Exam Code: AB-731

Exam Name: Microsoft Certified AI Transformation Leader

Question 1

Which cost model is commonly used by Azure AI for generative AI solutions?

Correct answer

A. Pay-as-you-go with token-based billing.

Explanation

Azure AI commonly uses a pay-as-you-go model with token-based billing for generative AI solutions. This model allows users to pay for the they consume, such as the number of tokens used for training and inference, providing flexibility and cost-effectiveness for varying usage levels.

B. One monthly subscription that covers all models with unlimited usage.

Explanation

A monthly subscription that covers all models with unlimited usage is not a commonly used cost model by Azure AI for generative AI solutions. This type of subscription model may not align with the pay-as-you-go approach typically used for AI solutions, where costs are based on actual usage rather than a fixed monthly fee.

C. One-time purchase for lifetime access to a single model.

Explanation

A one-time purchase for lifetime access to a single model is not a commonly used cost model by Azure AI for generative AI solutions. Generative AI solutions often require ongoing

training, maintenance, and updates, making a one-time purchase model less suitable for these dynamic and evolving AI applications.

D. A fixed annual enterprise license with unlimited usage and no usage tracking.

Explanation

A fixed annual enterprise license with unlimited usage and no usage tracking is not a commonly used cost model by Azure AI for generative AI solutions. This model lacks the granularity and flexibility needed for AI solutions, as it does not track usage or provide insights into resource consumption for optimization and cost management.

Overall Explanation

Azure's generative AI services (such as Azure OpenAI Service) commonly use a pay-as-you-go model, where you pay based on usage. For language models, this is often measured in tokens processed (input + output). This makes costs scale with actual consumption.

Why the other options are incorrect

- Option B: There may be enterprise agreements, but the standard, widely used model is usage-based, not "one flat monthly fee with unlimited usage" across all models.
- Option C: Generative AI is delivered as a cloud service, not a one-time software purchase. Costs are ongoing and tied to consumption.
- Option D: Some enterprise agreements can simplify billing, but you still have usage tracking and limits; it is not simply "unlimited usage with no tracking."

Domain

AB-731 Fundamentals

Question 2

What is a key security practice for protecting sensitive information in AI systems?

A. Allowing open access to enterprise data for faster AI responses.

Explanation

Allowing open access to enterprise data for faster AI responses is not a key security practice for protecting sensitive information in AI systems. Open access can lead to unauthorized access and potential data breaches, compromising the security of sensitive information.

Correct answer

B. Encrypting data at rest and in transit.

Explanation

Encrypting data at rest and in transit is a key security practice for protecting sensitive information in AI systems. Encryption helps to secure data both when it is stored and when it is being transmitted, ensuring that only authorized parties can access and decipher the information.

C. Storing sensitive data in unstructured formats without access controls.

Explanation

Storing sensitive data in unstructured formats without access controls is not a key security practice for protecting sensitive information in AI systems. Unstructured data and lack of access controls can make it easier for unauthorized users to access and misuse sensitive information.

D. Sharing API keys and secrets in prompts for easier integration.

Explanation

Sharing API keys and secrets in prompts for easier integration is not a key security practice for protecting sensitive information in AI systems. Sharing sensitive information in this manner can lead to unauthorized access and potential security breaches. It is important to securely manage and protect API keys and secrets to maintain the security of AI systems.

Overall Explanation

A fundamental security practice is ensuring encryption at rest and in transit so that data is protected when stored and while moving across networks. In Azure, many services encrypt data at rest by default, and TLS is used for data in transit; organizations can also use customer-managed keys for stricter control.

Why the other options are incorrect

- Option A: “Open access” increases the risk of data leakage and misuse; least privilege and access controls are best practice.
- Option C: Sensitive data must be protected with access controls, classification, and encryption, not left unmanaged.
- Option D: API keys and secrets should never be placed in prompts or shared in logs; they must be stored securely (for example, in Azure Key Vault).

Domain

AB-731 Fundamentals

Question 3

How can grounding improve the reliability of AI solutions?

Correct answer

A. By connecting AI responses to verified data sources.

Explanation

Grounding in AI refers to connecting AI responses to verified data sources, which helps ensure that the AI solution is based on accurate and reliable information. By grounding AI responses in verified data sources, the reliability of the AI solution is improved as it reduces the risk of errors or biased outcomes.

B. By increasing the size of the original training dataset.

Explanation

Increasing the size of the original training dataset may improve the performance of an AI model in terms of generalization and accuracy, but it does not directly relate to improving the reliability of AI solutions through grounding. Grounding specifically focuses on connecting AI responses to verified data sources to enhance reliability.

C. By automating prompt creation.

Explanation

Automating prompt creation may streamline the AI development process and improve efficiency, but it does not directly contribute to improving the reliability of AI solutions through grounding. Grounding involves connecting AI responses to verified data sources to ensure accuracy and reliability in the outputs.

D. By disabling all human review of AI outputs.

Explanation

Disabling all human review of AI outputs can lead to potential errors, biases, or inaccuracies in the AI solutions. Grounding, on the other hand, aims to improve the reliability of AI solutions by connecting AI responses to verified data sources to ensure accuracy and reduce the risk of errors.

Overall Explanation

Grounding (often via Retrieval-Augmented Generation, RAG) involves providing the model with relevant, trusted data at query time—for example, enterprise documents or knowledge bases. This helps the model generate answers that are more factual and aligned to your current data, reducing hallucinations.

Why the other options are incorrect

- Option B: Grounding does not retrain the model or change its training dataset; it adds retrieved context to the prompt at runtime.
- Option C: Prompt automation can be useful, but it's not what "grounding" means.
- Option D: Grounding works best with human review in high-risk scenarios, not instead of it.

Domain

AB-731 Fundamentals

Question 4

A company is using AI to filter job applications. Which practice best aligns with the principle of inclusiveness?

A. Standardizing all job applications to exclude applicants with gaps in employment history.

Explanation

Standardizing all job applications to exclude applicants with gaps in employment history goes against the principle of inclusiveness. It may unintentionally discriminate against individuals who have valid reasons for employment gaps, such as caregiving responsibilities or health issues.

Correct answer

B. Including people with disabilities in system testing to validate usability for the widest audience.

Explanation

Including people with disabilities in system testing to validate usability for the widest audience aligns with the principle of inclusiveness. By involving individuals with diverse abilities in testing, the AI system can be optimized to be accessible and usable for a broader range of applicants.

C. Using only the top 10% of resumes for training the AI model.

Explanation

Using only the top 10% of resumes for training the AI model may lead to biased outcomes and exclude qualified candidates who do not fit within that narrow selection criteria. This practice does not align with the principle of inclusiveness, as it may perpetuate existing biases in the hiring process.

D. Restricting the AI to process applications only from regions with historically high hiring rates.

Explanation

Restricting the AI to process applications only from regions with historically high hiring rates can perpetuate regional biases and limit opportunities for candidates from underrepresented or marginalized communities. This practice does not align with the principle of inclusiveness, as it may exclude qualified individuals based on their geographical location.

Overall Explanation

Inclusiveness means designing AI systems so they are accessible, usable, and beneficial to a broad and diverse set of people. Including people with disabilities and other underrepresented groups in testing helps ensure the system works for them and does not inadvertently exclude them.

Why the other options are incorrect

- Option A: Excluding applicants with gaps can disproportionately affect certain groups and reduce inclusiveness.

- Option C: Training only on “top” resumes can lock in historical bias and narrow what “good” looks like.
- Option D: Limiting regions can embed geographic or socioeconomic bias rather than broadening opportunity.

Domain

AB-731 Fundamentals

Question 5

Which metric best helps leaders understand whether AI is delivering business value rather than just technical performance?

Correct answer

A. Reduction in average handling time for a customer service process.

Explanation

Reduction in average handling time for a customer service process is a key metric that directly reflects the impact of AI on business operations and customer satisfaction. It indicates the effectiveness of AI in improving efficiency and delivering value to the business.

B. Number of GPUs used by AI workloads.

Explanation

The number of GPUs used by AI workloads is more of a technical performance metric rather than a business value metric. While it may be important for optimizing AI performance, it does not directly measure the impact of AI on delivering business value.

C. Total number of models deployed in production.

Explanation

The total number of models deployed in production is a metric related to the technical aspects of AI implementation rather than its business value. While deploying multiple models may be necessary for various tasks, it does not necessarily indicate the overall impact of AI on business outcomes.

D. Number of prompts sent to a generative AI model per day.

Explanation

The number of prompts sent to a generative AI model per day is a technical metric related to the usage and activity of the AI model. It does not directly measure the business value delivered by AI, as it focuses more on the operational aspects of the AI model rather than its impact on business outcomes.

Overall Explanation

Metrics like reduction in handling time, revenue uplift, conversion rate changes, or error reduction tie AI directly to business outcomes. These are the kinds of value-focused metrics leaders need to see to understand whether AI is improving the business, not just running more technology.

Why the other options are incorrect

- Option B: Infrastructure metrics (GPUs, cores) show scale, not business value.
- Option C: More models in production doesn't automatically mean more impact; some may add little value.
- Option D: Prompt volume is an adoption indicator, but doesn't prove outcome-level value on its own.