PassQueen
a.k.a dumpsqueen.com

CompTIA SecAI+ v1

CompTIA CY0-001

Version Demo

Total Demo Questions: 16

Total Premium Questions: 164

Buy Premium PDF

https://passqueen.com

support@passqueen.com

Topic Break Down

Topic	No. of Questions
Topic 1, AI Security Fundamentals	20
Topic 2, AI Threats and Vulnerabilities	32
Topic 3, AI Security Architecture	45
Topic 4, AI Security Operations	44
Topic 5, AI Governance, Risk, and Compliance	22
Topic 6, Mix Questions	1
Total	164

QUESTION NO: 1

An architect is using the firm ' s recommended large language model (LLM) to find an internal solution for content management.

Given the following:

Which of the following controls is the best for mitigating this issue?

A. Model training

B. Response validation

C. Access controls

D. Integrity monitoring

ANSWER: B

Explanation:

Response validation is the best control because it evaluates LLM-generated content before the organization relies on it to select or implement an internal content-management solution. An LLM can produce an answer that appears authoritative while containing unsupported claims, obsolete information, incorrect technical details, or fabricated references. Validation establishes a control point that checks whether the response is accurate, relevant to the organization's requirements, and supported by approved internal sources.

In practice, this can include grounding answers in curated enterprise repositories, requiring verifiable citations, checking recommendations against approved architecture standards and configuration data, applying confidence thresholds, and routing uncertain or high-impact responses to a qualified human reviewer. These measures help ensure that generated recommendations are treated as content requiring verification rather than as inherently trustworthy facts. The OWASP guidance for LLM applications identifies misinformation and overreliance as key risks and recommends mechanisms that improve the factuality and trustworthiness of outputs: OWASP Top 10 for LLM Applications. Similarly, the NIST AI Risk Management Framework emphasizes managing AI-system risks related to valid, reliable, and trustworthy outputs. Response validation directly operationalizes those principles for this internal decision-making use case.

QUESTION NO: 2

An internal user enters a client credit card number into an internal generative machine learning (ML) model:

#User prompt: Customer Jane Doe has a new credit card that she wants to add to her account. The number is 5555-5555-5555-5555

Which of the following is the most effective way to prevent prompt injection attacks against a large language model (LLM)?

A. Guardrails

B. Antivirus

C. Web application firewall (WAF)

D. Role-based access control

ANSWER: A

Explanation:

Guardrails are the most effective control because they are designed specifically to constrain how an LLM handles untrusted input and what it is permitted to return or do. In an LLM application, user-provided text must be treated as untrusted data rather than trusted instructions. Guardrails enforce that separation through policy-based input inspection, instruction filtering, prompt-template controls, content and sensitive-data detection, tool-use restrictions, and output validation. These controls can prevent a malicious or embedded instruction from overriding the application's intended system instructions or causing unauthorized actions and disclosures.

Effective guardrails are implemented as defense in depth rather than as a single prompt. For example, the application can identify payment-card data before it reaches the model, restrict the model to an approved task and data scope, limit access to connected tools and knowledge sources, and validate the response before it is shown to a user or used by another system. This approach directly addresses prompt injection, which occurs when untrusted content attempts to alter an LLM's behavior. OWASP identifies prompt injection as a major LLM application risk and recommends measures that constrain model behavior and validate inputs and outputs. See OWASP LLM01: Prompt Injection and Amazon Bedrock Guardrails documentation.

QUESTION NO: 3

An organization collects threat-intelligence feeds to train an AI system that identifies malicious domains. The security architect is concerned that untrusted or altered source data could affect model behavior.

Which of the following controls should the architect implement? (Choose two.)

A. Data provenance tracking

B. Data validation

C. Output token controls

D. Session expiration

E. Content delivery network (CDN)

ANSWER: A B

Explanation:

Data provenance tracking and **data validation** are appropriate controls because they help establish that training data originates from approved sources and meets defined quality and integrity requirements. Provenance tracking records the origin, ownership, collection method, and transformation history of each dataset. This supports investigation when suspicious or inaccurate records are discovered.

Data validation checks incoming records against expected schemas, formats, ranges, and business rules before they are accepted into the training pipeline. These controls can identify malformed, incomplete, unexpected, or unauthorized data that may indicate an error or poisoning attempt. NIST describes the importance of managing data quality and provenance throughout the AI lifecycle in the NIST AI Risk Management Framework.

QUESTION NO: 4

Developers introduce new features to their generative AI product in an effort to stand out from the competition and offer more value to customers.

Which of the following most accurately explains the risks when enabling more functionality?

A. The risks remain the same as before the new features were added.

B. The risks increase when new features are added.

C. The risks are measured qualitatively.

D. The risks are proportional to the model's capabilities.

ANSWER: D

Explanation:

The risks are proportional to the model's capabilities. Each new capability can expand the system's attack surface, its ability to affect systems or users, and the consequences if it is misused or behaves unexpectedly. A generative AI system limited to producing draft text has a substantially different risk profile from one that can access sensitive enterprise data, browse external content, invoke tools, execute code, make API calls, or initiate actions on a user's behalf. The appropriate level of security controls, human oversight, authorization, logging, testing, and continuous monitoring should therefore scale with the functionality and autonomy being enabled.

This is not simply a matter of counting features. Risk depends on the context and impact of each capability, including the data it can reach, the privileges granted to it, the decisions it can influence, and the reversibility of its actions. NIST's AI Risk Management Framework describes AI risk as context-dependent and calls for risk management across the AI lifecycle. OWASP also identifies excessive agency and insecure tool or plugin integration as important risks for large language model applications, particularly where models can take consequential actions. See the NIST AI Risk Management Framework and the OWASP guidance on excessive agency.

QUESTION NO: 5

A data science team suspects that an attacker has added mislabeled records to a training dataset used by a phishing-detection model. The team must identify and reduce the impact of poisoned data before retraining the model.

Which of the following actions should the team take? (Choose two.)

A. Reviewing data-source provenance

B. Performing statistical outlier analysis

C. Increasing output token limits

D. Enabling chat-history memory

E. Implementing a content delivery network (CDN)

ANSWER: A B

Explanation:

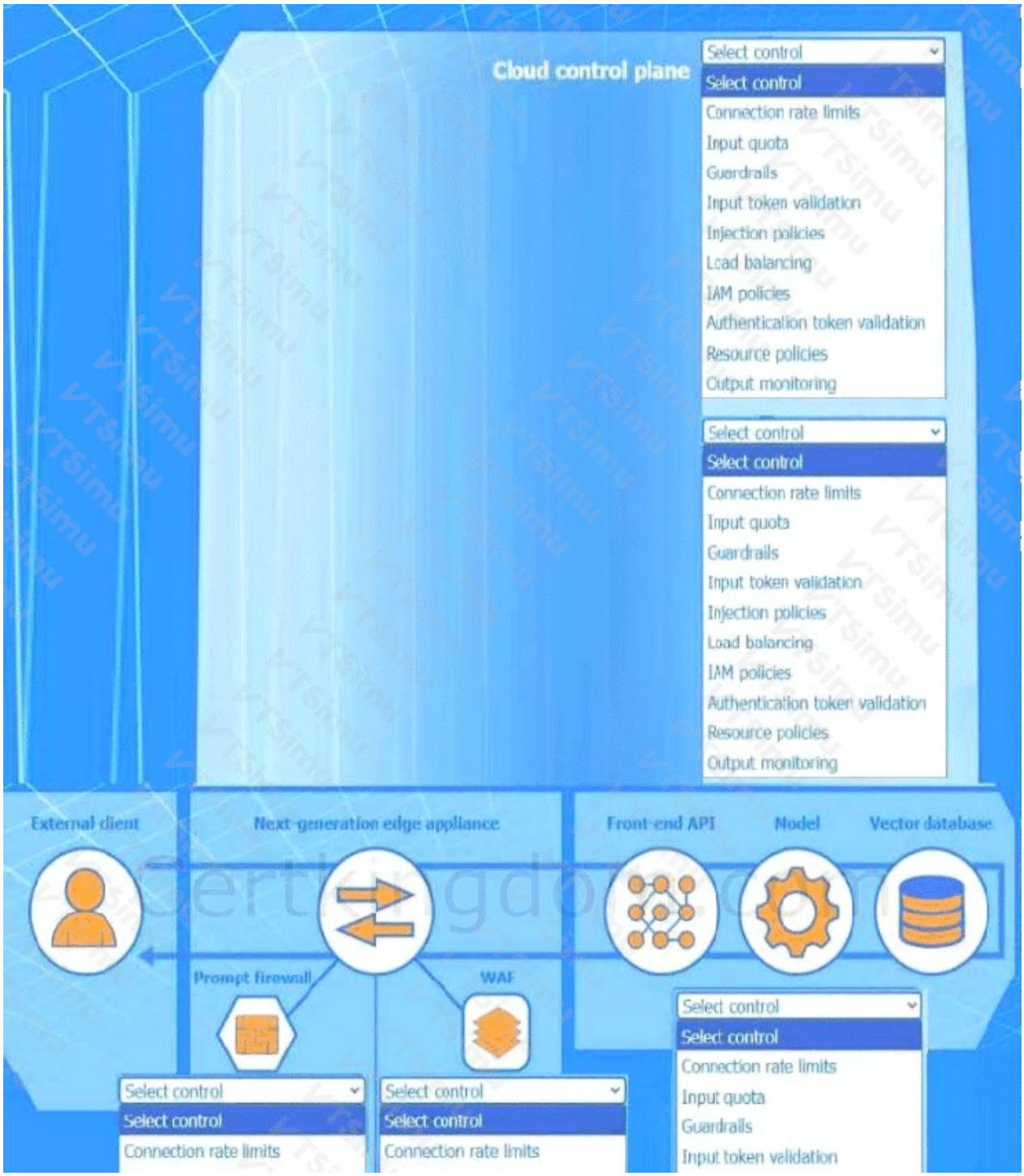
Reviewing data-source provenance and **performing statistical outlier analysis** are appropriate actions because they help identify suspicious records before they influence model training. Provenance review determines whether records originated from approved sources and whether unauthorized changes occurred during collection, transfer, or preprocessing.

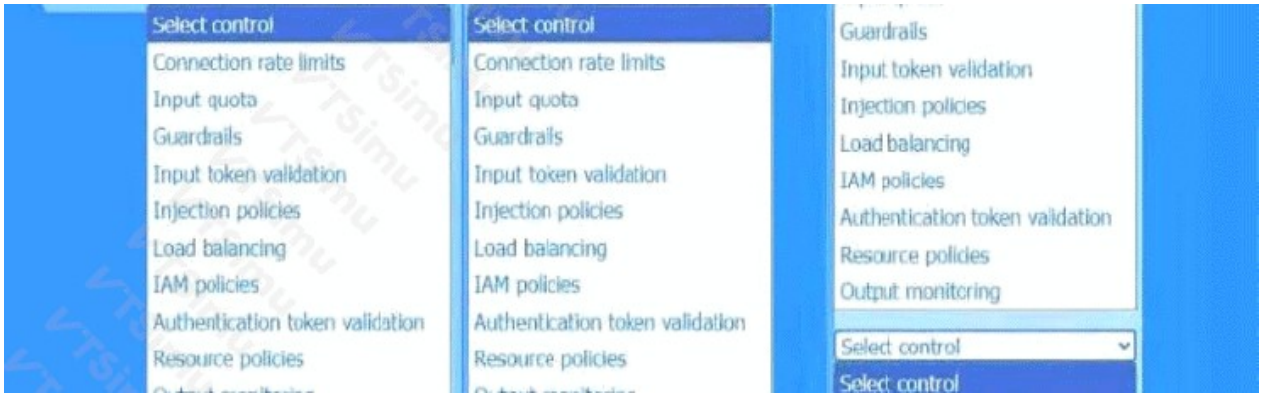
Statistical outlier analysis can identify unusual feature values, label distributions, or record clusters that differ significantly from expected phishing data. Suspicious records should be investigated and excluded or corrected before the model is retrained. Data poisoning is an attack against the training process, so protecting data integrity and validating training inputs are essential. MITRE includes data poisoning techniques in the MITRE ATLAS knowledge base.

QUESTION NO: 6 - (HOTSPOT)

Instructions: Use the drop-down menus to define two appropriate security controls for each component of the AI system. Each control may be used only once.

An engineer is deploying a new AI system and wants to integrate it into the core system through an API.


Cloud control plane
Select control
Connection rate limits
Input quota
Guardrails
Input token validation
Injection policies
Load balancing
IAM policies
Authentication token validation
Resource policies
Output monitoring
External client
Next-generation edge appliance
Front-end API
Model
Vector database
Prompt firewall
WAF


Select control
Connection rate limits
Input quota
Guardrails
Input token validation
Injection policies
Load balancing
IAM policies
Authentication token validation
Resource policies
Output monitoring

ANSWER:

Explanation:

The appropriate selections implement defense in depth by applying each control at the layer where it can most effectively govern access, validate requests, preserve availability, or constrain AI behavior. In the cloud control plane, select **IAM**

policies and **resource policies**. These controls establish least-privilege permissions for identities and define permitted access to cloud-hosted resources that support the AI deployment. Cloud policy enforcement is the proper foundation for administrative and service-to-service authorization. Guidance on policy-based authorization is available from AWS IAM policies and permissions.

At the prompt firewall, select **input token validation** and **injection policies**. A prompt firewall sits immediately before model processing, making it the right place to parse and validate incoming prompt content and apply rules intended to identify or block prompt-injection patterns. This protects the model before untrusted text can influence its instructions, tool usage, or data access. Prompt injection is a central LLM security concern described by the OWASP guidance on prompt injection.

At the WAF, select **authentication token validation** and **connection rate limits**. The edge layer should verify that API callers present valid credentials and limit excessive connection behavior before traffic is allowed farther into the application. For the front-end API, select **load balancing** and **input quota**. These availability controls distribute legitimate workloads and ensure that no client or tenant consumes an unfair or unsafe share of request-processing capacity.

For the model, select **guardrails** and **output monitoring**. Guardrails constrain model behavior according to defined safety and operational rules, while output monitoring reviews generated content for unsafe, sensitive, or policy-violating responses. Together, these measures provide a final control layer around model behavior and generated output, consistent with the governance and monitoring principles in the NIST AI Risk Management Framework.

QUESTION NO: 7

A company is preparing to release a large language model (LLM) application that summarizes documents uploaded by external users. The security team needs to evaluate whether malicious content embedded in documents can influence the application or connected tools.

Which of the following testing activities should the security team perform? (Choose two.)

A. Indirect prompt-injection testing

B. Tool-use authorization testing

C. Storage-capacity testing

D. Model quantization testing

E. Screen-resolution testing

ANSWER: A B

Explanation:

Indirect prompt-injection testing and **tool-use authorization testing** are appropriate because uploaded documents are untrusted content that may contain instructions intended to influence model behavior. Indirect prompt-injection testing determines whether malicious instructions within a document can override the application's intended behavior, cause disclosure of protected information, or alter the model's response.

Tool-use authorization testing verifies that the model cannot invoke connected services with privileges beyond those required for the user request. Tools should enforce authorization independently of model output and should not treat generated text as trusted instructions. OWASP identifies prompt injection and excessive agency as key risks for LLM applications in the OWASP Top 10 for LLM Applications.

QUESTION NO: 8

A human resources officer is using AI to evaluate resumes and help select candidates that meet minimum criteria. To improve the results, the human resources officer adjusts the query parameters and includes an example resume that matches a successful candidate.

Which of the following best describes this query?

A. Distillation

B. Prompt template

C. One-shot prompting

D. System role

ANSWER: C

Explanation:

One-shot prompting is correct because the query provides the AI with one concrete example of the desired kind of output or evaluation target: a resume representing a successful candidate. This example supplies in-context guidance that helps the model infer how the stated minimum criteria should be interpreted and applied to subsequent resumes. Rather than only stating abstract requirements, the human resources officer includes a single representative sample that demonstrates the qualities, experience, formatting, or qualifications associated with a suitable candidate.

In prompt-engineering terminology, zero-shot prompting supplies instructions without examples, one-shot prompting supplies exactly one example, and few-shot prompting supplies multiple examples. The scenario explicitly identifies one example resume, so it matches one-shot prompting. The adjusted query parameters can further refine the instruction, but the defining feature is the single exemplar embedded in the request. Using an example in this way can improve consistency in an AI-assisted screening workflow by grounding the model's assessment in a stated reference profile. See the Prompt Engineering Guide's discussion of few-shot prompting, which includes one-shot prompting as the single-example case, and Google Cloud's prompt design guidance for guidance on using examples to shape model responses.

QUESTION NO: 9

A security operations center (SOC) analyst needs to automate multiple security tasks by breaking them down into smaller parts.

Which of the following AI tools is the best for this task?

A. Agentic AI

B. Retrieval-augmented generation (RAG) AI

C. Generative AI

D. Chatbot

ANSWER: A

Explanation:

Agentic AI is the best fit because it is designed to pursue a defined objective through planning and coordinated execution. A key capability of an AI agent is task decomposition: it can take a larger SOC objective—such as investigating a suspicious alert—and organize it into smaller actions, such as collecting relevant telemetry, enriching indicators with threat-intelligence data, correlating findings, determining a priority, documenting the result, and initiating an approved response workflow. This is particularly useful in security operations because investigations often involve multiple systems, repeated decisions, and conditional next steps rather than a single generated answer.

Agentic systems can also use tools and act on intermediate results. For example, an agent may query a SIEM, retrieve endpoint details, assess whether an indicator is known malicious, and then continue the investigation or trigger an escalation based on defined policies. This goal-oriented orchestration distinguishes agentic AI as the appropriate tool for automating multi-step security work. IBM describes agentic AI as systems that autonomously plan and execute complex workflows, while Google Cloud explains that AI agents reason, plan, and use tools to complete tasks. See IBM's agentic AI overview and Google Cloud's AI agents overview.

QUESTION NO: 10 - (SIMULATION)

An engineer is analyzing findings from a penetration test that indicate insufficient data encryption. The report also indicates that additional controls must be placed on the data. To protect against loss of intellectual property, the engineer must implement data security.

Part 1: Use drop-down menu to select the most appropriate protocol or cipher for each system component.

Part 2: Use the drop-down menu to select the most appropriate technique to apply to the modified data.

ANSWER: See the explanation for the answer

Explanation:

Part 1 — Protocol/cipher selections

TLS 1.2 protects API communications in transit. CAMELLIA is a symmetric block cipher appropriate for encryption of database data at rest. HMAC-SHA512 provides an additional integrity/authenticity control for the stored AI model, helping identify unauthorized model changes.

Part 2 — Select the techniques in the displayed top-to-bottom row order

QUESTION NO: 11

Which of the following describes the number of training cycles used in an AI model for threat detection?

A. k-means clustering

B. Tokens

C. Temperature

D. Epoch

ANSWER: D

Explanation:

Epoch is correct because an epoch represents one complete training cycle: the model processes the full training dataset once and uses the resulting feedback to update its learned parameters. Therefore, the number of epochs specifies how many complete passes an AI threat-detection model makes through its available training examples. These examples may include normal activity, malicious events, phishing indicators, malware attributes, anomalous network behavior, or other labeled security telemetry. Repeating full passes allows the model to progressively reduce training error and recognize relevant patterns more reliably. Epoch count is a key training hyperparameter because it affects both the duration and effectiveness of training. A sufficiently selected value gives the model repeated opportunities to learn the relationships in its data, while training monitoring can help practitioners determine when additional cycles no longer improve generalization. Google's machine-learning glossary defines an epoch as a full training pass over the entire training set. TensorFlow similarly

documents the `epochs` setting as the number of times a training process iterates over the input data. See the Google Machine Learning Glossary and TensorFlow Keras training guide.

QUESTION NO: 12

A SOC analyst identifies that a user extracted the full system prompt from the company's chatbot by prompting it to repeat the last query and provide the entire conversation context. Which of the following mitigations reduces the risk to the AI system?

A. Restricting the LLM's access to internal services

B. Using data version control to detect content manipulation

C. Enhancing model guardrails

D. Segregating and identifying external content

ANSWER: C

Explanation:

Enhancing model guardrails is the appropriate mitigation because this incident is a system-prompt extraction attack: a user manipulated the chatbot into disclosing instructions and conversation context that should remain inaccessible. Guardrails are application and model-layer controls that enforce boundaries on what the chatbot can process and return. In practice, they can identify requests for hidden prompts, internal instructions, prior conversation data, policy content, or protected context; apply explicit refusal behavior; and inspect generated output before it is displayed to the user. Effective guardrails should be implemented as defense in depth, including clear system-level policies, input classification, output filtering, secure handling of conversation state, and testing against indirect prompt-injection and extraction attempts. These measures reduce the likelihood that a model follows an untrusted user instruction that conflicts with protected application instructions. OWASP identifies prompt injection and sensitive-information disclosure as significant risks for LLM applications, reinforcing the need for controls around model inputs, outputs, and privileged context. NIST's AI Risk Management Framework likewise supports managing operational AI risks through documented controls, monitoring, and ongoing assessment. See the OWASP guidance on prompt injection and the NIST AI Risk Management Framework.

QUESTION NO: 13

A company is deploying a retrieval-augmented generation (RAG) assistant that searches internal engineering documents. The company must prevent employees from using the assistant to retrieve documents outside of their authorized projects.

Which of the following controls should the company implement? (Choose two.)

A. Document-level access controls

B. Identity propagation

C. Increased output token limits

D. Public document indexing

E. Model quantization

ANSWER: A B

Explanation:

Document-level access controls are necessary because a RAG system must enforce a user's existing authorization before content is retrieved from the knowledge base. The retrieval component should filter documents according to the authenticated user's permissions so unauthorized content is never added to the model context.

Identity propagation is also necessary because the RAG application must pass the requesting user's identity and entitlements to the retrieval layer. This permits the data source or authorization service to make an identity-aware access

decision rather than granting the AI application broad access on behalf of all users. These controls implement least privilege and prevent the RAG interface from bypassing existing data permissions. NIST access-control guidance is available in NIST SP 800-53 Rev. 5.

QUESTION NO: 14

An airline corporation wants to implement a chatbot application using a large language model (LLM) so its customers can ask questions and receive answers about flight details and have the option to upload files.

Which of the following security controls should the airline use to protect against malicious input and unauthorized use beyond the service-level agreement? (Choose two.)

A. Prompt guardrails

B. Role-based access controls

C. Firewall rules

D. Model token quotas

ANSWER: A D

Explanation:

Prompt guardrails are appropriate because a public chatbot processes untrusted customer prompts and potentially untrusted uploaded-file content. Guardrails apply policy checks before and after model processing to detect or block prompt-injection attempts, instruction overrides, unsafe requests, and content that is outside the chatbot's approved airline-support scope. They can also constrain the model's responses and tool use so that customer-provided text does not cause unintended behavior. This is particularly important where uploaded documents may contain embedded or indirect instructions intended to manipulate the model. Prompt injection is recognized as a major application-security concern for LLM deployments by the OWASP GenAI Security Project.

Model token quotas protect the service against unauthorized or excessive consumption beyond the agreed service limits. Quotas limit tokens used in requests and generated responses, and can be applied alongside request-rate limits to control repeated interactions, unusually large prompts, and file-derived content. This reduces exposure to resource exhaustion, unexpected inference costs, and denial-of-wallet-style abuse while helping maintain predictable availability for legitimate airline customers. LLM platform quota controls are documented by Microsoft Azure OpenAI quotas and limits.

QUESTION NO: 15

A security team is preparing to test an ML malware-classification model before deployment. The team wants to identify weaknesses that an adversary could exploit to cause incorrect classifications.

Which of the following testing activities should the team perform? (Choose two.)

A. Adversarial example testing

B. Data-poisoning simulation

C. Increasing the model temperature

D. Expanding output token quotas

E. Disabling validation data

ANSWER: A B

Explanation:

Adversarial example testing evaluates whether carefully modified inputs can cause the model to misclassify malicious or benign files. This helps the team assess the model's robustness against evasion attempts that preserve much of an input's original appearance or function while changing the model's prediction.

Data-poisoning simulation evaluates the effect of malicious or corrupted training data on the model's behavior. This testing can reveal whether an attacker could manipulate labels, inject biased samples, or insert backdoor patterns that affect later inference results.

Both activities focus on AI-specific adversarial risks that should be evaluated before production deployment. MITRE documents adversarial ML tactics and techniques in the MITRE ATLAS knowledge base.

QUESTION NO: 16

A SOC deploys an AI agent that can gather endpoint information and submit containment requests to the SOAR platform. The organization must maintain accountability for the agent's activities during incident response.

Which of the following should the SOC implement? (Choose two.)

A. Comprehensive audit logging

B. Human approval for high-impact actions

C. Unlimited tool permissions

D. Shared administrative credentials

E. Disabled incident records

ANSWER: A B

Explanation:

Comprehensive audit logging provides an accountable record of the agent's requests, tool calls, data accessed, recommendations, approvals, and resulting actions. These records allow the SOC to reconstruct events, investigate errors, and demonstrate that actions followed approved processes.

Human approval for high-impact actions ensures that an authorized analyst evaluates consequential actions, such as isolating a critical host or disabling an executive account, before the SOAR platform executes them. This maintains human oversight while still allowing the agent to accelerate enrichment and triage.

These measures support controlled automation and reduce operational risk. NIST identifies audit and accountability requirements in NIST SP 800-53 Rev. 5, and the NIST AI Risk Management Framework emphasizes appropriate human-AI oversight.