch to improve document retrieval.
D. Pre-process the medical records in LangChain to remove irrelevant information before sending the query to Elasticsearch.

Answer: A

Question: 168

You are tasked with developing a generative AI application for customer service, using a Watsonx model to generate responses based on user queries. The dataset includes user-provided data such as past interaction histories, query types, and feedback ratings. Which two approaches should you implement to optimize the usage of these data elements for generating highly relevant responses? (Select two)

A. Filter out user interaction history data that is more than six months old to focus on recent patterns. B. Normalize feedback ratings data by converting them into discrete categories (e.g., low, medium, high).
C. Prioritize the use of user feedback ratings over query type for generating more personalized responses.
D. Apply data augmentation techniques to artificially expand the dataset with new, synthetic interaction histories.
E. Use query type as an additional input feature to ensure the model tailors responses accordingly.

Answer: B,E

Question: 169

You are part of a team building an AI-powered assistant that helps software developers by answering technical queries. To handle the vast amount of technical documentation efficiently, the team has chosen to implement the Retrieval-Augmented Generation (RAG) pattern. LangChain is used to construct the retrieval-generation pipeline, and SingleStore is employed to store the document embeddings. You need to ensure that the integration of LangChain with SingleStore allows for real-time retrieval of relevant technical documentation. Which of the following configurations should be used to ensure an optimal RAG implementation with LangChain and SingleStore?

A. Use LangChain to embed both queries and documents into the same vector space and store these embeddings in SingleStore. During inference, retrieve documents using SQL queries based on keyword similarity.
B. Use LangChain to embed documents into a vector space and store these embeddings in SingleStore. During inference, embed the query in the same vector space and retrieve documents using vector similarity search.
C. Use LangChain to retrieve previously generated embeddings from SingleStore, then match these embeddings against the query using cosine similarity before passing them to the generative model. D. Use LangChain to retrieve pre-generated embeddings from SingleStore, and then fine-tune the generative model based on the retrieved embeddings to improve response accuracy.

Answer: B

Question: 170

You're developing a generative AI system for a medical diagnosis application that uses patient data. Your responsibility includes designing prompts that extract valuable insights without exposing sensitive patient information. Which of the following steps is the most effective way to reduce model risks related to privacy while ensuring useful outputs from the AI?

A. Restrict the model's output length to reduce the risk of sensitive information leakage.
B. Increase the length of the prompts to provide more context, ensuring more accurate results.
C. Utilize a smaller model to minimize the likelihood of overfitting sensitive data.
D. Employ differential privacy techniques to add noise to the model's outputs.

Answer: D

Question: 171

In a scenario where a large language model (LLM) is integrated into a customer support application, the model is designed to retrieve relevant product information to answer complex user queries. The dataset consists of diverse product documents, including PDFs, user manuals, and website pages. Which of the following best describes when to use a vector database as part of the Retrieval-Augmented Generation (RAG) approach?

A. When there is a need to perform efficient keyword-based search on highly structured documents.
B. When there is a requirement to process large volumes of streaming data in real-time, and exact matching is the priority.
C. When the data consists of diverse unstructured documents, and you need to retrieve semantically similar content using dense vector representations.
D. When the dataset consists mainly of structured tabular data and relational queries.

Answer: C

Question: 172

You are tasked with designing a prompt using a few-shot strategy for Watsonx AI to generate a legal contract summary. You provide two examples of contract summaries before asking the model to summarize a new contract. Which of the following options best demonstrates an effective few-shot prompt for this task?

A. "Provide a summary of the following contract: [New Contract].
Example 1: [First Contract]
Summary: [First Contract Summary]
Example 2: [Second Contract]
Summary: [Second Contract Summary]."
B. "Generate a summary for [New Contract]."
C. "Example 1: [First Contract]
Summary: [First Contract Summary]
Example 2: [Second Contract]
Summary: [Second Contract Summary].
Now generate a summary of the first contract."
D. "Example 1: [First Contract]
Summary: [First Contract Summary]
Example 2: [Second Contract]
Summary: [Second Contract Summary]
Generate a summary of the following contract: [New Contract]."

Answer: D

Question: 173

You are optimizing a generative AI model using prompts. You are tasked with choosing between a hard prompt and a soft prompt for generating a technical report. Which option best describes a soft prompt in

this context?

A. A manually written prompt that is optimized using a grid search of the best keywords and patterns for the task.
B. A prompt that influences the model by embedding learned vectors into the input, modifying the generation behavior indirectly.
C. A prompt that directly instructs the model with domain-specific keywords and structured language to steer the model's output.
D. A prompt that restricts the model's output using hard-coded rules to ensure specific behavior during generation.

Answer: B

Question: 174

You are tasked with explaining the outcomes produced by a Watsonx Generative AI model based on specific prompts. Which of the following approaches is most effective in ensuring transparency and understanding of how the model arrives at its decisions?

A. Explaining the optimization process that minimized the model's loss function
B. Providing an interpretable

Answer: B

Question: 175

You are fine-tuning a machine learning model using IBM Watsonx with a dataset that includes sensitive information. You decide to enable differential privacy while generating synthetic data to ensure the privacy of individual records. What key feature of differential privacy ensures that the synthetic data does not leak private information from the original dataset?

A. Using clustering techniques to group similar data points, preventing individual-level data from being exposed.
B. Limiting the number of data points generated to avoid overfitting the synthetic data.
C. Masking sensitive data fields before creating the synthetic data, ensuring no private information is directly used.
D. Adding controlled noise to the data, ensuring that no individual's data point is easily distinguishable from aggregate data.

Answer: D

Question: 176

During prompt engineering for IBM Watsonx, you need to understand how the decoding process works when generating responses. Which of the following best describes a high-level overview of the decoding process in generative AI?

A. Decoding occurs only in reinforcement learning, where the model refines its responses based on

user feedback over multiple generations.
B. Decoding is the process where the model generates a response token-by-token, choosing each token based on the probability distribution over all possible tokens.
C. Decoding involves translating the input data into a format that the AI model can understand before generating an output.
D. Decoding is the final step in training, where the AI model verifies the accuracy of its outputs against a predefined set of labels.

Answer: B

Question: 177

In the context of Retrieval-Augmented Generation (RAG) models, embeddings play a crucial role in retrieving relevant documents. Which of the following best describes the purpose of embeddings in GenAI, particularly within a RAG system?

A. Embeddings are used to compress large documents into a smaller format, enabling faster storage and retrieval within a RAG system.
B. Embeddings are primarily used to measure the grammatical correctness of the generated responses by checking the sentence structure in the generated text.
C. Embeddings represent each word in a fixed-length vector based on its semantic meaning, which allows for precise retrieval of documents by directly matching words between the query and the documents.
D. Embeddings are dense vector representations of words, phrases, or entire documents that capture semantic relationships, making it easier to find similar content in the knowledge base during retrieval.

Answer: D

Question: 178

What is a key benefit of using Prompt Lab in IBM Watsonx to build reusable prompts for generative AI applications?

A. Prompt Lab uses machine learning to automatically improve prompts over time without the need for developer intervention.
B. Prompt Lab uses predefined responses to train the model, eliminating the need for further prompt experimentation.
C. Prompt Lab automatically adjusts prompts based on the user's previous interactions, ensuring personalized output.
D. Prompt Lab allows for the creation of templates with dynamic placeholders, making it easier to reuse prompts in various contexts without modification.

Answer: D

Question: 179

During the fine-tuning of a large language model (LLM) with InstructLab for a legal document classification task, you notice that the model performs exceptionally well on the training set but poorly on

the validation set. What could be done to address the overfitting issue and improve the model's generalization? (Select two)

A. Increase the size of the model to better capture complex patterns in the training data.
B. Remove all regularization and fine-tune the model on the training set until convergence.
C. Introduce dropout regularization during fine-tuning to prevent overfitting.
D. Use early stopping based on the validation set performance.
E. Fine-tune the model for more epochs to ensure the model has fully learned from the training data.

Answer: C,D

Question: 180

You are deploying a generative AI model to summarize legal documents. During testing, you observe that the model sometimes produces inaccurate or fabricated information about legal clauses that do not exist in the original document. Which of the following strategies would be the most effective in reducing hallucinations in the model's output?

A. Reduce the maximum token limit and set a higher repetition penalty.
B. Use beam search with a high beam width to generate the most likely sequence.
C. Implement a post-processing step that filters for factual accuracy using a knowledge base.
D. Increase the model’s temperature to allow for more diverse outputs.

Answer: C

Question: 181

You are designing a Retrieval-Augmented Generation (RAG) model within IBM watsonx to assist in generating responses for a customer service chatbot. The model needs to leverage a knowledge base (KB) of articles to enhance the accuracy of responses. Which of the following correctly describes how the RAG model can be implemented to achieve this goal?

A. The RAG model first generates a response from the language model and then retrieves related articles from the knowledge base to augment the output.
B. The RAG model retrieves entire documents from the knowledge base and directly outputs them as the response to the user’s query.
C. The RAG model retrieves relevant knowledge from the knowledge base using a search or retrieval mechanism before generating a response, which is then influenced by this retrieved information.
D. The RAG model first generates an intermediate query based on the input question, which is then passed through a retrieval system to generate the final response.

Answer: C

Question: 182

A team is tasked with selecting a vector database to store and search through millions of document embeddings efficiently. They need to ensure that the chosen database supports operations such as approximate nearest neighbor (ANN) search and scalability. Which of the following considerations is

most important when selecting the appropriate vector database for this application?

A. The database should limit the embedding size to 256 dimensions to reduce storage costs and ensure scalability.
B. The vector database should allow for both ANN search and customizable distance metrics, such as cosine or Euclidean distance.
C. The database should store embeddings in relational format (rows and columns) for compatibility with traditional SQL queries.
D. The database must support exact nearest neighbor (ENN) searches to ensure high-precision results.

Answer: B

Question: 183

A generative AI model is given the following prompt: "Translate the following sentence into French: 'The sun is shining brightly today.'" No additional context or examples are provided. This is an example of which type of prompting and why is it likely to succeed?

A. Zero-shot prompting, but it will only succeed if the prompt includes a few additional translation examples.
B. Zero-shot prompting, because the model is expected to perform the task without any example.
C. Zero-shot prompting, but it will likely fail since translation tasks always need examples for accuracy.
D. Few-shot prompting, because the task requires translation examples to guide the model.

Answer: B

Question: 184

In the IBM watsonx Prompt Lab, you are tasked with improving a customer-facing chat model by editing and refining prompts. Which of the following is an effective prompt editing option to fine-tune a chatbased generative AI model?

A. Altering the tone of the prompt to make it more aligned with the brand's voice and customer expectations.
B. Modifying the stop sequence to ensure the model responds with shorter answers for concise customer interactions.
C. Increasing the temperature parameter to make the model more deterministic and consistent in its responses.
D. Decreasing the number of tokens allowed in the output to encourage more creative responses.

Answer: A

Question: 185

You are designing a document search application using IBM Watson's RAG architecture. You need to generate vector embeddings for your document corpus, which consists of unstructured text data. The vector embeddings will be used for similarity search in conjunction with a retriever model. Which of the

following steps is a prerequisite to generating effective vector embeddings from the unstructured text using an embedding API?

A. Ensuring that the text data is fully normalized, including removing punctuation and converting all text to lowercase
B. Choosing a transformer-based model pre-trained on a domain-specific corpus to generate the vector embeddings
C. Preprocessing the text by tokenizing it into individual words and removing stop words
D. Compressing the text data to reduce its dimensionality before passing it to the embedding API

Answer: B

Question: 186

In the context of using Tuning Studio in IBM watsonx, which of the following is the primary benefit of fine-tuning a pre-trained Generative AI model using this tool?

A. Reduces the time needed to generate text responses by simplifying the model's architecture.
B. Improves model performance on domain-specific tasks without retraining from scratch.
C. Allows you to train the model from scratch on a new dataset.
D. Increases the size of the model to handle more complex prompts.

Answer: B

Question: 187

Which of the following factors is most likely to contribute to biased outputs in a generative AI system?

A. Prompt length that exceeds the model's context window.
B. Insufficient diversity in the training dataset used to train the model.
C. The use of too many input parameters in the prompt.
D. Repeatedly using the same prompt for multiple generations.

Answer: B

Question: 188

You are using the IBM watsonx platform to fine-tune an existing generative AI model for a text-based customer support system. Due to privacy constraints, the company cannot use real customer data, so synthetic data must be generated using the platform's UI. Your task is to create and configure the synthetic data and fine-tune the model to optimize customer query handling. Which of the following actions will best ensure the synthetic data generation process leads to a well-tuned model that handles diverse customer queries effectively? (Select two)

A. Prioritize the generation of synthetically perfect customer queries with no grammatical errors to ensure high-quality data.
B. Generate a synthetic dataset that only covers the most frequent customer queries to reduce complexity in model training.

C. Leverage the user interface to create synthetic data that includes rare edge cases, such as technical support questions or multi-part inquiries.
D. Simulate both simple and complex customer queries, including ambiguous or vague requests, in the synthetic data.
E. Fine-tune the model immediately after generating synthetic data, without further inspection, to maintain efficiency.

Answer: C,D

Question: 189

In the context of large-scale synthetic data generation for fine-tuning a generative AI model, which of the following practices can lead to data that effectively improves the model's performance on downstream tasks?

A. Randomly generating sentences without adhering to the task-specific instructions
B. Using domain-specific templates to generate synthetic data that reflects the target use case
C. Generating synthetic data that overrepresents the simplest task cases to reduce computational load
D. Avoiding any post-processing or filtering of synthetic data to retain diversity

Answer: B

Question: 190

In a Retrieval-Augmented Generation (RAG) system, embeddings play a central role in linking input queries with relevant external knowledge. Different embedding models can be used to generate these embeddings. Which of the following embedding models is best suited for capturing semantic meaning in text for use in a RAG system?

A. Bag-of-Words (BoW)
B. One-Hot Encoding
C. Word2Vec
D. Latent Dirichlet Allocation (LDA)

Answer: C

Question: 191

You are using IBM's Tuning Studio to fine-tune a large-scale foundation model for a customer service chatbot. The goal is to optimize the model for performance in handling a wide variety of customer queries while minimizing computational costs. Before making any changes, you want to understand how Tuning Studio can help achieve your optimization goals. Which of the following is the most significant benefit provided by Tuning Studio when optimizing a generative AI model?

A. Tuning Studio reduces the dataset size needed for training by implementing automated data augmentation strategies.
B. Tuning Studio allows the user to implement custom model architectures from scratch to meet specific task requirements.

C. Tuning Studio automatically deploys the fine-tuned model to production environments without requiring further testing.
D. Tuning Studio provides real-time monitoring of model performance metrics during the fine-tuning process, allowing you to adjust hyperparameters effectively.

Answer: D

Question: 192

When optimizing a generative AI model using the Tuning Studio in IBM Watsonx, which two of the following actions can most effectively improve model performance when dealing with underfitting issues? (Select two)

A. Reduce the learning rate
B. Enable early stopping
C. Increase the number of training epochs
D. Increase the model's complexity by adding more layers
E. Decrease the batch size

Answer: C,D

Question: 193

While working with IBM Watsonx, you are asked to ensure that the synthetic data generated has a low privacy leakage probability. During the generation process, you must adjust settings related to privacy leakage probability. Which factor most directly affects the probability of privacy leakage when generating synthetic data with differential privacy?

A. The amount of noise added to the synthetic data during the differential privacy process.
B. The use of ensemble models to train on different portions of the dataset and reduce dependency on individual records.
C. The quality and variance of the synthetic data generated.
D. The size of the original dataset used for training.

Answer: A

Question: 194

IBM Watsonx Tuning Studio offers several benefits when fine-tuning pre-trained models for specific tasks. Which of the following is not a key benefit of using Tuning Studio?

A. Tuning Studio allows selective parameter updates, reducing the need to retrain the entire model for each new task.
B. Tuning Studio enables real-time, dynamic adjustments to the model's architecture during inference to handle new tasks.
C. Tuning Studio integrates with existing AI infrastructure to streamline model fine-tuning without requiring complex deployment processes.
D. Tuning Studio offers fine-tuning with minimal data, allowing users to adapt models to niche tasks

without needing large datasets.

Answer: B

Question: 195

In the context of a Retrieval-Augmented Generation (RAG) system using IBM Watsonx, which of the following is the correct process for generating vector embeddings for document retrieval?

A. Directly extract the vector embeddings from raw text without any tokenization or processing steps by passing the raw data to a pre-trained language model.
B. Tokenize the text data, then pass the tokens through a pre-trained language model to generate vector embeddings for each token, combining them into a final vector.
C. Use a pre-trained generative model to generate text predictions, then use these predicted tokens as the basis for vector embedding generation.
D. Load the text data into the model, tokenize it, and directly feed the tokens into a decoder for vector generation.

Answer: B

Question: 196

In planning the deployment of a generative AI model that relies on a large corpus of data, you need to organize and version the data repository used for training prompts. Which of the following approaches best ensures efficient data versioning, integrity, and easy rollback during prompt refinement?

A. Implement a data versioning system like DVC (Data Version Control) or MLflow, integrated with a cloud storage service, to track data changes and link specific data versions to corresponding model and prompt versions.
B. Store all datasets in a monolithic repository and append new data versions as additional files with timestamps.
C. Store data versions directly in the production environment without versioning to save on storage costs.
D. Use a relational database to store the corpus, but without tracking any changes to the datasets.

Answer: A

Question: 197

A business is implementing a RAG solution to enhance its chatbot capabilities. The chatbot needs to answer queries using a large collection of unstructured documents. Which scenario best highlights when to use a vector database to augment this system?

A. When performing traditional database operations like sorting and filtering based on numeric or categorical values.
B. When working with large amounts of unstructured text data, to enable semantic search through embeddings that represent the meaning of documents.
C. When you need to provide answers based on the keywords present in the documents, as vector

databases are designed for keyword-based retrieval.
D. When storing highly structured relational data, as a vector database excels at managing tabular information efficiently.

Answer: B

Question: 198

When using few-shot prompting, which of the following is a common limitation you may encounter while working with Watsonx AI for text generation tasks?

A. Few-shot prompting cannot handle open-ended tasks, as it only works for structured data inputs.
B. Few-shot prompting requires the model to be retrained with additional examples.
C. Few-shot prompting is only effective for classification tasks and cannot be used for text generation.
D. The model may overfit to the few examples provided, leading to overly specific responses.

Answer: D

Question: 199

A healthcare organization is deploying a generative AI model to assist doctors in generating patient reports based on medical notes and test results. The organization has strict compliance requirements regarding patient data privacy (e.g., HIPAA in the U.S.) and needs to ensure that the model outputs are governed under their AI policies. You are tasked with integrating AI governance policies into the deployment to meet both ethical and legal standards. Which AI governance measure is most important to implement during the deployment phase of this AI solution, given the healthcare organization's need for compliance with patient privacy regulations?

A. Implementing a system to track and log all model-generated inferences, flagging any that suggest potential bias or non-compliance with privacy regulations.
B. Ensuring that the model's training data includes only anonymized patient records to avoid potential data breaches.
C. Deploying an explainability module to allow doctors to see the reasons behind each AI-generated suggestion, ensuring ethical use of the model.
D. Requiring doctors to manually review every AI-generated report before it is finalized to meet compliance with medical standards.

Answer: A

Question: 200

When optimizing a model using soft prompts, which of the following is a potential advantage over hard prompts in the context of a generative AI application focused on creative content generation?

A. Soft prompts offer higher flexibility by dynamically adjusting the model’s behavior based on pre-learned embeddings rather than relying on static instructions.
B. Soft prompts produce more deterministic outputs because they are optimized for a specific task through manual input adjustment.

C. Soft prompts allow for more human oversight by embedding specific instructions into the model's generation process.
D. Soft prompts are preferable in scenarios where the task requires rigid structure and adherence to specific predefined guidelines.

Answer: A

Question: 201

Your team has developed multiple versions of a custom Watsonx Generative AI model to address different use cases. During deployment, the client requires that all model versions be deployed concurrently, allowing requests to be routed to the appropriate model based on the input data. What deployment strategy would best accommodate this requirement?

A. Blue-Green deployment
B. Multi-model serving with dynamic routing
C. A/B testing deployment
D. Canary deployment

Answer: B

Question: 202

You are working with IBM Watsonx to generate automated customer support responses. To ensure consistency and flexibility in responses across multiple product categories, you decide to use prompt variables. Which of the following best describes the benefits of using prompt variables in this scenario?

A. They enable the AI to better predict the intent of the customer query, reducing the need for explicit customer input.
B. They ensure that the AI generates responses with consistent tone and personality across all prompts, regardless of product category.
C. They allow for dynamic input fields in prompts, making it easier to tailor responses for different product categories without rewriting the entire prompt.
D. They automatically improve the accuracy of the AI's responses by allowing the system to learn from each generated prompt.

Answer: C

Question: 203

You are creating a prompt template for a generative AI model that helps technical support staff troubleshoot customer issues based on symptoms provided. The goal is to generate a clear, step-by-step diagnostic process. What elements would improve the effectiveness of the template? (Select two)

A. Set a high temperature value to explore more creative diagnostic approaches
B. Provide explicit instructions to break the response into steps
C. Incorporate complex technical jargon to ensure expert-level responses
D. Ask the model to generate multiple diagnostic paths for each issue

E. Include an instruction to ask clarifying questions when the issue is unclear

Answer: B,E

Question: 204

In the context of prompt engineering, how does the repetition penalty parameter influence the behavior of a generative AI model, and which of the following scenarios demonstrates its correct usage?

A. A repetition penalty higher than 1.0 will discourage the model from repeating the same words or phrases, while a value lower than 1.0 encourages repetition
B. Setting the repetition penalty to 1.0 disables any influence on repetition and allows the model to generate text without penalizing previously used tokens
C. A repetition penalty of 0.5 is ideal for generating creative text, as it promotes repetition of key phrases for emphasis
D. A repetition penalty of 1.0 encourages the model to introduce more varied word choices in its output

Answer: A

Question: 205

In what situation might greedy decoding fail to generate an optimal output, even though it consistently chooses the most probable token at each step?

A. Greedy decoding maximizes local probabilities but can lead to suboptimal global coherence
B. Greedy decoding is highly effective when multiple equally probable tokens are available at each step
C. Greedy decoding guarantees the highest overall probability for the output sequence
D. Greedy decoding works best when combined with temperature scaling to increase randomness

Answer: A

Question: 206

You are preparing a dataset to fine-tune a language model for sentiment analysis. The dataset consists of user reviews with a mix of neutral, positive, and negative sentiments. Which of the following strategies will best ensure that the model learns balanced sentiment detection?

A. Focus only on neutral and negative examples to challenge the model
B. Increase the number of positive sentiment examples in the dataset
C. Ensure an equal distribution of positive, negative, and neutral sentiment examples
D. Convert neutral examples into either positive or negative to simplify the task

Answer: C

Question: 207

You are training a generative AI model using IBM's Tuning Studio and want to optimize its performance. You aim to avoid both overfitting and underfitting by carefully selecting the appropriate number of

epochs. Which of the following strategies would best help you set the optimal number of epochs during the tuning process?

A. Gradually increase the number of epochs until the loss on the training set reaches zero
B. Use a single epoch to avoid overfitting
C. Set the number of epochs equal to the size of the dataset
D. Set a high number of epochs and use early stopping to determine the optimal point

Answer: D

Question: 208

In a scenario where a developer is creating reusable prompt templates for a Watsonx Generative AI project, what is the most effective method to track the usage and performance of these templates over time?

A. Embedding unique identifiers in the prompt templates and using Watsonx's logging mechanisms
B. Relying on manual tracking of templates in a spreadsheet for each generation
C. Using static prompt templates without any tracking, as tracking adds unnecessary overhead
D. Leveraging Watsonx's native analytics and monitoring tools with built-in prompt tracking features

Answer: A

Question: 209

You are tasked with integrating watsonx.ai into a legacy system that operates over HTTP and requires strict security and monitoring of the API calls. The legacy system lacks modern authentication mechanisms like OAuth. Which integration method would best suit the needs of this environment while ensuring security and efficient API management?

A. Directly embed the watsonx.ai SDK into the legacy system to handle API management and security, eliminating the need for external API calls.
B. Implement a GraphQL API to allow the legacy system to query specific fields and reduce payload size, thereby improving efficiency.
C. Use REST API with API key-based authentication, leveraging HTTPS for encrypted communication between the legacy system and watsonx.ai.
D. Use a REST API with OAuth authentication and token management to secure the communication.

Answer: C

Question: 210

You are working with IBM watsonx's generative AI model and wish to reduce the likelihood of generating rare, low-probability tokens while still retaining some level of creativity. You decide to use top-k sampling for this purpose. Which of the following settings for the top-k parameter would be most effective in achieving a balance between creativity and maintaining coherent outputs?

A. Settop-k to 50
B. Settop-k to 1000

C. Settop-k to 5
D. Settop-k to 1

Answer: A

Question: 211

IBM Watsonx Tuning Studio provides metering options to help monitor and optimize fine-tuning processes. Which of the following best describes how these metering options can optimize resource usage during fine-tuning?

A. Fine-tuning metering adjusts the learning rate dynamically to optimize both training speed and accuracy.
B. Fine-tuning metering enables hyperparameter search, automatically testing multiple configurations to find the most optimal one for the current task.
C. Fine-tuning metering options automatically halt the training process if no significant improvements in model accuracy are observed, reducing unnecessary resource usage.
D. Fine-tuning metering provides insights into token usage and computational costs, allowing users to set budget constraints and minimize expenses during the tuning process.

Answer: D

Question: 212

You are tasked with designing a generative AI model to assist users in filling out a form that collects personal information, such as email addresses and phone numbers. What is the most appropriate method to ensure the model can differentiate between required personal information and unnecessary sensitive data that should not be included in the output?

A. Train the model exclusively on datasets that contain no personal information
B. Leverage regular expressions to filter out personal information in the prompt
C. Use a post-generation filter to remove any text that appears to be personal information
D. Embed privacy-sensitive heuristics in the model's prompt to guide its behavior

Answer: D

Question: 213

You are working on a Retrieval-Augmented Generation (RAG) system using IBM watsonx. The system needs to retrieve relevant documents based on a user's query and generate a response using a language model. To optimize retrieval, you are tasked with generating vector embeddings for documents and queries using a pre-trained model. Your goal is to ensure that the embeddings are semantically meaningful to improve the retrieval accuracy. Which of the following steps should be taken to ensure the vector embeddings are correctly generated and effective for document retrieval in a RAG system? (Select two)

A. Manually adjust the embedding vectors to emphasize certain keywords that are more important for retrieval.

B. Use a generative language model to generate embeddings without any fine-tuning, as it captures all the necessary context.
C. Normalize the vector embeddings after generation to ensure they are comparable during retrieval.
D. Use a pre-trained model designed specifically for embedding generation rather than generalpurpose language models.
E. Generate embeddings for documents only and skip embeddings for user queries, relying on traditional keyword-based retrieval for queries.

Answer: C,D

Question: 214

Which quantization technique aims to optimize a model by converting weights and activations into 8-bit integers while minimizing the impact on the model's performance?

A. Hybrid quantization
B. Post-training static quantization
C. Quantization-aware training (QAT)
D. Post-training dynamic quantization

Answer: C

Question: 215

A financial institution is using a generative AI model to create reports based on transaction data. During deployment, the institution notices that the model sometimes fabricates trends or patterns that do not exist in the underlying data. This is an example of a hallucination. Which of the following techniques would best minimize this risk during inference?

A. Increase the top-p value to ensure more tokens are considered during generation.
B. Disable the model's autoregressive capability to prevent it from generating future predictions.
C. Reduce the model size to decrease its capacity to hallucinate complex patterns.
D. Use a retrieval-augmented generation (RAG) model that incorporates external financial data into the generation process.

Answer: D

Question: 216

You are deploying a generative AI model for a financial services company. The model is responsible for automating customer support and providing recommendations. Due to the sensitive nature of financial data, the company emphasizes the need for robust AI governance. What governance mechanism should you prioritize to ensure compliance with data privacy regulations and maintain trust in AI outputs?

A. Implementing role-based access control (RBAC) to restrict who can interact with the model.
B. Ensuring model version control to track changes and updates made to the model during the deployment process.

C. Regularly retraining the model to avoid performance degradation due to data drift.
D. Using AI explainability techniques to make the model's decisions transparent to regulators and customers.

Answer: D

Question: 217

In IBM Watsonx's Prompt Lab, example input prompts can be used to improve the effectiveness of generated responses. Which of the following best describes a key benefit of utilizing example input prompts in Prompt Lab?

A. Example input prompts help generate responses that are more aligned with the specific context or style intended by the user.
B. Example input prompts allow the model to learn new concepts and update its knowledge base dynamically.
C. Example input prompts guarantee consistency across all outputs, regardless of variability in user-provided data.
D. Example input prompts automatically adjust the model's training dataset for more accurate predictions.

Answer: A

Question: 218

You are tasked with fine-tuning a pre-trained large language model (LLM) for a customer service chatbot using IBM's InstructLab. You need to customize the LLM to improve its ability to handle specific user instructions related to order management, such as tracking orders, processing returns, and issuing refunds. Which of the following components in InstructLab is the most critical for guiding the LLM to respond appropriately to these specific instructions?

A. The feedback loop system for real-time user input validation.
B. The data loader module for transforming data into tokenized inputs.
C. The pre-processing pipeline for normalizing and standardizing the dataset.
D. The prompt engineering interface for designing task-specific instructions.

Answer: D

Question: 219

As an IBM Watsonx Generative AI engineer, you are tasked with creating a chatbot for a public-facing service. One key concern is ensuring that the model does not generate or propagate hate speech, abusive content, or profanity. To mitigate these risks, you must implement appropriate controls. Which of the following is the best approach to mitigate hate speech, abuse, and profanity from being generated by your AI model?

A. Apply a simple word-level blacklist filter to detect and remove harmful content from the model

output.
B. Train the model only on data that excludes all user-generated content to prevent exposure to harmful language.
C. Fine-tune the model with a highly curated dataset that contains labeled examples of hate speech, abuse, and profanity for the model to learn to avoid.
D. Use IBM Watsonx's HAP (Hate, Abuse, and Profanity) filter to dynamically detect and block harmful content at inference time.

Answer: D

Question: 220

After completing a prompt-tuning experiment, you notice that the model's accuracy in generating relevant responses is high, but the fluency and grammatical correctness of the outputs seem to be suboptimal. What statistical metric would most directly indicate this issue, and what action should you take to improve the output?

A. F1 score; increase the training dataset size to improve overall accuracy.
B. ROUGE score; adjust the token generation limit to ensure longer outputs.
C. BLEU score; improve prompt engineering to ensure that the model focuses on fluency.
D. Perplexity score; apply additional language model fine-tuning on grammatical correctness.

Answer: D

Question: 221

In the context of decoding methods in IBM Watsonx Generative AI, both top-k and top-p sampling are used to control the output of the model. Which of the following statements correctly distinguishes between top-k and top-p sampling?

A. Both top-k and top-p sampling ensure that no tokens are selected with a probability below the set threshold, regardless of their context.
B. Top-k sampling samples from a dynamic range of tokens, while top-p sampling restricts choices to a fixed number of tokens.
C. Top-p sampling allows for a more flexible token selection process based on cumulative probabilities, while top-k limits token selection to a fixed number of top choices.
D. Top-p sampling ensures deterministic outputs, whereas top-k sampling guarantees random selection from the entire vocabulary.

Answer: C

Question: 222

You are tasked with integrating IBM watsonx with a third-party customer relationship management (CRM) system to enhance customer interactions through conversational AI. Which of the following approaches best ensures real-time responses while minimizing the latency caused by data exchange between systems?

A. Utilize watsonx's real-time streaming API with pre-configured webhooks in the CRM system for instant data retrieval and response.
B. Set up periodic data synchronization between the CRM system and IBM watsonx using FTP file transfers.
C. Employ an asynchronous message queue between IBM watsonx and the CRM system to handle customer requests.
D. Use IBM watsonx API for batch processing of customer requests, triggering the CRM system via a scheduled job.

Answer: A

Question: 223

You are tasked with setting up a pipeline that uses an embedding API for a RAG system. Before the API can be used to generate vector embeddings for large-scale document retrieval, certain prerequisites must be met. Which of the following is a valid prerequisite for using the embedding API efficiently?

A. Precomputing the embeddings manually to avoid real-time API calls
B. Securing the API access with proper authentication and authorization measures
C. Ensuring that all data sources are available in a structured format such as JSON or CSV
D. Training the API to generate embeddings for every possible word in the corpus

Answer: B

Question: 224

You are tasked with building a generative AI model to help create automated marketing copy for a business. A key concern is the potential generation of biased or legally sensitive content, which could negatively impact the company's reputation. Which of the following strategies would be the most effective in mitigating these model risks?

A. Implement a post-processing filter to remove any potentially offensive or legally sensitive content. B. Include fairness metrics in the model evaluation stage to monitor for biased outputs.
C. Use a comprehensive training dataset that includes diverse business domains to reduce biases.
D. Use reinforcement learning to fine-tune the model based on user feedback to eliminate bias in the long term.

Answer: B

Question: 225

When crafting prompts for a generative AI model, readability is crucial to ensure clarity for both the model and human collaborators. You are asked to optimize the prompt to improve both the generation's accuracy and usability. Which strategy would most effectively balance readability with optimal model performance?

A. Craft technical prompts that focus solely on model parameters, ignoring human readability for performance gains.

B. Create longer, detailed prompts that cover all edge cases to reduce the need for multiple training iterations.
C. Use simple, concise instructions that avoid ambiguity but ensure all necessary constraints are included.
D. Focus on minimal prompts to reduce computational load, even if it sacrifices some clarity.

Answer: C

Question: 226

You are using IBM's Tuning Studio to optimize a generative AI model. The model performance on your validation set has plateaued, and you suspect that tuning certain parameters in the Tuning Studio will improve the outcome. Which of the following actions would be most effective in improving the model's performance?

A. Decrease the batch size
B. Use a larger validation set
C. Increase the number of epochs
D. Enable early stopping

Answer: D

Question: 227

You have successfully deployed a custom model using IBM Watson Machine Learning. Now, an application needs to access the model for inference. Which of the following configurations will ensure that the application has appropriate access to the deployed model?

A. Use the default IAM policies for all users in the IBM Cloud project
B. Assign the "viewer" role to the application's service account in Watson Machine Learning
C. Implement OAuth2 authentication using IBM Cloud IAM for securing the API.
D. Create a public-facing API endpoint without authentication

Answer: C

Question: 228

When planning data elements for optimizing a generative AI model's performance in IBM watsonx, which of the following strategies should be prioritized to ensure data quality and model accuracy?

A. Ensure all data used is unstructured to allow for maximum model flexibility.
B. Prioritize balanced datasets to minimize model bias and enhance generalization.
C. Leverage data augmentation techniques only on the smallest available dataset.
D. Focus on using only labeled data, as generative AI models do not perform well on unlabeled data.

Answer: B

Question: 229

You want a generative AI model to summarize a lengthy text in one sentence. You provide the following prompt: "Summarize the following paragraph in one sentence: 'Artificial intelligence (AI) is the simulation of human intelligence in machines that are programmed to think and learn. The field of AI includes everything from speech recognition to problem-solving and robotics.'" No prior examples are given. What type of prompting is being used, and what are the expectations?

A. Zero-shot prompting, but the model may fail without detailed instructions on what aspects of the text to focus on.
B. Few-shot prompting, but the model will likely struggle without a few examples of summaries provided.
C. Zero-shot prompting, but it requires the addition of a few examples for the model to generate a summary.
D. Zero-shot prompting, with the expectation that the model can summarize common concepts like AI due to pre-training.

Answer: D

Question: 230

You are developing a natural language generation (NLG) system for financial report summaries. The system needs to generate concise, high-quality summaries for different financial instruments. Tuning Studio is available as a tool to improve the performance of the model. Which of the following capabilities of Tuning Studio is most helpful for improving the NLG system's performance in this domain-specific application?

A. It allows you to visually inspect and analyze the model's token-by-token generation process.
B. It helps in selecting the most relevant datasets from a pre-existing library for fine-tuning the model.
C. It provides real-time feedback on model performance during text generation.
D. It enables the fine-tuning of pre-trained models on custom datasets to improve accuracy and relevance to the specific domain.

Answer: D

Question: 231

A company is considering using IBM Watsonx for two different use cases: (1) automating email responses for routine inquiries from customers, and (2) generating creative marketing copy for new product campaigns. Which of the following best describes the model selection process for these use cases?

A. Use the smallest possible model to reduce computational costs, regardless of the use case.
B. A single model should be used for both use cases since Watsonx can handle any type of text generation task equally well.
C. Use a large language model fine-tuned for email response generation for both tasks, as the fine-tuning process will enable the model to handle creative tasks as well.

D. Choose a domain-specific model for automating email responses and a more generalized model with creative capabilities for generating marketing copy.

Answer: D

Question: 232

A client needs a Generative AI solution to summarize large legal documents into concise briefs. The solution must capture the critical legal arguments while preserving the formal language required in legal contexts. Additionally, the client wants the model to identify key legal clauses and ensure their inclusion in the summaries. You have a pre-trained LLM that was trained on general text, and now you must design a generative solution to meet the client's needs. What would be your next step in analyzing and designing the most effective solution?

A. Use a zero-shot approach, prompting the model to summarize legal documents without further fine-tuning.
B. Use prompt engineering to instruct the model to focus on key legal clauses and adjust the output to match the legal context.
C. Apply model quantization to optimize the LLM for handling long legal documents more efficiently. D. Fine-tune the pre-trained LLM on a dataset of legal documents, specifically focusing on case law, contracts, and formal briefs.

Answer: D

Question: 233

A company is building a conversational AI system using a Retrieval-Augmented Generation (RAG) architecture. They need to store and retrieve large amounts of unstructured data efficiently, ensuring that their model can retrieve semantically similar documents based on user queries. When is the use of a vector database most appropriate in this scenario?

A. When the data consists of text, images, or other unstructured content, and the goal is to retrieve items based on semantic similarity rather than exact matches.
B. When you need to store a relatively small dataset (under 1,000 records) and can perform bruteforce search without significant performance issues.
C. When you need to perform frequent joins and aggregations across multiple tables of structured data.
D. When the data is highly structured and queries are focused on exact matches like numeric ranges or specific dates.

Answer: A

Question: 234

Which prompt engineering strategy is most effective for reducing the risk of generating biased content in a generative AI model?

A. Adjusting the model's temperature to a lower value to generate more deterministic and consistent

outputs.
B. Using a larger number of examples in the prompt to steer the model towards safer and more neutral outputs.
C. Designing prompts that explicitly ask for balanced perspectives, such as "Provide both pros and cons of this issue."
D. Including specific disclaimers or guidelines in the prompt instructing the model to avoid biased language.

Answer: C

Question: 235

You are tasked with creating a prompt-tuned model that generates optimal, task-specific responses for a financial advisory chatbot. Your goal is to improve the model's accuracy in answering financial queries, and you need to determine the right parameters to focus on during the tuning process. Which two of the following strategies are most effective in optimizing prompt-tuned models for accuracy? (Select two)

A. Include domain-specific financial terms in the prompt-tuning data to help the model specialize in accurate financial advice generation.
B. Increase the number of layers fine-tuned in the model to capture deeper contextual information from financial data.
C. Apply a low temperature setting (e.g., 0.2) during inference to ensure more deterministic and precise responses.
D. Choose an initial learning rate that is high to encourage faster convergence during the fine-tuning process.
E. Use a beam search decoding algorithm with a large beam width to generate a variety of response candidates for each query.

Answer: A,C

Question: 236

You are fine-tuning the output behavior of a generative AI model in IBM Watsonx for creative content generation. You decide to adjust the temperature parameter to influence the randomness of the model's output. Which of the following best describes the effect of increasing the temperature value?

A. Raising the temperature makes the model more likely to repeat tokens, reducing variability in its responses.
B. Increasing the temperature makes the model generate more deterministic responses by always selecting the most probable token at each step.
C. A higher temperature setting reduces the length of the generated output by limiting the number of tokens in each response.
D. Raising the temperature encourages the model to consider less likely tokens, leading to more diverse and creative outputs.

Answer: D

Question: 237

You are optimizing a generative AI prompt for creative content generation, but you want to ensure that outputs with the same prompt can vary slightly across multiple runs to maintain freshness in the responses. Which parameter should you adjust to strike a balance between random variability and consistent quality?

A. Max Tokens = 300, Random Seed set to 42
B. Set a fixed Random Seed to zero
C. Temperature = 0.1, Random Seed unset
D. Top-p = 0.9, Temperature = 1.0, Random Seed unset

Answer: D

Question: 238

While generating synthetic data using IBM Watsonx's User Interface to supplement your fine-tuning dataset for a customer service chatbot, you notice some responses are incoherent or contain contradictions. What steps should you take to maintain the quality and relevance of the synthetic data?

A. Enable auto-validation in the UI to automatically discard any generated data that doesn't align perfectly with training data patterns.
B. Only use synthetic data that matches predefined patterns exactly to avoid any incoherence or errors.
C. Generate as much data as possible, then use external validation tools to filter out problematic responses.
D. Manually adjust the temperature parameter in the UI to lower values to reduce randomness and increase coherence in the generated text.

Answer: D

Question: 239

In the lifecycle of deploying a prompt template for a generative AI solution, which of the following best describes the stage where user feedback is integrated to refine the template's performance?

A. Initial testing on synthetic datasets and model validation
B. Deployment to production with regular monitoring and logging
C. Iterative prompt tuning based on A/B test results and feedback loops
D. Retraining the model based on emerging trends in data

Answer: C

Question: 240

You are developing a legal research assistant that uses Retrieval-Augmented Generation (RAG) to help lawyers find relevant case law. The system must provide case law that is not only relevant to the query but also semantically similar in meaning to the legal terminology used. Given this scenario, which of the following retrievers would be the most appropriate choice for this application?

A. A sparse retriever that integrates exact matching of legal terms with synonym matching.
B. A sparse retriever based on keyword matching and BM25 algorithm.
C. A rule-based retriever optimized for simple, pre-defined keyword mappings.
D. A dense retriever based on embeddings from a fine-tuned BERT model.

Answer: D

Question: 241

Which of the following techniques is the most effective for reducing bias in generative AI models through prompt engineering?

A. Allowing the model to auto-correct its own responses by cross-referencing with other outputs
B. Applying Greedy Decoding to ensure the most likely tokens are selected during generation
C. Using neutral and carefully phrased prompts to avoid triggering biased outputs
D. Fine-tuning the model using training data that explicitly includes examples of biased outputs

Answer: C

Question: 242

You are working on a large-scale enterprise application using IBM watsonx and need to ensure that different versions of your generative AI model prompts are properly managed for deployment. Which of the following is the most appropriate action when planning the deployment of prompt versions?

A. Use a deployment space in IBM watsonx to version your prompts, assigning unique tags to each version and ensuring rollback capabilities.
B. Store all prompt versions directly in the model's code repository, updating the main branch with each new version.
C. Embed the prompt version directly into the API request body so that the deployed model can select the correct prompt dynamically at runtime.
D. Keep prompt versions in an external document management system and manually track which versions are deployed in the application.

Answer: A

Question: 243

You are tasked with creating structured prompts in IBM watsonx Prompt Lab to ensure consistency and reliability in generating responses from a machine learning model. What is one major benefit of using structured prompts when experimenting with different prompts in Prompt Lab?

A. Using structured prompts makes AI responses deterministic, ensuring that no creativity is applied in the generated text.
B. Structured prompts provide a template that ensures consistent intent, which can be adjusted with dynamic variables for flexibility.
C. Using structured prompts allows the AI to autonomously generate diverse outputs with minimal user input.

D. Structured prompts ensure that the AI generates longer and more detailed responses by default.

Answer: B

Question: 244

You are tasked with improving the performance of a generative AI model used for customer service automation. The model needs to respond quickly and with high accuracy, particularly for complex queries. You have access to Tuning Studio as part of your optimization toolkit. Which of the following is a primary benefit of using Tuning Studio to optimize the model in this scenario?

A. It provides automated fine-tuning of the model's hyperparameters to improve performance on domain-specific tasks.
B. It enables the creation of new datasets by generating synthetic data based on prompts.
C. It automates the process of cleaning and preprocessing the input data before model training.
D. It allows you to manually edit the output tokens to ensure correctness.

Answer: A

Question: 245

You are tasked with inspecting and validating a dataset using IBM Watson's Data Refinery tool. The dataset will be used for an AI application that performs natural language processing (NLP) on customer support tickets. Upon inspection, you discover that the 'ticket_description' field contains numerous instances of null values and inconsistent formats. What is the most appropriate action to take next using Data Refinery?

A. Use Data Refinery's "Data Quality" feature to generate a profiling report and assess the scope of the issues.
B. Delete all rows containing null values in the 'ticket_description' field.
C. Proceed with training the model without any changes, as missing data will not impact NLP models significantly.
D. Replace all null values with zeros, as this will standardize the dataset.

Answer: A

Question: 246

: You are tasked with fine-tuning a large language model (LLM) to perform sentiment analysis on product reviews. The dataset contains customer reviews, but some reviews are very short, and others contain irrelevant data like product specifications or spam. You want to prepare the dataset for fine-tuning by ensuring the data is clean, relevant, and representative of the task at hand. Which of the following steps is most critical to ensure the dataset is suitable for fine-tuning?

A. Randomly remove 10% of the dataset to reduce noise
B. Filter the dataset to remove irrelevant or outlier reviews
C. Increase the learning rate to adjust for inconsistencies in the dataset
D. Increase the number of epochs to handle the noisy data

Answer: B

Question: 247

In a Retrieval-Augmented Generation (RAG) system, you are tasked with generating vector embeddings for a large corpus of documents. You plan to use a pre-trained transformer-based model to generate these embeddings. What is the most important factor to consider when choosing a pre-trained model for generating embeddings in this scenario?

A. The model should have been trained on a similar task (e.g., document retrieval), as this ensures the embeddings will be relevant for your corpus.
B. The model should generate embeddings based on sentence-level inputs only, as document-level embeddings are always too large for effective retrieval.
C. The model's size (in terms of parameters) should be minimized to reduce memory usage, even if it impacts embedding quality.
D. The model's embeddings should always be fine-tuned on your specific corpus before use, as pre-trained embeddings are too general for most tasks.

Answer: A

Question: 248

While preparing a fine-tuning project using IBM watsonx, you want to generate synthetic data via the User Interface (UI) to supplement your existing dataset. Which of the following describes an option supported by IBM watsonx for synthetic data generation through the UI?

A. Uploading real-time streaming data for immediate synthetic generation
B. Using pre-built templates to generate domain-specific synthetic data
C. Manually coding data augmentation scripts within the UI
D. Directly modifying the architecture of the model to generate task-specific synthetic data

Answer: B

Question: 249

You are developing a chatbot application that uses IBM watsonx to assist users with customer support. The chatbot needs to respond with accurate, up-to-date information from a large corpus of documents. The documents include unstructured data, such as support tickets, product guides, and troubleshooting steps. Due to the need for highly specific answers, the system should be able to retrieve relevant information from the document base, while also generating human-like responses using a large language model (LLM). In this scenario, which configuration should be used to ensure the chatbot retrieves the most relevant context before generating a response, and why would this approach be beneficial?

A. Use a dense retriever with a vector database for embedding-based retrieval, followed by using the LLM for context generation.
B. Use a term-based retriever with an inverted index and perform keyword-based search, followed by minimal model tuning for response generation.

C. Use a sparse retriever and a standard SQL database, then fine-tune the language model to perform context matching.
D. Use a knowledge graph-based retriever integrated with the LLM, leveraging structured relationships between data points.

Answer: A

Question: 250

How can developers leverage Prompt Lab in IBM Watsonx to build reusable prompts for a conversational AI system that handles multiple topics?

A. By generating a unique prompt for each possible user query to ensure maximum precision.
B. By designing prompts with dynamic variables for user-specific information, allowing for flexibility across different topics.
C. By utilizing Prompt Lab's automatic suggestion feature, which builds new prompts based on the system's training data.
D. By creating a single prompt template that adapts to all user inputs regardless of the subject.

Answer: B

Question: 251

You are implementing a Retrieval-Augmented Generation (RAG) system to enhance a large language model's (LLM) ability to answer questions based on an external document store. What role do embeddings play in the RAG architecture, and how can you optimize them for more relevant document retrieval? (Select two)

A. The cosine similarity metric is typically less effective than Euclidean distance for comparing embeddings in RAG systems.
B. Increasing the dimensionality of embeddings always improves retrieval accuracy by providing a more detailed representation of the documents.
C. Fine-tuning the embedding model with task-specific data can lead to more accurate retrievals.
D. Embeddings are vector representations of text, used to measure the similarity between queries and documents.
E. Embeddings are only used during the generation phase of RAG to enhance language model outputs.

Answer: C,D

Question: 252

When fine-tuning a model in Tuning Studio, which of the following is a key advantage of this tool in reducing resource costs while improving model performance?

A. It expands the model's architecture to handle larger datasets.
B. It allows incremental training, saving computational resources by reusing checkpoints.

C. It automatically increases the number of layers for more complex tasks.
D. It optimizes the model for multilingual capabilities by adding new embeddings.

Answer: B

Question: 253

A generative AI team is optimizing a model for text summarization. The team uses both hard prompts and soft prompts during training. What is the primary reason a soft prompt might provide less explainability compared to a hard prompt?

A. Soft prompts involve complex embedding vectors, which are learned and difficult for humans to interpret directly.
B. Soft prompts are used exclusively in unsupervised learning settings, where explainability is inherently lower due to lack of labeled data.
C. Soft prompts include manually designed templates that change dynamically, reducing transparency.
D. Soft prompts are explicitly written by humans and include domain-specific instructions, making them harder to explain.

Answer: A

Question: 254

You are using IBM Watsonx to control the randomness of a language model's output by adjusting the top-k parameter. What happens when you reduce the top-k value from 50 to 5 during text generation?

A. Reducing the top-k value increases the temperature of the model, making the outputs more creative and diverse.
B. The model will now consider only the top 5 most probable tokens at each step, making the output more deterministic and focused.
C. Lowering the top-k value forces the model to generate shorter outputs by limiting the number of tokens available for selection.
D. Reducing the top-k value reduces the model's ability to predict rare or uncommon words, leading to less accurate outputs.

Answer: B

Question: 255

A healthcare organization is deploying a generative AI model to assist in summarizing patient records. Given the sensitivity of patient data, the model must not output Personally Identifiable Information (PII) such as names, addresses, or Social Security numbers. Which of the following strategies would best help prevent the model from unintentionally generating PII?

A. Fine-tune the model specifically on anonymized datasets with PII removed.
B. Increase the temperature and top-k values to reduce the probability of PII being generated.
C. Disable the model's ability to use context from previous tokens to prevent PII from being included in the output.

D. Use a post-processing step to detect and mask PII after the model generates the text.

Answer: A

Question: 256

You are tasked with securing an endpoint for a generative AI model that interacts with external applications. Which of the following practices is MOST effective in ensuring both security and stability of the model's endpoint in a production environment?

A. Implement strict IP whitelisting and authentication, but allow unfiltered prompts for flexibility.
B. Allow all external applications unrestricted access to the endpoint to facilitate wider model usage.
C. Focus exclusively on endpoint authentication without enforcing any rate limits to prevent user frustration.
D. Limit the number of API calls per minute, restrict prompt lengths, and apply input validation to prevent malicious queries.

Answer: D

Question: 257

You are tasked with improving the performance of a generative AI model that generates personalized marketing emails. The client wants the model to produce more relevant and targeted emails based on user behavior while keeping token usage and computational costs low. You decide to use Tuning Studio to achieve this. Which of the following is a key benefit of using Tuning Studio in this scenario?

A. It provides tools for manual annotation of data to improve model accuracy.
B. It automatically generates custom datasets for training without needing labeled data.
C. It allows the fine-tuning of the model’s hyperparameters based on the specific domain, improving relevance and reducing token generation costs.
D. It increases the model's maximum token limit, allowing for more extensive outputs without sacrificing performance.

Answer: C

Question: 258

Consider an organization implementing a RAG system to enhance the accuracy of their internal documentation search tool. The retriever is responsible for fetching relevant documents based on user queries. What is the core capability of the retriever in this context?

A. To perform entity recognition and classify documents based on specific keywords, ignoring the overall document meaning.
B. To return relevant documents or passages from a knowledge base by calculating semantic similarity between the query and documents.
C. To generate new responses based on training data, without relying on external data sources.
D. To perform pre-defined template matching on queries to retrieve documents that exactly match pre-configured templates.

Answer: B

Question: 259

You are tasked with building a Retrieval-Augmented Generation (RAG) system to assist users in retrieving relevant documents from a vast knowledge base. The first step in this process is to generate vector embeddings for the documents using a pre-trained model. After generating embeddings, you notice that the model is sometimes failing to retrieve semantically similar documents. Which of the following is the most appropriate approach to ensure that semantically similar documents are retrieved effectively?

A. Use Greedy Decoding during the embedding generation to avoid irrelevant tokens in the vectors. B. Choose a model with a smaller embedding dimension to reduce the memory footprint of embeddings.
C. Convert all documents into embeddings using cosine similarity directly instead of using a vector search algorithm.
D. Fine-tune the model on a task-specific dataset to improve the quality of the embeddings for your domain.

Answer: D

Question: 260

When preparing a dataset for fine-tuning a large language model for a named entity recognition (NER) task, which of the following preprocessing steps is most critical for ensuring accurate entity classification?

A. Remove rare entities to improve model performance on common entities
B. Randomly shuffle the dataset before training to increase diversity
C. Use sentence segmentation to isolate each named entity in its own sentence
D. Ensure proper tokenization of the dataset according to the model's vocabulary

Answer: D

Question: 261

You are tasked with optimizing a large language model (LLM) for deployment in a resource-constrained environment where memory usage and computational cost need to be minimized without significantly compromising model accuracy. Which quantization technique would be the most appropriate to achieve this balance?

A. Full integer quantization with 8-bit precision
B. Mixed-precision floating-point quantization
C. Post-training dynamic quantization
D. Post-training weight clustering

Answer: C

Question: 262

You are tasked with building a question-answering system using IBM WatsonX's LLM integrated with LangChain. The system needs to retrieve relevant documents from a corpus of research papers and generate accurate, contextually aware responses. Which of the following steps is most critical when implementing a RAG pattern in this environment?

A. Configuring LangChain to use WatsonX's LLM for exact keyword matching during document retrieval.
B. Using WatsonX LLM to pre-process the research papers before storing them in LangChain's memory.
C. Setting up the vector database to store embeddings of the research papers for semantic search.
D. Fine-tuning WatsonX's LLM on the research papers before enabling retrieval through LangChain.

Answer: C

Question: 263

You are analyzing prompts submitted to a Generative AI model used for summarizing long research papers. One user submits the following prompt:
"Summarize this 40-page research paper on quantum computing, including details on every section and subsection, providing a detailed description of key points, methodologies, results, discussions, and future work. The summary should be at least 5 pages long."
Why is this prompt considered inefficient, and how should it be optimized?

A. The prompt is inefficient because it does not specify a character limit, which means the model might generate overly verbose output.
B. The prompt is inefficient because it requests a 5-page summary, which is unnecessary for summarizing the key information from a research paper.
C. The prompt is inefficient because it asks for too much detail across all sections, leading to excessive token usage and unnecessary information in the output.
D. The prompt is efficient because it clearly outlines the expectations and ensures a comprehensive summary.

Answer: C

Question: 264

You are tasked with integrating IBM watsonx with an existing enterprise application that uses a custom-trained Large Language Model (LLM) to answer complex customer queries. The enterprise application requires real-time responses from the LLM, and the integration must allow for scalable, low-latency interactions across multiple customer channels, such as email and live chat. You need to ensure that the data flowing into the LLM is preprocessed appropriately and that the orchestration between different Watson services and the LLM is efficient. What is the best approach for integrating IBM watsonx to meet these requirements?

A. Use IBM watsonx's Generative AI API and directly integrate it with the application via REST,

ensuring the LLM receives real-time data from each channel.
B. Integrate IBM watsonx Assistant to handle multi-channel inputs and orchestrate LLM responses, while using IBM Event Streams to handle real-time scalability across channels.
C. Directly implement IBM watsonx Machine Learning models into each communication channel to ensure low-latency interactions with the LLM.
D. Employ IBM watsonx's Data Refinery tool to preprocess incoming data from each channel and orchestrate data flow through Apache Kafka for real-time processing.

Answer: B

Question: 265

When addressing bias in a generative AI model, which of the following strategies is least likely to be effective in reducing biased outputs during text generation?

A. Training the model on a diverse and representative dataset
B. Using temperature control during generation to manage diversity in responses
C. Leveraging prompt rephrasing techniques to remove bias-inducing keywords or phrases
D. Incorporating fairness constraints during the model's training phase

Answer: B

Question: 266

You are deploying a new version of a generative AI model in IBM Watsonx, and you want to maintain the integrity of prompt versioning throughout the deployment lifecycle. Which of the following methods is the most effective for ensuring that the correct prompt version is used with the corresponding model version in production?

A. Implement model registry tags that associate a specific prompt version with each model version during deployment.
B. Use semantic versioning for both the model and the associated prompts, and track them independently in separate systems.
C. Use the latest available prompt version for every deployment, without specifying an exact version D. Hard-code the prompt within the model deployment script to ensure that the correct prompt is always used

Answer: A

Question: 267

A data scientist is choosing between using hard prompts and soft prompts in a generative AI project. Which of the following best explains why hard prompts might be more suitable for scenarios where explainability is crucial?

A. Hard prompts are based on learned embeddings, which offer better model understanding due to their complexity.
B. Hard prompts allow for a clear, human-readable set of instructions that directly guide the model's

behavior.
C. Hard prompts reduce the model's flexibility by making the output deterministic, which enhances explainability.
D. Hard prompts dynamically adjust the model's internal representations, providing more clarity in complex situations.

Answer: B

Question: 268

Your team is tasked with enhancing a document search engine using IBM watsonx Discovery. The search engine will be used to help employees quickly find relevant internal documentation, such as policy guides, project specifications, and technical manuals. You have decided to use Retrieval- Augmented Generation (RAG) to enhance this search engine's performance. The current focus is on selecting the best retriever for the task and setting up a vector database to store document embeddings. Which of the following retriever types is most suitable for this RAG-based search engine setup, considering the diverse and unstructured nature of the documents, and why?

A. Sparse retriever with BM25 ranking in combination with a traditional relational database.
B. Sparse retriever with BM25 ranking in combination with a traditional relational database.
C. Hybrid retriever that combines sparse retrieval methods with dense embeddings stored in a flat-file system.
D. Rule-based retriever using manually crafted rules to find documents based on predefined logic.

Answer: A

Question: 269

You are tasked with managing multiple versions of a generative AI model in a production environment. The team has decided to utilize IBM watsonx deployment spaces to version control both models and prompts. What is the most efficient way to organize these assets within the deployment space?

A. Use a third-party version control tool like Git to manage prompt and model versions, syncing with the deployment space only for final deployment.
B. Create separate deployment spaces for each model and prompt version, linking the spaces manually to ensure consistency between model and prompt versions.
C. Store all versions of both models and prompts in a single deployment space, and use comments to track differences between versions.
D. Assign version tags to both models and prompts within a single deployment space, using metadata and associations to track prompt-model pairings for each version.

Answer: D

Question: 270

Which of the following best describes the difference between a hard prompt and a soft prompt in the context of generative AI optimization?

A. A hard prompt is used for deterministic output, while a soft prompt allows for random sampling.
B. A hard prompt directly modifies the model architecture, while a soft prompt changes the model's training data.
C. A hard prompt encodes the model's hyperparameters, while a soft prompt adjusts the learning rate dynamically.
D. A hard prompt consists of static, predefined instructions, while a soft prompt is optimized using continuous input embeddings that adapt based on the training context.

Answer: D

Question: 271

In a generative AI model, you are tasked with producing creative yet coherent text for a marketing campaign. You want to ensure that the output contains varied word choices and diverse sentence structures while still maintaining some degree of logical consistency. Which of the following settings for the temperature parameter would most likely achieve this balance?

A. Temperature = 2.0
B. Temperature = 0.7
C. Temperature = 0.2
D. Temperature = 0.0

Answer: B

Question: 272

During the deployment of a generative AI model using IBM Watsonx, you are responsible for ensuring that the model does not generate content containing hate speech, abusive language, or profanity. However, in some scenarios, the model might bypass basic filtering mechanisms and produce subtle or context-specific harmful language. Which of the following best addresses edge cases where hate speech or abuse may slip through more basic filtering mechanisms?

A. Use IBM Watsonx's HAP (Hate, Abuse, and Profanity) filter, which is designed to handle context-specific and subtle forms of harmful language.
B. Implement regular expression-based filters that cover a wide range of variations in harmful language.
C. Use sentiment analysis to detect and block negative content that may indicate hate speech or abuse.
D. Continuously monitor and retrain the model on datasets containing edge cases of hate speech, abuse, and profanity.

Answer: A

Question: 273

You are tasked with optimizing a generative AI model in IBM watsonx.ai for an NLP-based application. During the planning stage, which of the following data elements is most important to ensure the model generalizes well to real-world application usage?

A. High-dimensional feature vectors from a small, well-curated dataset
B. Only the structured data relevant to specific business rules
C. A large, diverse dataset with both structured and unstructured data
D. Pre-processed data with stop words removed to reduce noise

Answer: C

Question: 274

You are refining a prompt for an AI model that generates financial reports using IBM watsonx Prompt Lab. To ensure the AI produces structured outputs such as sections on “Revenue,” “Expenses,” and “Net Profit,” which of the following is the most effective prompt editing option to achieve this goal?

A. Increase the randomness parameter to encourage diverse sections in the report structure.
B. Use static prompt text and avoid variables to maintain a consistent structure for each report.
C. Adjust the temperature setting to generate more creative and detailed reports.
D. Incorporate specific section headers as part of the prompt to guide the model in generating the desired report format.

Answer: D

Question: 275

How can developers best ensure that prompt templates are version-controlled and any changes to the templates can be tracked efficiently in a collaborative team environment?

A. Implement a versioning system for prompt templates in a source control tool like Git
B. Manually keep a changelog of prompt template modifications in a shared document
C. Use a centralized cloud storage system that automatically saves all versions of the prompt templates
D. Update prompt templates directly in production and rely on error logs for tracking changes

Answer: A

Question: 276

Which of the following best describes the benefit of using prompt variables when developing generative AI models in IBM Watsonx?

A. Prompt variables ensure that the same text input will always yield the same output, improving consistency.
B. Prompt variables allow for dynamic input customization without needing to manually modify the core prompt, increasing flexibility.
C. Using prompt variables in Watsonx allows the model to learn from real-time data input, enabling self-training over time.
D. Prompt variables reduce computational load by optimizing model performance, improving overall system efficiency.

Answer: B

Question: 277

You are managing a generative AI model deployment in IBM Watsonx and need to implement prompt versioning to ensure traceability and reproducibility of model behavior over time. Which of the following strategies best enables versioning of prompts during deployment?

A. Relying on model checkpointing to manage both model weights and prompts.
B. Storing prompts in a flat file system and manually tracking versions.
C. Using a source control system (e.g., Git) to track prompt changes alongside model code.
D. Disabling versioning for prompts since it is not required for generative models.

Answer: C

Question: 278

You are implementing a RAG system that uses vector embeddings to retrieve relevant information from a large corpus of documents. To store and efficiently query the vector embeddings, you need to choose a suitable vector database. Which of the following is a key capability that a vector database must support to ensure efficient retrieval in this context?

A. Built-in support for Approximate Nearest Neighbor (ANN) search algorithms
B. The ability to store and query data only in JSON format
C. The ability to store sparse vectors instead of dense vectors
D. Automatic generation of vector embeddings from raw text data

Answer: A

Question: 279

Which two of the following are key benefits of using IBM Watsonx's Tuning Studio for fine-tuning generative AI models on specific tasks? (Select two)

A. Automatically selects the best model architecture for the task
B. Significantly reduces the risk of overfitting without manual intervention
C. Automatically generates prompt templates based on data input
D. Offers real-time validation of tuning experiments to optimize model performance
E. Provides automated hyperparameter search and tuning strategies

Answer: D,E

Question: 280

You are designing a prompt for a generative AI model in a customer support chatbot. The goal is to help users troubleshoot issues with a software installation. Which of the following prompts would be the most effective for guiding the AI to deliver clear, actionable troubleshooting steps?

A. "Can you tell the user about software installation?"

B. "Guide the user through troubleshooting software installation issues, starting from common problems like insufficient disk space and incompatible OS."
C. "Tell the user what software installation is and how it works."
D. "Please list all the possible errors that occur during software installation."

Answer: B

Question: 281

You are implementing a Retrieval-Augmented Generation (RAG) system using IBM Watsonx to improve your generative AI model. The system retrieves relevant information from a large corpus and augments it into the generative process. In this context, what role do embeddings play in a RAG-based system?

A. Embeddings ensure that only syntactically correct documents are retrieved, without regard to semantic content.
B. Embeddings are used to retrieve relevant documents by calculating the semantic similarity between user queries and the stored documents.
C. Embeddings reduce the size of the generative model by compressing the parameters into a smaller representation.
D. Embeddings store the entire content of documents, which is then directly passed to the generative model.

Answer: B

Question: 282

When working with IBM Watsonx Generative AI models, it's important to configure proper stopping criteria to control when the model should terminate the text generation process. You are developing a chatbot where responses should stay within a manageable length without losing coherence. Which configuration best represents an effective stopping criterion to ensure coherent responses without abrupt truncation?

A. Beam search decoding with a stop sequence of "END" and a maximum tokens limit of 50.
B. Greedy decoding with maximum tokens set to 20 and a stop sequence of "END".
C. Greedy decoding with temperature set to 2.0 and no stop sequence.
D. Greedy decoding with no stop sequence and maximum tokens set to 200.

Answer: A

Question: 283

A company is using IBM's InstructLab to fine-tune a large language model (LLM) to perform customer support tasks, such as answering frequently asked questions (FAQs) and troubleshooting product issues. Which of the following components of InstructLab plays the most crucial role in ensuring the model learns to align its responses with the specific format and tone required for customer interactions?

A. The prompt-tuning engine, which fine-tunes model outputs based on pre-defined instructions.
B. The instruction optimizer, which tunes hyperparameters to improve task-specific performance.

C. The user simulation environment, which provides real-time testing of model outputs.
D. The evaluation metrics dashboard, which tracks model performance on customer interaction tasks.

Answer: A

Question: 284

You are configuring an LLM for a product recommendation chatbot. The goal is to balance creativity and relevance, ensuring the chatbot suggests diverse but appropriate products. Which combination of model parameters will best achieve this? (Select two)

A. Set a high penalty for repetition to encourage varied recommendations
B. Set the temperature to 0.1 for highly deterministic responses
C. Increase the temperature to 1.5 to maximize creativity in suggestions
D. Apply a low top-k value (e.g., k=10) to restrict randomness
E. Use a top-p (nucleus) sampling value of 0.95 for diverse, relevant outputs

Answer: D,E

Question: 285

In a Retrieval-Augmented Generation (RAG) system, the retriever plays a crucial role in retrieving relevant documents or passages from a knowledge base. What is the primary function of a retriever in this architecture?

A. The retriever generates the final response based on the retrieved documents and the user's query. B. The retriever finds and retrieves documents from a knowledge base that are semantically similar to the input query, often using vector embeddings to compare similarity.
C. The retriever performs tokenization and splitting of large documents into smaller chunks, making them easier to index.
D. The retriever modifies the transformer-based model's weights in real-time to improve the relevance of future retrievals.

Answer: B

Question: 286

You are working with a healthcare provider to deploy a generative AI model that assists in diagnosing patient conditions based on medical history and symptoms. The healthcare provider is concerned about ethical use, bias, and compliance with health data regulations such as HIPAA. Which governance approach is the most critical to ensure both compliance and ethical AI deployment?

A. Ensure data encryption during model inference to protect patient information from unauthorized access.
B. Implement a bias detection framework to identify and mitigate any biases present in the model's outputs.
C. Apply a federated learning approach to train the model without directly accessing the healthcare

provider's patient data.
D. Use a cloud-based infrastructure with HIPAA certification to deploy the model, ensuring compliance with healthcare regulations.

Answer: B

Question: 287

You are tasked with improving the performance of a Retrieval-Augmented Generation (RAG) system in IBM watsonx. Part of this improvement involves selecting the right embedding model for document retrieval. Which of the following is the best description of the differences between various embedding models, and how would you choose the most suitable model for your task?

A. Word2Vec embeddings capture only the syntactic relationships between words, while BERT embeddings focus on both syntax and semantic context, making BERT more suitable for complex retrieval tasks in a RAG system.
B. BERT embeddings are context-independent, which makes them less useful for a RAG system than Word2Vec or GloVe, which focus on learning semantic relationships between words.
C. TF-IDF is an advanced embedding model that captures both the frequency and semantic meaning of words, making it more effective than deep learning-based models like BERT for retrieval in RAG systems.
D. Word2Vec, GloVe, and BERT are all embedding models, but BERT embeddings capture richer context by considering the entire sentence rather than just the local context, making it more effective for generating semantically relevant embeddings.

Answer: D

Question: 288

When tuning the generative model parameters, which of the following scenarios describes an appropriate use of the maximum tokens setting, and how will it influence the model's output?

A. Setting the maximum tokens to 100 will prevent the model from generating coherent sentences, as it cuts off abruptly after 100 tokens
B. Setting the maximum tokens to 500 will guarantee that the model always generates exactly 500 tokens
C. Setting the maximum tokens to 0 will allow the model to generate an unlimited amount of text
D. Setting the maximum tokens to 100 ensures that the model's output will not exceed 100 tokens, but it may stop earlier if the generation ends naturally

Answer: D

Question: 289

Which of the following is the most effective approach when planning for data elements to optimize application usage in IBM watsonx generative AI models?

A. Ensure that all features are included in the model to capture as much data context as possible,

regardless of their relevance.
B. Select only high-dimensional features to increase the complexity of the model and boost its predictive power.
C. Use feature selection techniques to reduce dimensionality, enhancing model efficiency without sacrificing performance.
D. Aggregate similar data types to minimize the need for feature selection during model optimization.

Answer: C

Question: 290

Which of the following represents the most effective use of example input prompts within IBM Watsonx's Prompt Lab for generating a high-quality response?

A. Using example prompts that introduce multiple topics at once to evaluate how well the model handles multitasking during response generation.
B. Writing prompts that are extremely vague, allowing the model to freely interpret the input and generate diverse responses.
C. Crafting prompts with very specific and detailed instructions, ensuring the model follows a strict framework for the desired response.
D. Using prompts with complex sentence structures and advanced terminology to challenge the model's understanding capabilities.

Answer: C

Question: 291

When creating a prompt-tuned model, which of the following factors is most critical to ensure that the model generates accurate and contextually relevant responses?

A. Ensuring that the prompt length does not exceed the model's input token limit.
B. Employing low learning rates to prevent the model from adapting too quickly to new prompts.
C. Matching the training data prompts as closely as possible to real-world deployment scenarios.
D. Using a diverse set of unrelated tasks in the tuning dataset to increase model generalization.

Answer: C

Question: 292

You are tuning a generative AI model to control the length of the generated responses. Which of the following parameter configurations will ensure that the model generates responses that are at least 50 tokens long but no longer than 150 tokens?

A. Setting the minimum tokens to 150 and maximum tokens to 50
B. Setting no minimum token value but configuring the maximum tokens to 150
C. Setting the minimum tokens to 50 and maximum tokens to 150
D. Setting the maximum tokens to 50 and minimum tokens to 150

Answer: C

Question: 293

A business wants to deploy a customer service chatbot using IBM watsonx, integrated with multiple back-end systems including ERP, CRM, and a payment gateway. To manage this complex integration, the chatbot should dynamically switch between these systems based on the customer's intent. Which architecture best supports this requirement while ensuring scalability and minimal orchestration overhead?

A. Set up IBM watsonx to queue customer interactions and process them sequentially using a single system at a time.
B. Utilize IBM watsonx's API orchestration layer to dynamically route requests to the correct system based on intent analysis.
C. Implement a monolithic architecture where all system interactions are handled within a single service.
D. Use synchronous REST APIs for all back-end systems, relying on watsonx's built-in intent recognition to manage API routing.

Answer: B

Question: 294

You are working on a generative AI model that helps customers generate personalized responses to legal queries. The model is trained on a large corpus of publicly available legal documents. However, users often input personal information when interacting with the AI. What is the most effective strategy to mitigate the risk of exposing personal information in the model's responses?

A. Limit the model's ability to retain memory of previous user inputs by resetting its state after every query.
B. Apply a rule-based content filter to the model's outputs to remove any phrases that appear to contain personal information.
C. Use a privacy-preserving tokenization method to mask personal data in the input before feeding it into the model.
D. Train the model on anonymized data to ensure that personal information is never present in the training set.

Answer: C

Question: 295

You are working on optimizing a large language model (LLM) using quantization techniques. Your goal is to reduce memory usage while maintaining as much of the model's original accuracy as possible. What is a common challenge faced when applying quantization to LLMs, and how can it be mitigated?

A. Quantization causes a drastic reduction in training time, but increases memory usage. Use sparse quantization to address this.
B. Embedding layers in LLMs are difficult to quantize, so it's best to skip quantization for these layers

entirely.
C. Quantization may cause a significant drop in model accuracy, especially in embedding layers. Using quantization-aware training can help mitigate this.
D. LLMs are inherently resistant to quantization, so switching to a smaller model architecture is the only viable solution.

Answer: C

Question: 296

You are tasked with creating a prompt template for IBM Watsonx that will help generate product reviews for a new line of smartphones. The prompt needs to be adaptable across various product features and sentiment (positive, neutral, or negative) while maintaining a consistent structure. Which of the following prompt templates would be the most effective for generating well-structured, detailed reviews?

A. “Generate a review mentioning the best features of the smartphone.”
B. “Write a [sentiment] review of the [product name], focusing on the [specific feature]. Include reasons for the sentiment and detailed examples.”
C. “Write a review about the smartphone.”
D. “Tell a story about the user’s experience with the smartphone.”

Answer: B

Question: 297

In the context of a Retrieval-Augmented Generation (RAG) system, which type of retriever is best suited for retrieving documents based on semantic similarity in a vector space?

A. Boolean retriever, which uses logical operators (AND, OR, NOT) to filter documents based on keyword presence.
B. Hierarchical retriever, which first applies keyword-based filters and then ranks documents using machine learning models.
C. Dense retriever, which converts both the query and documents into vectors and retrieves based on similarity in the vector space.
D. Exact match retriever, which retrieves documents that exactly match the query string.

Answer: C

Question: 298

A financial institution is deploying a generative AI model to generate loan approval recommendations based on applicant profiles, including factors like income, credit score, and employment history. The organization is concerned about ensuring that the model does not introduce bias in its recommendations, particularly related to gender and race. You have been asked to design a process to evaluate the model's inferences during deployment and mitigate any potential bias. Which method would be most effective for evaluating the model's inferences for bias in this deployment scenario?

A. Implement a fairness audit, where a sample of the model’s inferences is checked for disparate

impact across protected groups such as gender and race.
B. Manually review all loan decisions generated by the model for signs of bias before releasing them to customers.
C. Use greedy decoding in the inference phase to ensure deterministic outputs, avoiding potential bias from probabilistic sampling methods.
D. Periodically retrain the model with updated datasets that exclude sensitive attributes such as gender and race.

Answer: A

Question: 299

You are tasked with building a Retrieval-Augmented Generation (RAG) system for answering legal questions. The legal documents vary significantly in complexity and structure. How would you optimize embeddings in this domain to ensure the system retrieves the most relevant documents? (Select two)

A. Rely solely on word-level embeddings to capture the meaning of legal phrases and concepts.
B. Train a domain-specific embedding model using legal documents to better capture the nuances of legal terminology.
C. Apply dimensionality reduction techniques like PCA to compress embeddings and improve retrieval speed.
D. Use an unsupervised learning approach to generate embeddings, as labeled data is not necessary for improving retrieval performance.
E. Integrate additional metadata (e.g., document date, author) into the embedding representation to improve retrieval.

Answer: B,E

Question: 300

You are training a chatbot to handle customer inquiries for a telecommunications company. To augment your training data, you decide to generate synthetic data using IBM's InstructLab platform. You aim to improve the model's ability to handle rare or edge-case scenarios, such as technical issues with specific device models. You are following the LAB (Large-scale Alignment for chatBots) methodology to ensure alignment of the chatbot's responses with company policies. Which of the following steps is most aligned with LAB methodology principles for generating synthetic data in this case?

A. Generate synthetic data without aligning it with company policies to maximize diversity
B. Generate synthetic customer inquiries and responses based on company policy guidelines
C. Manually write every possible customer inquiry scenario to ensure quality
D. Use random generation to create a diverse range of responses without constraints

Answer: B

Question: 301

You are developing a solution that requires generating various reports using IBM Watsonx. To maintain efficiency and avoid redundant prompt creation, you decide to implement reusable prompts. Which of the

following best summarizes the value of using reusable prompts in this context?

A. They allow for faster deployment of new AI-driven solutions since the same prompt templates can be adapted for multiple tasks with minimal changes.
B. Reusable prompts eliminate the need for monitoring model performance since they consistently deliver accurate results.
C. Reusable prompts automatically adjust their structure depending on the input, eliminating the need for dynamic parameters or customization.
D. Reusable prompts ensure that the AI model continues to evolve without additional training, as the prompts themselves become more refined over time.

Answer: A

Question: 302

When leveraging existing data for fine-tuning an LLM in IBM watsonx, you want to optimize the model for a highly specialized domain. You also want to generate additional synthetic data to augment your dataset. Which of the following approaches would best help you achieve your goal?

A. Using the watsonx UI to generate synthetic data that mirrors your existing dataset, filling any data gaps
B. Relying exclusively on pre-trained general models without making domain-specific modifications
C. Manually crafting complex datasets by sampling individual instances from unrelated domains
D. Using unsupervised learning on your existing dataset without adding synthetic data

Answer: A

Question: 303

When deploying a generative AI model using IBM watsonx, which of the following actions should be prioritized to ensure high availability and minimize downtime in production environments? (Select two)

A. Disabling failover mechanisms to reduce deployment complexity
B. Configuring auto-scaling policies based on usage patterns and resource demand
C. Setting up redundant storage for the model's dataset but not for the model itself
D. Implementing containerization for the model through Kubernetes
E. Deploying the model on a single cloud-based server with load balancing enabled

Answer: B,D

Question: 304

You are designing a customer support chatbot using watsonx.ai as the primary generative model. You want to enhance the chatbot's capabilities by integrating it with IBM Watson Assistant to handle structured conversations while allowing watsonx.ai to generate responses for open-ended queries. Which integration approach would most effectively combine both services while maintaining optimal performance and accuracy?

A. Create separate chat interfaces for Watson Assistant and watsonx.ai, and allow the user to choose

which system to query based on their needs.
B. Train Watson Assistant to handle both structured and unstructured conversations, while using watsonx.ai only for rare edge cases that Watson Assistant cannot manage.
C. Implement Watson Assistant for structured conversations and use a middleware layer that dynamically routes complex, open-ended queries to watsonx.ai, returning the results within the same session.
D. Use Watson Discovery to preprocess all queries before routing them to either Watson Assistant or watsonx.ai based on query complexity.

Answer: C

Question: 305

You are fine-tuning a generative AI model using Tuning Studio for a legal document analysis task. The model needs to perform well in summarizing long, complex legal texts while minimizing the amount of computational resources used. You are tasked with optimizing the tuning process to ensure maximum efficiency and model accuracy. Which of the following actions would most effectively optimize the tuning process in Tuning Studio for this task?

A. Adjusting the learning rate dynamically during training based on the model's real-time loss values.
B. Using a fixed number of training epochs to ensure consistent results across tuning iterations.
C. Reducing the number of layers in the model to decrease computational overhead.
D. Increasing the batch size to reduce the number of gradient updates during training, thus speeding up the process.

Answer: A

Question: 306

You are optimizing the generative AI model in IBM watsonx to balance creativity and coherence. You want to use a decoding method that dynamically adjusts the token probability threshold based on cumulative probabilities, thus ensuring the model generates coherent outputs while still allowing for some creativity. Which parameter should you adjust, and what is the optimal setting?

A. Set temperature to 1 and top-p to 0.2
B. Set temperature to 1.5 and top-p to 0.9
C. Set top-p to 0.9 and temperature to 0.7
D. Set temperature to 0 and top-p to 0.5

Answer: C

Question: 307

You are working on a task that involves generating marketing copy using IBM Watsonx. The goal is to craft a prompt that leads to detailed and persuasive content about a new product launch. Which of the following approaches would most likely result in high-quality, detailed, and contextually appropriate content?

A. Use a very short prompt: “Generate marketing copy for a product launch.”
B. Avoid specifying any constraints and rely on Watsonx’s default model behavior: “Write product launch marketing content.”
C. Use complex, technical jargon to generate highly specific content: “Produce syntactically dense prose with multifaceted aspects of ecological ramifications and commodification for a consumer base.” D. Provide specific context and audience information: “Generate marketing copy for a new eco-friendly water bottle targeting health-conscious consumers. Include persuasive language and focus on the sustainability features of the product.”

Answer: D

Question: 308

In the context of Generative AI, which of the following is an appropriate method for setting a stopping criterion to prevent a model from generating excessively long sequences?

A. Allowing the model to continue until it reaches the minimum token length specified
B. Stopping the generation when the model reaches the end of its training dataset
C. Stopping the model based on the number of unique tokens generated
D. Setting a maximum token length for the output sequence

Answer: D

Question: 309

You are working as a generative AI engineer and have developed a custom large language model (LLM) optimized for a specific use case. You are tasked with deploying this model on the IBM Watsonx platform. Which of the following steps is most essential to ensure the successful deployment of your custom model, given that the model uses a third-party transformer architecture?

A. Containerize the model using Docker or an equivalent containerization tool, ensuring that all required dependencies, such as transformers, tokenizers, and necessary packages, are included. B. Set up auto-scaling in the IBM Watsonx environment to handle large numbers of simultaneous model inference requests.
C. Modify the model to use IBM's proprietary transformer architecture, as third-party architectures are not supported by Watsonx.
D. Ensure that the model's training data is in a proprietary IBM format, as only Watsonx-specific formats are supported for custom model deployments.

Answer: A

Question: 310

You are building a RAG system in IBM watsonx for a legal document retrieval platform. The platform deals with highly structured legal texts and unstructured client queries. Your task is to ensure that both structured documents and unstructured queries can be efficiently retrieved using vector embeddings generated by the system. Which of the following actions will help you optimize the generation of vector embeddings to support accurate retrieval in the RAG system, considering the diverse data types?

(Select two)

A. Use a separate embedding model for legal documents and client queries to account for differences in data structure.
B. Fine-tune the embedding model on legal text data to ensure embeddings capture the specific structure and semantics of the documents.
C. Ensure that embeddings for legal documents are based solely on the document metadata, as it contains the most important information.
D. Generate embeddings for legal documents at the sentence level rather than the document level, as this increases precision in retrieval.
E. Rely on pre-trained general-purpose embeddings without any domain-specific adaptation, as they provide broad coverage across many industries.

Answer: A,B

Question: 311

You are tasked with integrating third-party embedding models into a Retrieval-Augmented Generation (RAG) system for document retrieval. Several models offer pre-trained embeddings that can be leveraged for a variety of downstream tasks. Which of the following third-party models is designed for generating embeddings that capture semantic meaning and context, making it ideal for a RAG-based GenAI system?

A. WordNet
B. BERT
C. GPT-3 Embeddings
D. Naive Bayes Classifier

Answer: B

Question: 312

In the context of AI governance, what is the most important aspect of managing model performance in a production environment to ensure compliance with regulatory and ethical guidelines?

A. Ensuring traceability of model decisions and providing auditability for each inference
B. Deploying the model only in secure, on-premises environments to prevent data breaches
C. Minimizing the model's inference time to optimize user experience
D. Maximizing the number of datasets the model is trained on to cover more use cases

Answer: A

Question: 313

When using Tuning Studio in IBM watsonx, which of the following types of models is most suitable for fine-tuning in a domain-specific task to optimize performance?

A. A pre-trained model that has been trained on a large, general-purpose dataset.

B. An untrained model with a minimal number of layers to reduce complexity.
C. A model trained from scratch on a small, specific dataset.
D. A pre-trained model with very few layers to ensure fast processing speeds.

Answer: A

Question: 314

You've conducted a prompt-tuning experiment, and after reviewing the generated outputs, you observe issues such as incomplete responses, irrelevant content, and occasional factual inaccuracies. What is the most appropriate action to address these data quality problems?

A. Increase the length of the input prompt to ensure that responses are more complete.
B. Introduce temperature tuning to adjust the randomness of the model's output and reduce irrelevant content.
C. Fine-tune the model on domain-specific data to improve factual accuracy and relevance.
D. Lower the model's perplexity score to improve both completeness and factual accuracy.

Answer: C

Question: 315

In the context of generative AI optimization, you are tasked with improving the model's response accuracy across different domains. One suggestion is to use soft prompts for enhanced performance. How does a soft prompt differ from a hard prompt in terms of implementation and flexibility?

A. Hard prompts are stored as vector embeddings within the model, while soft prompts are manually crafted by users to optimize generation.
B. Soft prompts are embedded during the model's training and rely on learned representations, while hard prompts are explicit text-based inputs given during inference.
C. Soft prompts use a rule-based system to guide the model, while hard prompts depend on statistical techniques to generate responses.
D. Soft prompts can dynamically change during inference based on the input data, whereas hard prompts are fixed throughout the entire generation process.

Answer: B

Question: 316

A developer is using a GitHub Code Retrieval API to help build a search engine that can locate relevant code snippets from public repositories. The API is designed to retrieve code based on the semantic similarity of the query (e.g., a description of what the code does) to the code itself. What is the primary advantage of using a vector-based approach for code retrieval in this scenario?

A. It provides real-time updates to the code embeddings, ensuring the latest code is always retrieved.
B. It speeds up retrieval by limiting the search to repositories where the user has committed code in the past.
C. It retrieves code snippets based on the semantic similarity between the query and the code,

enabling the discovery of relevant code even when the query and code do not use the same keywords.
D. It ensures that the code snippets returned are exact matches to the keywords in the query, avoiding irrelevant code.

Answer: C

Question: 317

You are using IBM watsonx's generative AI model to generate responses for a chatbot, and you want to ensure that the model stops generating text when it encounters a specific phrase like "End of Response." Which of the following settings for stop sequences is most appropriate to achieve this goal?

A. Set the stop sequence to "\ n\ n"
B. Set the stop sequence to "<|stop|>"
C. Set the stop sequence to "End of Response"
D. Set the stop sequence to "STOP"

Answer: C

Question: 318

You are tasked with generating a product description for an e-commerce platform using a generative AI model. However, you notice that the generated text tends to repeat phrases excessively, leading to verbose output. To address this, you decide to adjust the model's temperature parameter. Which of the following changes would help reduce the repetitiveness of the generated text while maintaining a balance between creativity and coherence?

A. Increase the temperature from 0.5 to 1.5
B. Decrease the temperature from 0.9 to 0.3
C. Set the temperature to 0.0
D. Decrease the temperature from 0.8 to 0.6

Answer: B

Question: 319

You are building a Retrieval-Augmented Generation (RAG) system where documents are converted into embeddings. You decide to use a transformer-based model to convert your text into embeddings. The embeddings will later be used in a vector search engine. After generating the embeddings, you observe that similar documents are not being clustered closely in the vector space, leading to poor retrieval. What could be a likely reason for this behavior, and how can you address it?

A. The model has not been fine-tuned for generating document embeddings, leading to inaccurate representations.
B. The model's output is not normalized, which can cause issues during the vector search process.
C. The vector search algorithm is using Euclidean distance, which is inappropriate for high-dimensional embedding spaces.
D. The embedding vectors are being generated using sentence-level embeddings instead of word-level

embeddings.

Answer: B

Question: 320

You are tasked with fine-tuning a pre-trained large language model (LLM) using synthetic data generated through the IBM watsonx user interface. Which of the following steps should you follow to ensure the model is fine-tuned correctly and the synthetic data is used effectively?

A. Select the pre-trained model, generate synthetic data, inspect the generated data for quality, and fine-tune the model by adjusting hyperparameters and training settings.
B. Directly upload synthetic data without inspecting or validating it and initiate the fine-tuning process.
C. Select the pre-trained model, generate synthetic data, and fine-tune the model using default parameters without further customization.
D. Use synthetic data as a replacement for real-world data without cross-validation or any quality control measures.

Answer: A

Question: 321

You are designing a question-answering system that can provide responses based on a vast corpus of legal documents. Your solution leverages the Retrieval-Augmented Generation (RAG) pattern using an IBM watsonx-based architecture. The goal is to ensure that the system retrieves highly relevant documents and uses the retrieved content to generate accurate responses. The team is considering various libraries to integrate dense retrieval and generation models into this architecture. Which of the following libraries or frameworks should you integrate to effectively implement the RAG pattern in this scenario, considering the need for both document retrieval and natural language generation?

A. Hugging Face Transformers for language generation and FAISS for embedding-based document retrieval.
B. Keras for handling sparse retrieval and pre-trained generation models.
C. TensorFlow for dense retrieval and document embedding generation.
D. PyTorch for implementing rule-based retrieval and generative model fine-tuning.

Answer: A

Question: 322

You are working on tuning a generative AI model in IBM watsonx.ai for better performance in generating conversational responses. You have been asked to add new data to the project. What is the most effective approach to adding data for model tuning?

A. Add unstructured data with no labels for semi-supervised learning
B. Add only outlier data to improve the model's ability to handle edge cases
C. Add only the test set data to ensure robust evaluation of the model
D. Add new labeled training data that closely resembles the use case scenarios

Answer: D

Question: 323

You are working with a Generative AI model to generate a summary of a large financial report. To reduce costs, you are exploring different model parameters such as minimum and maximum token limits. Which configuration would help minimize generation costs while ensuring an accurate summary of the document?

A. Set both the minimum and maximum token limits to 1,000 to ensure the entire report is captured in detail, with no important information left out.
B. Set the maximum token limit to 500 and the minimum token limit to 250 to ensure a balanced summary of key sections with some detailed
C. Set the maximum token limit to 300 and the minimum token limit to 0, allowing the model to generate a brief summary of the most relevant points while avoiding excessive verbosity.
D. Set the maximum token limit to 700 and the minimum token limit to 600 to ensure a well-rounded summary with all necessary sections and detailed information included.

Answer: C

Question: 324

You have just finished a prompt tuning experiment for a large language model (LLM) to optimize its output for generating customer support summaries. The tuning results show that while the accuracy of the generated summaries is high (95%), the response time for generating them has significantly increased. The experiment data suggests that increasing the maximum token length during tuning led to better quality summaries but with slower generation. Which parameter should you adjust to improve the model's response time without sacrificing too much summary quality?

A. Decrease the maximum token length
B. Change the decoding strategy from beam search to greedy decoding
C. Increase the batch size
D. Decrease the number of epochs

Answer: B

Question: 325

In the context of IBM Watsonx Generative AI models, hallucinations refer to outputs where the model generates text that is factually incorrect or not grounded in the provided input or training data. Understanding the underlying causes of hallucinations is critical for maintaining the reliability of the model. Which of the following best describes a primary cause of hallucinations in generative models?

A. The model's training on incomplete or unstructured datasets leading to incorrect generalizations.
B. The model's use of a greedy decoding strategy without beam search.
C. The model's incapacity to follow the temperature parameter settings.
D. The model's over-reliance on token repetition to form coherent sentences.

Answer: A

Question: 326

You are tasked with reconstructing a prompt used in an AI-based customer support chatbot. The current prompt generates lengthy, detailed answers that are often overly verbose and unnecessary for the customer's inquiries. Your objective is to optimize this prompt to reduce model usage costs without compromising the quality of the responses. Which of the following strategies is the most effective in reducing the cost of using a Generative AI model while maintaining response relevance and clarity?

A. Splitting the prompt into multiple sub-prompts to generate responses for different sections separately.
B. Using temperature scaling to increase randomness and reduce token usage.
C. Revising the prompt to make it more specific by narrowing the scope of expected responses.
D. Reducing the token limit in the model configuration to restrict the length of responses.

Answer: C

Question: 327

You are tasked with building a customer service chatbot powered by a generative AI model for a large financial institution. Customers often provide sensitive personal and financial information in their queries. What is the best method to ensure the model does not generate responses containing sensitive personal data?

A. Limit the model’s ability to generate personalized responses by focusing on generalized outputs, reducing the risk of data exposure.
B. Restrict the model to only respond to predefined template-based answers to eliminate the risk of personal data being generated.
C. Use supervised fine-tuning to train the model on data that excludes all personally identifiable information (PII).
D. Implement differential privacy to reduce the likelihood that sensitive data from user inputs is exposed in the model’s responses.

Answer: D

Question: 328

You are tasked with deploying two generative AI systems: one for document summarization and another for question-and-answer (Q&A). Both systems need to handle large volumes of unstructured text, but they differ in their response time requirements and interaction with users. Which deployment strategy would be most appropriate for this scenario?

A. Use a single deployment strategy where both the summarization and Q&A systems are integrated into a shared model that processes all requests through the same inference pipeline.
B. Deploy the summarization model as a batch processing system, while the Q&A system is deployed as an on-demand service using a low-latency inference model.

C. Use a rule-based model for summarization and a machine learning-based model for Q&A, combining them into a single service endpoint for deployment.
D. Deploy both systems as a single pipeline, with the summarization model providing input to the Q&A model to improve the quality of answers.

Answer: B

Question: 329

In the context of Generative AI models, which scenario best illustrates the use of greedy decoding for determining the output sequence?

A. Randomly selecting tokens from the top-k highest probability tokens at each step
B. Averaging the probabilities of multiple tokens at each step to generate a sequence
C. Selecting the token that maximizes the cumulative probability of the sequence, considering future tokens
D. Selecting the token with the highest probability at each step, without considering future possibilities

Answer: D

Question: 330

In a project where watsonx.ai is deployed as the core generative model for text generation, you need to augment its capabilities by integrating it with IBM Watson Discovery to access a large corpus of documents for fact-checking and data retrieval. What is the best way to integrate these two services to ensure that generated responses are both creative and factual?

A. Have Watson Discovery preprocess user inputs to identify relevant documents, and then feed the retrieved information into watsonx.ai as a context for text generation.
B. Train watsonx.ai to automatically search Watson Discovery for factual information during the text generation process, eliminating the need for an external integration layer.
C. Deploy a custom API layer that combines watsonx.ai with Watson Assistant, letting the latter manage queries and retrieve documents from Watson Discovery on demand.
D. Use watsonx.ai to generate all responses, and periodically train the model using data indexed by Watson Discovery.

Answer: A

Question: 331

Your team is building a natural language processing pipeline using IBM watsonx components, where data from multiple external APIs and user inputs needs to be transformed, analyzed, and routed through various AI models. The process should involve the dynamic selection of models based on input data characteristics. The goal is to minimize latency while maintaining accuracy across tasks like sentiment analysis, text summarization, and query generation. Which IBM watsonx service would you use to implement a flexible, model orchestration pipeline that meets these requirements, and why?

A. IBM watsonx Data Refinery, as it can preprocess and analyze incoming data, and use its rules-

based engine to route it to different models.
B. IBM watsonx Model Management to dynamically select and orchestrate the models for different tasks based on real-time data analysis.
C. IBM watsonx Orchestrator, which allows for the integration and management of multiple AI models and can dynamically route inputs to the appropriate model based on predefined criteria.
D. IBM watsonx API Gateway to handle external data inputs, route them to different models, and ensure that each input is preprocessed in a low-latency manner.

Answer: C

Question: 332

You are tasked with optimizing a Generative AI model by reducing the cost per inference without sacrificing the quality of output. Which of the following prompt modifications is the most effective approach to achieve this goal?

A. Use more descriptive language in the prompt to improve output quality.
B. Simplify the prompt by reducing the number of tokens used.
C. Repeat the main instruction several times to ensure the model understands it.
D. Add additional context to the prompt to guide the model more effectively.

Answer: B

Question: 333

You are designing a generative AI system using the Retrieval-Augmented Generation (RAG) pattern. Your goal is to improve the accuracy of the generated content by incorporating relevant external knowledge in real-time. Which deployment strategy would best support this system architecture?

A. Deploy a single generative model and rely on post-processing to add any external information after the initial generation is complete.
B. Deploy the model with pre-configured static datasets for retrieval, updating the dataset manually on a regular basis.
C. Deploy the RAG system within a single containerized model that combines retrieval and generation functionalities in one inference step.
D. Use a multi-deployment model architecture where a retriever model and a generator model are deployed independently and communicate via API calls.

Answer: D

Question: 334

You are using a generative AI model in a healthcare application to generate personalized treatment recommendations based on patient data. Which of the following scenarios represent valid concerns related to model risks when deploying the AI in this setting? (Select two)

A. The model fails to generate treatment recommendations for some patients due to exceeding token limits during inference.

B. The model includes probabilistic estimates for treatment outcomes, which adds uncertainty to the recommendations and reduces their usability.
C. The model produces highly creative treatment recommendations that are not based on standard medical guidelines.
D. The model generates text that includes private patient information, violating data privacy regulations.
E. The model generates biased recommendations based on incomplete or skewed training data, which disproportionately impacts certain patient demographics.

Answer: D,E

Question: 335

You are optimizing a generative AI chatbot for concise responses to user queries, ensuring that it doesn't over-generate unnecessary content. However, you observe that the model occasionally stops prematurely, cutting off relevant information. What configuration best addresses this issue without allowing for excessive output?

A. Stop Sequence = ‘
B.
C. ’, Max Tokens = 150
D. Stop Sequence = ‘
E. ’, Max Tokens = 500
F. Stop Sequence = ‘. ’ (period followed by a space), Max Tokens = 300
G. Stop Sequence = ‘</end>’, Max Tokens = 250

Answer: G

Question: 336

Which of the following practices are best suited to optimize the performance of a deployed generative AI model in IBM watsonx under real-world traffic conditions? (Select two)

A. Relying on a single model configuration across all hardware types to simplify deployment
B. Using batch processing instead of real-time inference for all requests
C. Employing model quantization techniques during deployment
D. Loading the entire model into memory at runtime to avoid latency issues
E. Monitoring and adjusting resource allocation dynamically based on usage statistics

Answer: C,E

Question: 337

You are using IBM Watsonx to generate answers for a customer service chatbot. However, the responses generated sometimes include fabricated details that are inaccurate (hallucinations). Which of the following strategies will best mitigate the risk of hallucination when designing your prompt?

A. Rely on the AI model to fact-check and correct hallucinations in its responses automatically: “What

can you tell me about our product?"
B. Supply explicit product details and constraints in the prompt: "List the latest features of the XYZ smartphone released in 2024, including its AI-powered camera and battery-saving mode."
C. Use subjective language in your prompts: "Can you describe why our product is the best on the market?"
D. Ask open-ended questions without providing reference information: "What are the latest features of our product?"

Answer: B

Question: 338

In developing an LLM-based conversational AI application using LangChain, you want the AI to perform complex tasks, such as answering questions based on dynamic knowledge from multiple sources (e.g., databases, APIs, etc.). Which approach using LangChain best supports this requirement by combining various tools into a structured workflow for the AI to follow?

A. Build a chain of multiple LangChain agents, each handling a specific task (e.g., querying an API, accessing a database) to ensure data from various sources is used effectively.
B. Use a single LangChain agent to directly query all external data sources, allowing it to gather information on demand.
C. Create a LangChain chain that connects different tools (e.g., API access, database queries) in a sequential or branching manner to process and combine data dynamically.
D. Integrate LangChain memory with an agent to handle all external data retrieval without needing to build complex chains.

Answer: C

Question: 339

You are designing a prompt template for IBM Watsonx to generate various types of content for a company's internal communications. The goal is for the template to be reusable across different formats (emails, announcements, reports). Which of the following best describes an effective template for this scenario?

A. "Generate a formal communication piece for the company."
B. "Write the content for the communication based on the provided information."
C. "Create a [content type] for [audience], including key points on [subject]. Use a [tone] style and keep the length around [length]."
D. "Produce a report with all necessary information."

Answer: C

Question: 340

You are working on a generative AI model for a wide variety of language generation tasks. Your team is debating whether to use soft prompts for optimizing the model's performance. Which of the following is an accurate benefit of using soft prompts, and what is a potential drawback?

A. Benefit: Soft prompts reduce the amount of explicit data needed for training. Drawback: Soft prompts might not generalize well across domains.
B. Benefit: Soft prompts offer more control over the output. Drawback: Soft prompts require additional manual engineering for each task.
C. Benefit: Soft prompts allow for fine-grained control during inference. Drawback: Soft prompts may lead to lower performance with large datasets.
D. Benefit: Soft prompts can optimize performance without retraining the model. Drawback: They are computationally expensive to tune during inference.

Answer: A

Question: 341

You are implementing a Retrieval-Augmented Generation (RAG) system using LangChain, IBM WatsonX, and a vector database. The system needs to answer complex technical questions by retrieving relevant technical documents and generating a coherent response. Which of the following best describes LangChain's role in the RAG pattern implementation?

A. LangChain optimizes document retrieval by pre-generating responses based on common queries and caching them in memory for future use.
B. LangChain serves as an intermediary, facilitating communication between the vector database and WatsonX's LLM, ensuring that retrieved documents are contextually relevant for the LLM's response generation.
C. LangChain is responsible for storing and retrieving documents using a sparse keyword-based retriever from a traditional relational database.
D. LangChain allows WatsonX's LLM to fine-tune its parameters based on user interactions to improve future retrieval and generation accuracy.

Answer: B

Question: 342

You are tasked with optimizing a prompt-tuned large language model (LLM) using IBM Watsonx for a customer service chatbot. The chatbot needs to handle a variety of tasks, such as answering frequently asked questions (FAQs), providing detailed product descriptions, and troubleshooting user issues. What is the most appropriate task to focus on during the initial tuning experiment?

A. Tune the model for text summarization, condensing user queries into shorter forms.
B. Fine-tune the model to generate product descriptions using longer contextual prompts.
C. Optimize the model for extractive question-answering from a predefined knowledge base.
D. Focus on prompt-tuning the model for multi-turn dialogue to simulate more natural conversations.

Answer: C

Question: 343

In the context of Generative AI (GenAI), various embedding models are used to represent textual data. Which of the following best describes the difference between Word2Vec, BERT, and Sentence-BERT

embedding models?

A. Word2Vec captures both word and sentence meanings in a single vector space, BERT generates only word embeddings, and Sentence-BERT generates embeddings for entire documents.
B. Word2Vec captures contextual relationships between words, while BERT and Sentence-BERT generate sentence-level embeddings based on the overall document length.
C. Word2Vec creates static word embeddings, BERT generates dynamic embeddings based on context, and Sentence-BERT produces embeddings specifically optimized for sentence-level tasks like semantic similarity.
D. Word2Vec uses a transformer architecture for embedding generation, whereas BERT and Sentence-BERT use neural networks to model context.

Answer: C

Question: 344

You are building a generative AI conversational model to act as a virtual assistant that helps users schedule meetings. The goal is to create a prompt that initiates a conversation in a way that guides the user to provide relevant scheduling information. Which prompt would be the most effective for this use case?

A. "Please provide the user's calendar availability."
B. "Provide a list of all upcoming events in the user's calendar and ask if they want to add another."
C. "Guide the user through scheduling a meeting by asking about their preferred date, time, and attendees."
D. "Ask the user when they want to schedule their meeting."

Answer: C

Question: 345

You are fine-tuning a generative AI model in IBM Watsonx and need to define appropriate stopping criteria to ensure the generated text is relevant and coherent. Which of the following would be an example of a valid stopping criterion for a text generation task?

A. The model will stop generating text once the average probability of each word in the output falls below a 50% threshold.
B. The model will stop generating text when it encounters a special end-of-sequence token or punctuation such as a period or question mark.
C. The model will stop generating text once a predefined number of characters is reached, ensuring that the output does not exceed a set length.
D. The model will stop generating text once it detects a repetitive sequence of words or phrases, automatically cutting off redundancy.

Answer: B

Question: 346

You are tasked with deploying a suite of AI assets, including a pretrained generative language model and

a set of custom-trained models. Your company operates in an environment where scalability and adaptability are key, as customer needs vary significantly across regions and sectors. You need to ensure that the deployment of these AI assets can meet varying demand, maintain performance, and allow for customization based on specific client needs. Which deployment strategy best balances scalability, performance, and adaptability for this suite of AI assets?

A. Deploy all AI assets as a single monolithic model on a multi-GPU setup, allowing scalability across multiple customers without the need for further customization.
B. Fine-tune the pretrained model for each client individually, deploy as separate models for each use case, and store them locally for each region.
C. Deploy the AI assets on edge devices near the customer, reducing latency and ensuring consistent performance without relying on cloud infrastructure.
D. Containerize each AI asset separately, deploy them as microservices on a cloud platform, and use load balancing to manage varying demand across regions.

Answer: D

Question: 347

You are tasked with fine-tuning a pre-trained language model for a customer support chatbot. The dataset you're using is mostly unstructured text from chat logs. What steps should you take to prepare the dataset for fine-tuning to ensure optimal model performance?

A. Normalize the text by removing all punctuation, special characters, and converting text to lowercase.
B. Randomly split the data into training, validation, and test sets.
C. Use data augmentation techniques like paraphrasing to artificially increase the dataset size.
D. Use domain-specific tokenization to better capture important keywords and phrases relevant to customer support.

Answer: D

Question: 348

You are designing prompt templates for an automated content generation tool that your team uses for marketing copy. The tool must balance cost efficiency with the generation of high-quality outputs. The prompts should be structured to avoid excessive token usage while ensuring that the model provides coherent and actionable content. Which of the following techniques would best optimize the design of cost-effective prompt templates?

A. Using placeholders in the prompt templates for dynamic insertion of variables specific to each task. B. Embedding explicit stop words within the prompt to control the length of generated outputs.
C. Creating templates that provide extensive background information to give the model context.
D. Designing prompts that incorporate redundancy to ensure model outputs address all possible user needs.

Answer: A

Question: 349

You are tasked with fine-tuning a generative AI model for text data using synthetic data created through the IBM watsonx platform's user interface. The data you are working with is skewed, containing mostly outliers, and you need to ensure that the synthetic data mimics the distribution accurately. Which algorithm would be most appropriate for generating synthetic data that mirrors the original distribution, considering the Anderson-Darling test for normality?

A. Bootstrapping
B. Generative Adversarial Networks (GANs)
C. Decision Trees
D. Anderson-Darling Based Synthetic Data Generation (ADS-DG)

Answer: D

Question: 350

You are tasked with integrating a generative AI model on watsonx.ai into a custom business workflow. The workflow requires complex prompt chains and interaction with external APIs. Which of the following best describes how you should approach the integration using watsonx.ai and LangChain?

A. Implement LangChain to handle complex multi-step workflows, using watsonx.ai's APIs to generate responses at specific stages in the chain.
B. Directly integrate the external APIs with watsonx.ai without any intermediate framework, since LangChain would add unnecessary overhead.
C. Use only watsonx.ai's built-in APIs and SDKs for integration, as LangChain is not required for chaining multiple prompts.
D. Write custom scripts to manage all prompt sequences manually, leveraging watsonx.ai's SDK to call the generative AI model at every step.

Answer: A

Question: 351

You are tasked with developing a RAG system that integrates a transformer-based language model with a large document corpus. To speed up the development process, you are considering using specialized libraries designed for RAG. Which of the following reasons best explains why these libraries are essential for your development process?

A. They provide pre-trained retrieval models and generators, eliminating the need for any fine-tuning or customization of the system.
B. They allow for the retrieval of documents based purely on keyword search, optimizing for exact match over semantic similarity.
C. They provide a graphical interface for users to build RAG systems without requiring any programming knowledge.
D. They streamline the integration of retrievers and generators, providing out-of-the-box support for embedding models and vector databases.

Answer: D

Question: 352

You're designing a prompt that should generate text for product descriptions using a generative AI model. You want to limit the model's word choices to ensure coherence while still allowing some flexibility for creativity. You decide to use Top-P (nucleus) sampling to achieve this. Which of the following settings for the Top-P parameter is most appropriate to strike a balance between creativity and coherence?

A. Top-P = 1.0
B. Top-P = 0.1
C. Top-P = 0.9
D. Top-P = 0.3

Answer: C

Question: 353

You are selecting a model to fine-tune using Tuning Studio for a financial application that requires high accuracy and domain-specific language understanding. Which type of model should you select to maximize fine-tuning efficiency and performance?

A. A pre-trained model that has been optimized for creative text generation.
B. A small pre-trained model specifically designed for open-domain tasks.
C. A large pre-trained language model that has been fine-tuned on generic business communication.
D. A model pre-trained on financial and business data but with limited language capabilities.

Answer: D

Question: 354

A machine learning engineer is optimizing a generative AI model for creative writing. They are debating the use of soft prompts over hard prompts. What is the primary advantage of using soft prompts in this context, despite their complexity?

A. Soft prompts improve the simplicity of the model's overall structure, making it easier to debug and interpret the generation process.
B. Soft prompts enforce stricter generation outputs due to the deterministic nature of the learned embeddings, leading to higher consistency in results.
C. Soft prompts allow for more nuanced control of the model's behavior through learned embeddings, which can adapt to a variety of tasks without requiring explicit human intervention.
D. Soft prompts provide more direct control over the model's behavior by offering explicit, human-readable instructions.

Answer: C

Question: 355

You have been assigned the task of fine-tuning a large language model (LLM) for a chatbot that will assist users with technical troubleshooting. The goal is to ensure the chatbot responds accurately to user queries, but also in a specific tone and format. Which of the following steps is the first critical phase in the InstructLab workflow to ensure successful customization of the model?

A. Running evaluation metrics on the baseline model to measure initial performance.
B. Defining task-specific instructions and fine-tuning them through prompt design.
C. Deploying the model in a real-time environment for user feedback collection.
D. Pre-processing and augmenting the training data to improve the model's generalization capabilities.

Answer: B

Question: 356

What is a key advantage of using prompt variables in IBM Watsonx for a chatbot application that needs to handle multiple user intents?

A. Using prompt variables ensures that the model will always select the most relevant response based on past conversations.
B. Prompt variables allow for the dynamic injection of user-specific details into the response, improving personalization.
C. Prompt variables enable developers to predefine multiple static responses for each possible user input.
D. Prompt variables allow the model to automatically detect the user's intent, ensuring accurate responses.

Answer: B

Question: 357

You are deploying an AI model to a production environment using watsonx.ai. The model is a fine-tuned GPT-based generative model designed to support real-time customer interactions. The deployment must ensure scalability, security, and maintain high availability. Which deployment strategy should you choose to best meet these requirements?

A. Deploy the model on multiple virtual machines, manually balancing traffic between instances.
B. Use a containerized deployment with Kubernetes on a multi-node GPU cluster to support auto-scaling and fault tolerance.
C. Deploy the model on an on-premises server with an isolated environment to ensure full control and security.
D. Deploy the model on a single cloud-based GPU instance, and enable scheduled maintenance to minimize downtime.

Answer: B

Question: 358

When selecting parameters to optimize a prompt-tuned model experiment in IBM watsonx, which parameter is the most critical for controlling the model's ability to generate coherent and contextually accurate responses?

A. Gradient clipping
B. Batch size
C. Learning rate
D. Dropout rate

Answer: C

Question: 359

You are using InstructLab to fine-tune a large language model (LLM) for generating technical documentation. The model's output is inconsistent, sometimes too verbose and other times lacking critical details. Which of the following actions within InstructLab will best help customize the model to consistently produce balanced, concise, yet informative outputs?

A. Use prompt engineering to provide more explicit instructions for the model
B. Adjust the token length limit in the model's configuration
C. Increase the number of fine-tuning epochs to ensure the model converges
D. Lower the batch size during fine-tuning to force the model to focus on smaller chunks of data

Answer: A

Question: 360

You are tasked with fine-tuning a large language model (LLM) on a specific industry dataset using the IBM watsonx user interface. Due to the lack of labeled data, your team decides to generate synthetic data to supplement the training set. The objective is to ensure the fine-tuned model can generalize effectively to real-world scenarios in this industry. You need to configure synthetic data generation and perform the fine-tuning. Which of the following actions should you take to ensure the synthetic data is suitable for fine-tuning the model and does not lead to overfitting or model bias? (Select two)

A. Configure the synthetic data to exclude rare or uncommon events, as these are not representative of the overall dataset.
B. Generate only a small amount of synthetic data to minimize computational costs and training time.
C. Use the IBM watsonx Data Refinery tool to inspect and balance the synthetic data before finetuning.
D. Rely solely on synthetic data for fine-tuning, as it is fully representative of the industry domain. E. Ensure that the synthetic data covers edge cases as well as common industry scenarios.

Answer: C,E

Question: 361

After conducting a prompt tuning experiment in IBM Watsonx, which two statistical metrics are most

indicative of a model's ability to generalize well to unseen data? (Select two)

A. Small difference between training and validation loss
B. High training accuracy
C. Low validation loss
D. Low training loss
E. Large difference between training and validation accuracy

Answer: A,C

Question: 362

When deploying a machine learning model in a highly regulated industry (e.g., healthcare or finance), which strategy is most effective to ensure ongoing model performance while adhering to AI governance standards?

A. Implement a model performance monitoring framework with fairness and bias detection metrics
B. Perform real-time continuous training of the model using live data from the production environment
C. Deploy the model with hard-coded rules to ensure it does not drift from expected behavior
D. Ensure model interpretability is maximized by simplifying the architecture to a linear model

Answer: A

Question: 363

You are tasked with designing a prompt for a sentiment analysis model based on a large language model (LLM). The goal is to generate a coherent response from the model that aligns with a particular sentiment (positive, negative, or neutral) for customer reviews of a product. Which of the following prompt designs are best suited to generate a positive review response? (Select two)

A. "Write a review about the product that highlights both its pros and cons."
B. "Describe the product as if you were a very satisfied customer, and you were recommending it to a friend."
C. "Generate a positive review about the product, focusing on the key strengths and avoiding any negative aspects."
D. "Analyze the product based on the customer feedback and write a review that covers all sentiments."
E. "Write a neutral review, neither praising nor criticizing the product."

Answer: B,C

Question: 364

A financial services company is building a generative AI model to assist with customer support. The company is concerned about potential legal liabilities if the model generates customer information, such as bank account numbers or personal identification data, as part of its responses. Which of the following techniques would best mitigate the risk of generating Personally Identifiable Information (PII) during

inference?

A. Use greedy decoding to ensure the model generates only the most probable tokens, which are less likely to include PII.
B. Set a strict token limit to prevent the model from generating long sequences, assuming PII tends to appear in longer outputs.
C. Implement a real-time PII filter that detects and removes sensitive data before the output is presented to the user.
D. Train the model on sensitive customer data but ensure that the temperature is set low to avoid generating diverse outputs.

Answer: C

Question: 365

Which of the following best describes the effect of controlling model parameters during the decoding process in IBM Watsonx's generative AI models?

A. Setting model parameters guarantees that the model will generate the most grammatically correct response, regardless of context.
B. Model parameters control the model's ability to self-learn from previous outputs, improving its performance over time.
C. Adjusting model parameters helps control the randomness in the output generation process, enabling a balance between creativity and accuracy.
D. Controlling model parameters allows the model to generate only the shortest possible responses, ensuring concise outputs.

Answer: C

Question: 366

You are tasked with implementing a Retrieval-Augmented Generation (RAG) pipeline for a customer service chatbot. Your goal is to ensure that the model can query a vast knowledge base, retrieve relevant documents, and generate responses that reference these documents. You are using the transformers and faiss libraries in Python to achieve this. After retrieving the relevant passages, the generative model will then process them to formulate a final response. What is the most appropriate sequence of steps for implementing this RAG pipeline?

A. Encode the query using FAISS -> Use the transformer model to retrieve documents from the knowledge base -> Generate a response using the encoded documents.
B. Use the FAISS library to create an index -> Encode the input query using the transformer model -> Retrieve the most relevant documents -> Pass the retrieved documents to the generative model for response generation.
C. Initialize the generative model -> Use FAISS to encode the retrieved documents -> Pass the encoded documents to the transformer model for response generation.
D. Generate the response first using the transformer model -> Retrieve documents from the knowledge base -> Encode the documents using FAISS -> Use the retrieved documents to fine-tune the generated response.

Answer: B

Question: 367

You are tasked with fine-tuning prompts for a customer support chatbot built using IBM Watsonx. You decide to leverage Prompt Lab to improve the model's responses. Which of the following best describes the key benefits of using Prompt Lab for this task?

A. Prompt Lab provides an environment to experiment with different prompt structures and analyze their impact on model outputs in real-time, helping optimize responses.
B. Using Prompt Lab guarantees that the model will never produce biased responses, regardless of the input data used.
C. Prompt Lab offers pre-trained prompts specific to industry verticals, making it unnecessary to create customized prompts.
D. Prompt Lab enables you to train the model with new data, ensuring continuous learning and improved performance over time.

Answer: A

Question: 368

You are developing a tuned language model for a healthcare chatbot that provides concise responses to patient inquiries. Using Tuning Studio, you want to ensure the model is well-optimized for generating responses specific to medical terminology while maintaining efficiency. Which of the following represents the correct workflow to create a tuned model using Tuning Studio?

A. Input the dataset, manually adjust the learning rate and batch size, and export the fine-tuned model without evaluation.
B. Select a model, upload the dataset, and let Tuning Studio automatically generate synthetic data to improve model training.
C. Select a pre-trained model, upload the custom medical dataset, fine-tune the hyperparameters, and evaluate the model's performance.
D. Load the model, automatically adjust its architecture, and deploy it to production.

Answer: C

Question: 369

In IBM Watsonx Generative AI, controlling model parameters is crucial for managing the output generation process. Which of the following statements correctly describes how adjusting the temperature parameter influences the model's response?

A. Decreasing the temperature parameter produces more creative and diverse responses.
B. The temperature parameter has no impact on the diversity or creativity of the model's output but controls response length.
C. Increasing the temperature parameter results in more deterministic and repetitive responses.
D. A higher temperature parameter introduces more randomness, leading to less predictable

responses.

Answer: D

Question: 370

You are tasked with deploying a foundation model for text generation on the IBM Watsonx platform. The foundation model has been pre-trained on a large corpus but has not been fine-tuned for your specific use case. What is the most critical factor to consider when deploying this model to ensure it performs optimally on the Watsonx platform?

A. You must ensure that the model is compatible with IBM's foundation model API, as Watsonx only allows deployment of models that integrate with its foundation model architecture.
B. The model's size should be reduced to fit within the memory constraints of the deployment environment, even if it leads to a loss of precision in its outputs.
C. The model should be fine-tuned with domain-specific data before deployment, as Watsonx requires fine-tuning for all foundation models before they can be deployed.
D. You should enable quantization of the model to optimize inference performance without compromising accuracy in a production environment.

Answer: D

Question: 371

A team is implementing a Retrieval-Augmented Generation (RAG) system for document search and retrieval. Their goal is to enable users to retrieve contextually relevant documents from a large, unstructured text corpus. They are considering using a vector database to handle this task. In which scenario is a vector database the most appropriate choice for storing and retrieving documents?

A. When the system must support real-time updates and frequent data modifications, ensuring that query results always reflect the latest state of the data.
B. When documents need to be retrieved based on semantic similarity to the user’s query, even if the exact terms are not matched.
C. When exact match search is required, such as finding documents containing specific keywords or phrases.
D. When all documents are structured and can be queried using traditional SQL queries, focusing on specific fields and categories.

Answer: B

Question: 372

You are working with IBM Watsonx to develop a generative AI solution that automatically generates product descriptions for an e-commerce website. The descriptions need to be concise, factual, and include important product features like size, color, and material. Which prompt design approach would best ensure the output meets these requirements?

A. "Generate a creative and imaginative product description for the items listed below."

B. "Generate a product description that highlights the unique aspects of the product and uses emotional language to engage the reader."
C. "Write a summary that provides information on each product, making the content engaging, humorous, and memorable."
D. "Provide a product description for the following items, ensuring it is factual, concise, and includes specific details such as size, color, and material."

Answer: D

Question: 373

Which of the following statements best describes the primary advantage of applying quantization to a large language model (LLM) during inference?

A. Quantization lowers computational requirements, enabling faster inference with minimal impact on model accuracy.
B. Quantization allows the model to learn more efficiently during training by focusing on fewer parameters.
C. Quantization automatically improves the accuracy of an LLM by converting all weights to higher precision.
D. Quantization primarily reduces the size of the training dataset required for an LLM.

Answer: A

Question: 374

You are enhancing an existing relational database system to support vector-based similarity search, integrating RAG into your infrastructure. Which of the following technologies or approaches represents a valid method for extending a traditional SQL database to handle vector embeddings and similarity searches?

A. Normalizing vector embeddings into relational tables with foreign key constraints
B. Using graph database extensions to enable vector embeddings within a SQL database C. Using a full-text indexing engine, such as Elasticsearch, to store the vector embeddings D. Adding a plugin that provides support for k-nearest neighbors (k-NN) search

Answer: D

Question: 375

You are designing a generative AI model to generate detailed technical

A. Stop Sequence = None, Max Tokens = 700
B. Stop Sequence = '<END>', Max Tokens = 100
C. Stop Sequence = 'END', Max Tokens = 500

D. Max Tokens = 500: This provides enough space for a detailed technical

Question: 376

In a RAG system, the retriever is responsible for fetching relevant documents or information from a knowledge base based on the input query. Different retriever types can be used depending on the nature of the task. Which retriever type is most suitable for a RAG system that requires efficient large-scale retrieval from a document corpus based on semantic similarity?

A. Lexical Retriever: Primarily returns results based on syntactic similarity, relying on word order and surface-level features of the input query.
B. Hybrid Retriever: Combines syntactic retrieval methods (like BM25) with semantic retrieval (like dense retrieval) but sacrifices retrieval speed for accuracy.
C. Exact-Match Retriever: Returns documents based solely on keyword matching and is optimized for highly structured, labeled datasets.
D. Dense Retriever: Uses vector embeddings to retrieve documents based on the semantic similarity of the input query and stored documents.

Answer: D

Question: 377

In which scenario would using a soft prompt be more beneficial than a hard prompt in optimizing generative AI outputs?

A. When the prompt needs to be manually adjusted by the user in real time during interaction with the AI.
B. When the task requires explicit and consistent user instructions to ensure deterministic outcomes.
C. When the model needs to generate a strictly factual output with minimal deviation from the prompt.
D. When fine-tuning a pre-trained model for domain-specific tasks, allowing the system to adapt its understanding through learned embeddings.

Answer: D

Question: 378

After prompt-tuning a generative AI model, you review its performance on multiple evaluation metrics. The metrics include accuracy, perplexity, and latency. Which combination of these metrics would most effectively allow you to optimize the model for both user experience and content quality?

A. High accuracy, low perplexity, high latency
B. Low accuracy, high perplexity, low latency
C. High accuracy, low perplexity, low latency
D. High accuracy, high perplexity, high latency

Answer: C