Sample Exam

Edition 202608

Copyright © EXIN Holding B.V. 2026. All rights reserved.
EXIN® is a registered trademark.

No part of this publication may be reproduced, stored, utilized or transmitted in any form or by any means, electronic, mechanical, or otherwise, without the prior written permission from EXIN.

Content

Introduction	4
Sample exam	5
Answer key	19
Evaluation	47

Introduction

This is the EXIN AI Security Professional based on the OWASP AI Exchange (AISP.EN) sample exam. The Rules and Regulations for EXIN's examinations apply to this exam.

This exam consists of 40 multiple-choice questions. Each multiple-choice question has a number of possible answers, of which only one is correct.

The maximum number of points that can be obtained for this exam is 40. Each correct answer is worth 1 point. You need 26 points or more to pass the exam.

The time allowed for this exam is 90 minutes.

Good luck!

Sample exam

1 / 40
An IT security manager works at a logistics company. The company has recently decided to adopt AI tools to optimize its delivery routes. The IT security manager leads the effort to organize AI security. The first two steps of the G.U.A.R.D. framework are already completed:

- The Govern step is done by setting up an AI security policy.
- The Understand step was done by mapping all AI assets and their risks.

What is the **next** step the IT security manager should take?

A) Adapt security practices to include AI-specific threat modeling, testing, and supply chain controls
B) Analyze the identified AI risks and establish a risk prioritization strategy based on their potential impact
C) Apply security controls across all AI systems to establish responsible AI security throughout the AI lifecycle
D) Assess the AI security policy created during the Govern step to include the mapped AI assets and their risks

2 / 40
A company reviews its AI strategy to ensure the AI strategy benefits the business while being ethical and operationally sound. The AI security officer categorizes the two main dimensions of responsible and trustworthy AI.

What is the **main** difference between responsible AI and trustworthy AI?

A) - Responsible AI focuses on ethics, society, and governance.
- Trustworthy AI focuses on technical and operational aspects, transparency, and explainability.
B) - Responsible AI focuses on regulatory compliance with relevant laws and legislation.
- Trustworthy AI focuses on data protection and cybersecurity to align with privacy principles.
C) - Responsible AI focuses on system accuracy and technical performance metrics.
- Trustworthy AI focuses on ethical guidelines and social impact to create consumer trust.
D) - Responsible AI focuses on vendor selection, risks, and cost management.
- Trustworthy AI focuses on safe and reliable user training and AI systems adoption.

3 / 40
A cybersecurity team already has a threat detection framework. Now they must also protect a new large language model (LLM) used in customer support for a chatbot.

The team wants to use the same tools they use for web applications:
- Signature-based detection rules
- Fixed input validation controls

Why is this approach **not** enough to secure the LLM?

A) Because AI systems have different attack surfaces and require specific controls for the LLM's behavior and data pipeline
B) Because the threat detection framework does not include incident response procedures, which makes it unsuitable for LLMs
C) Because LLMs generate outputs that are too variable for rule-based systems to evaluate and must be reviewed by a human
D) Because LLMs process requests faster than web applications, meaning signature-based rules cannot keep up with them
E) Because web application firewalls are not licensed for use with LLMs and require separate vendor agreements

4 / 40
When using AI, a company will have AI-specific assets. Each AI-specific asset also has its own key threats.

What is a **correct** combination of AI-specific asset and its key threat?

A) - Asset: audit logs generated by network firewalls
- Threat: AI resource exhaustion
B) - Asset: end-user authentication credentials
- Threat: output contains conventional injection
C) - Asset: software source code stored in a version control system
- Threat: direct runtime model poisoning
D) - Asset: training data used to build a model
- Threat: sensitive data disclosure in output

5 / 40
A law firm uses an AI system to do document review and analysis. The security team applies a risk management approach. They have completed an inventory of all potential threats to the AI system.

What is the **next** step the team should take?

A) Apply a threat modeling process to bridge the threats to a set of concrete and prioritized risks.
B) Assign a risk owner to each identified threat to ensure protection by design since the beginning.
C) Deploy security controls for every threat to mitigate further risks across the AI system lifecycle.
D) Share the full threat list with relevant stakeholders to raise awareness and establish transparency.

6 / 40
An AI security engineer reviews the access control design for an AI multi-agent system. Each agent in the system is provisioned with a shared service account. This account has read-and-write access to the production database, the internal file system, and the external API (application programming interface) gateway.

The different AI agents autonomously invoke tools and execute multi-step workflows across internal systems. The provisioning was done on a shared account to simplify credential management across the AI agent pipeline.

Why is this shared account design **particularly risky** in an agentic AI context?

A) Because AI agents generate higher volumes of API calls than human users, so it is more likely that rate-limiting controls are triggered and the shared account gets suspended
B) Because AI agents do not use multi-factor authentication, so the shared credential provides a single point of failure with no other verification layers, which an attacker can exploit
C) Because AI agents operate autonomously across multiple reasoning steps within a single session, so a compromised agent can chain actions across all accessible systems
D) Because regulatory frameworks for AI systems require that each AI agent is provisioned with a unique credential, so shared accounts are a direct compliance violation
E) Because sharing service accounts across multiple AI agents creates too much operational complexity during mandated credential rotation and activities in access management

7 / 40
A machine learning engineer at a financial institution is investigating two separate incidents involving the institution's fraud detection model.

- **Incident 1**: An external attacker, with no access to the model's architecture, training data, or parameters, submits thousands of crafted transactions. These are iteratively refined based on whether the fraud alert was triggered or not.
- **Incident 2**: A researcher uses a locally trained surrogate model with a similar architecture, to generate adversarial transactions. These successfully evade the production model without prior interaction.

Which two evasion strategies are these?

A) - Incident 1: evasion after poisoning
- Incident 2: perfect-knowledge evasion
B) - Incident 1: partial-knowledge evasion
- Incident 2: transfer attack
C) - Incident 1: perfect-knowledge evasion
- Incident 2: zero-knowledge evasion
D) - Incident 1: transfer attack
- Incident 2: partial-knowledge evasion
E) - Incident 1: zero-knowledge evasion
- Incident 2: transfer attack

8 / 40
Erick is a user of a social media platform which automatically moderates comments using an AI system. He tries to get a policy-violating comment past the platform's automated moderation because he wants to boost engagement on his post. Erick does not have information about the model's architecture, parameters, or training data, and he only sees a binary message when he tries to post his comment, either "blocked" or "posted". By repeatedly making small changes to the wording of his comment, Erick eventually finds a version that gets posted.

What type of evasion input threat is described here?

A) Zero-knowledge evasion
B) Partial-knowledge evasion
C) Perfect-knowledge evasion
D) Evasion after poisoning
E) Transfer attack

9 / 40
A legal services company develops a chatbot that answers user's questions and can retrieve information from internal and external resources, like websites and company documents. During a demo, the chatbot's behavior triggers an investigation by the AI security engineer. She identifies two separate threats:

- **Threat 1:** A message crafted to override the chatbot's instructions and safety rules can cause the chatbot to follow the tester's commands instead of its configured policy.
- **Threat 2:** A vendor webpage retrieved by the chatbot contains embedded text that acts like instructions. When the chatbot ingests that page, it treats the content as commands and deviates from its policy.

What types of threat are these?

A) - Threat 1: evasion
- Threat 2: data poisoning

B) - Threat 1: data poisoning
- Threat 2: direct prompt injection

C) - Threat 1: direct prompt injection
- Threat 2: indirect prompt injection

D) - Threat 1: indirect prompt injection
- Threat 2: evasion

10 / 40
A consulting company has an AI system in place to assist with day-by-day tasks. After a meeting with a client, the AI assistant summarizes the call. The AI assistant starts to insert a confidential internal links into the draft e-mail to be sent to the client. However, before sending, a self-operating anomaly detector pauses the sending and quarantines the draft.

The AI security architect wants to limit the AI meeting's assistant's access in advance only to the rights of the individuals being served.

Which prompt injection protection layers are illustrated here?

A) Automated oversight and user-based least privilege
B) Just-in-time authorization and human oversight
C) Model alignment and intent-based least privilege
D) Prompt injection input/output handling and rate limit

11 / 40
An auditor is assessing the privacy risks of a deployed healthcare large language model (LLM) and its compliance with data protection laws. To do that, the auditor probes the model with targeted queries about a specific patient and carefully matched synthetic non-patients. By observing differences in the model's response patterns and confidence scores, the auditor can reliably determine whether that patient's record was included in the training data, even though no actual medical details are revealed.

What type of attack is simulated here?

A) Data disclosure in model output
B) Indirect prompt injection
C) Membership inference
D) Model inversion

12 / 40
What is a **main** consequence of a successful model exfiltration attack?

A) The attacker can extract the raw training dataset used to build the original model, potentially including sensitive records.
B) The attacker gains direct control over the deployed model and can modify its weights and parameters without prior authorization.
C) The attacker obtains a replica of the model that can be used to craft evasion attacks without triggering the original's defenses.
D) The original model becomes unavailable to legitimate users due to resource exhaustion, causing service degradation and repeated timeouts.

13 / 40
A cloud-hosted large language model (LLM) is hit with a sponge attack. Attackers send very long or tricky inputs that force extra computation, driving up costs and causing slowdowns or outages. In order to address this situation, the AI security engineer wants to put two controls in place:

- **Control 1** will validate and sanitize inputs to reject or correct malicious content, such as abnormally large inputs.
- **Control 2** will restrict the computational resources consumed per model input to prevent overuse.

What are these two controls?

A) - Control 1: anomalous input handling
- Control 2: obscure confidence limits

B) - Control 1: denial-of-service (DoS) input validation
- Control 2: limit resources

C) - Control 1: model access control
- Control 2: evasion robust model

D) - Control 1: monitor use
- Control 2: rate limit

14 / 40
An AI security engineer identifies the following threats for a company:

- **Adversarial prompts** are crafted by attackers to cause misclassifications.
- A **compromised ingestion pipeline** introduces backdoored records into the corpus before the model is built.
- Attackers tamper with **labeling workflows** changing ground-truth annotations to bias model behavior.
- **Training data** is obtained from external or third-party sources without proper validation.

According to the OWASP AI Exchange, which threat does **not** involve risks of data poisoning during development-time?

A) Adversarial inputs
B) Compromised ingestion pipeline
C) Labeled workflows
D) Training data

15 / 40
According to the OWASP AI Exchange, what is an example of **direct development-time model poisoning**?

A) Attackers craft adversarial inputs at runtime to mislead the model
B) Attackers manipulate weights and a training pipeline to alter the model's behavior.
C) Sensitive training data is unintentionally exposed through model outputs.
D) Surge of traffic overwhelms the AI service causing it to slow down or have downtime.

16 / 40
According to the OWASP AI Exchange, what is an example of **supply-chain model poisoning**?

A) Attackers compromise supply-chain data, models, or components before integration.
B) Attackers manipulate inputs at runtime to trick the model into producing incorrect outputs.
C) The model produces incorrect results due to insufficient data or training quality.
D) The model unintentionally leaks sensitive training data through its responses.

17 / 40
An AI development team is conducting a risk assessment of their development environment. They identify two distinct threats:

Threat 1 involves unauthorized access to the datasets used to train and test their model.
Threat 2 involves unauthorized access to the model's parameters and weights stored in a compromised repository.

What types of threat are these?

A) - Threat 1: development-time data leak
- Threat 2: direct development-time model leak
B) - Threat 1: direct augmentation data leak
- Threat 2: repository leak
C) - Threat 1: direct runtime data leak
- Threat 2: direct runtime model leak
D) - Threat 1: input data leak
- Threat 2: source code/configuration leak

18 / 40
A security team reviews the runtime threat landscape of an AI system. The AI system is integrated into a web application. The team identifies two conventional security threats:

- **Threat 1:** An attacker can gain access to the production server's memory and extracts the model's weights and parameters.
- **Threat 2:** an attacker with access to the model registry replaces the deployed model artifact with a trojaned version.

What consequences do these conventional security threats have that are **specific** to AI systems?

A) - Threat 1 causes the system to auto-retrain on malicious data.
- Threat 2 forces the model to reject valid prompts.
B) - Threat 1 changes the model’s architecture at runtime.
- Threat 2 silently compresses weights to lower precision.
C) - Threat 1 disables safety filters making the system more permissive by design.
- Threat 2 makes outputs deterministic and predictable.
D) - Threat 1 enables model theft and potential training data inference.
- Threat 2 allows undetected model tampering.

19 / 40
A financial services company deploys a machine learning (ML) model to detect fraudulent transactions. The ML model is hosted in their production environment and continuously processes live transaction data. Two security incidents occur:

- **Incident 1**: An attacker gains unauthorized access to the deployed model environment and alters the model's parameters so that certain fraudulent transactions are no longer flagged. The manipulation is not detected, and the attacker benefits from this.
- **Incident 2**: An attacker exploits a vulnerability in the production system. This allows the attacker to replicate the model and analyze its behavior for further attacks.

Based on the OWASP AI Exchange, what type of security incidents are described?

A) - Incident A: data poisoning
- Incident B: direct runtime model poisoning

B) - Incident A: data poisoning
- Incident B: evasion attack

C) - Incident A: direct runtime model poisoning
- Incident B: direct runtime model leak

D) - Incident A: direct runtime model leak
- Incident B: evasion attack

20 / 40
A healthcare organization deploys an AI assistant to help clinicians summarize patient records and provide recommendations. The system integrates with internal databases and external tools, and clinicians rely on its generated outputs during daily operations. Two threats are observed:

- **Threat 1**: The AI assistant generates a response that includes a hidden malicious script embedded in HTML content. When the output is viewed, the script executes and sends session data to the attacker.
- **Threat 2**: An attacker exploits weak access controls in the system to intercept and retrieve sensitive patient data.

Based on the OWASP AI Exchange, what types of threat are described?

A) - Threat 1: data poisoning
- Threat 2: input data leak

B) - Threat 1: data poisoning
- Threat 2: model inversion attack

C) - Threat 1: output containing conventional injection
- Threat 2: input data leak

D) - Threat 1: output containing conventional injection
- Threat 2: model inversion attack

21 / 40
A company uses a retrieval-augmented generation (RAG) system to answer employee questions about internal policies. The AI system retrieves documents from a vector database and adds the documents to the prompt before sending them to the model.

A security review finds that an attacker gained unauthorized access to the vector database. The attacker viewed stored augmentation documents and replaced a legitimate policy document with another one, containing false information. Employees who ask the AI system about that policy now receive incorrect guidance.

Which security threats does this scenario illustrate?

A) - Threat 1: augmentation data leak
- Threat 2: output containing conventional injection

B) - Threat 1: augmentation data manipulation
- Threat 2: augmentation data leak

C) - Threat 1: membership inference
- Threat 2: augmentation data manipulation

D) - Threat 1: output containing conventional injection
- Threat 2: membership inference

22 / 40
General governance controls should be implemented to ensure effective security oversight of AI systems.

Which approach demonstrates this **best**?

A) Defining clear policies, roles, and risk management practices to guide the secure development, deployment, and monitoring of AI systems

B) Including security controls in the AI systems to apply security-by-design principles in the design stage of the AI system

C) Meeting minimum security requirements as AI systems are more secure by default and do not require multiple security layers

D) Separating governance controls for AI from information security controls to make sure these two different concepts are equally covered

23 / 40
A mid-size marketing agency has little AI experience and no formal AI governance yet. They want to use a generative AI tool to draft customer e-mails. The AI toll needs access to customer names and e-mails. The company should establish AI governance and establish security oversight before they can use the AI tool.

What is the **bare minimum** the company should do for security oversight?

A) Conduct red team testing of the generative AI e-mail system to mitigate security risks before launch

B) Hire an AI governance specialist to do a threat assessment and define policies and controls for deployment

C) Inventory the planned AI use and perform a risk analysis to identify threats, controls, and responsibilities

D) Request legal advice from a company specialized in AI and the related AI Act requirements

E) Start with a pilot and monitor the AI e-mail system for performance, misuses, and incidents with logs

24 / 40
How should the coverage of general governance controls in AI security be **best** understood?

A) As a collection of encryption methods used to secure AI outputs after generation
B) As a set of overarching controls that apply across all AI threats and lifecycle stages
C) As controls limited to the model training phase, focusing on dataset quality and tuning techniques
D) As optional measures that organizations may adopt if the AI systems do not process sensitive data
E) As technical safeguards applied exclusively at runtime to detect malicious inputs

25 / 40
A company in healthcare decides to use a ready-made AI model, provided by a third-party, for an application. The application processes user input and retrieves external data.

The AI security engineer is worried about the model's security. The security strategy that defines the division of security responsibilities between the AI model provider and the healthcare company is not defined.

What is **correct** about these responsibilities?

A) Most security controls can be delegated to the provider when using ready-made AI models, since the provider's training and hosting cover security risks.
B) The deployer is responsible for all security controls when they integrate the AI model into their systems and consider security before deployment.
C) The provider has trained the AI model, so they manage model-level controls, while the deployer should manage application-level controls.
D) The provider is responsible for all security controls because it trained the AI model and hosts it, so all risks and attacks should be handled by the provider.

26 / 40
An organization has deployed a customer-facing AI application using a large language model (LLM) supplied by a third-party provider. The model is accessed through a managed API (application programming interface) service and has not been further trained or modified by the organization.

The security team discovers that carefully crafted user inputs can cause the model to ignore its configured behavioral boundaries and produce outputs that violate organizational content restrictions. The team considers which control to implement that falls within the organization's responsibility as a deployer rather than the third-party provider's.

Which control should the security team implement to address the issue?

A) Insert an output validation layer that scans all model responses before they reach the end users and block outputs that violate the organization's content policy
B) Modify the provider's inference pipeline to enforce stricter generation constraints that limit response generation and prevent disallowed content from emerging
C) Request that the provider retrain and safety-tune the supplied model using additional refusal and policy-alignment data to improve resistance against jailbreak inputs
D) Require the provider to revise and apply strengthened, service-wide content moderation rules and enforcement so that unsafe outputs are filtered globally before delivery

27 / 40
A team is preparing training data for a generative AI assistant. The team finds that they do not need all the fields and data they currently use for the AI model's performance. As a next step, they are considering controls to limit sensitive data and improve confidentiality and integrity.

What is the **best** way to improve confidentiality and integrity?

A) Keep all the fields in the training data and encrypt the data at rest and in backups
B) Preserve identifiers in training data to simplify future retraining tasks and processes
C) Remove fields and data that are unnecessary for model training and performance
D) Use only synthetic data for all generative AI training and evaluation phases

28 / 40
A product team is training an AI model for customer support. The current training set includes full customer addresses, phone numbers, and legacy records that are not needed to meet performance goals but are likely useful in the future. After a red team exercise highlights a risk of data leakage and manipulation via memorized details, the team decides to reduce exposure. They do this by removing unnecessary fields and entire records so that sensitive information is not present in the dataset.

Which **general** control was implemented to limit sensitive data in this situation?

A) Allowed data
B) Data minimization
C) Obfuscate training data
D) Short retain

29 / 40
A retail company decides to stop storing customer payment details after each transaction is completed.

How does this action **reduce** the risk of data exposure?

A) By eliminating the need for encryption of data in transit or archived for auditability purposes
B) By ensuring compliance with all relevant data protection regulations, reducing complexity
C) By guaranteeing that no unauthorized user can access the company network and database
D) By preventing retained data from being leaked, reconstructed or inferred from the system

30 / 40
A regional hospital network uses an AI agent to read inbound vendor e-mails, draft replies, and automatically send responses. During a security testing following a recent phishing incident, testers found that prompt injection could make the AI agent ignore its rules and send unauthorized responses.

Which design decision works **best** to limit the effects of this unwanted behavior?

A) Disable all AI agent integrations and revert to fully manual human workflows for most processes
B) Expand user training for prompt safety, and emphasize oversight during agent-assisted e-mailing
C) Keep broad autonomous send permissions, and upgrade to a more advanced generative model
D) Restrict the AI agent permissions to task-specific scopes, and require approval for high-risk actions

31 / 40
What is the **best** way to minimize the effects of unwanted model behavior?

A) Emphasizing prompt hardening, role-based access, and scheduled audits of model outputs
B) Telying on automated output filters, exception logging, and occasional fairness spot-checks
C) Testing for unwanted bias and applying controls like oversight, least privilege, and continuous validation
D) Using strong input validation, constrained tool permissions, and periodic human-in-the-loop reviews

32 / 40
Limiting unwanted model behavior is a critical strategy for impact limitation and blast radius control.

How does this strategy **improve** performance and **reduce** risk?

A) By focusing on attacks that can affect AI systems and limiting oversight over system performance factors and accuracy
B) By implementing controls that remove the need for continuous testing and validation over the time to reduce complexity
C) By improving operational reliability and resource efficiency while minimizing incidents and wasted compute requests
D) By increasing model autonomy and allowing it to generate more outputs to continuously improve system performance

33 / 40
Within the broader scope of AI red teaming, what is the **primary** purpose of AI security testing?

A) Assessing whether an AI model is resilient to specific attacks by reproducing those attacks in a controlled environment
B) Evaluating whether an AI model produces accurate and traceable outputs when exposed to expected input data
C) Measuring whether an AI model demonstrates the expected computational efficiency when processing large volumes of requests
D) Verifying whether an AI model meets its defined business requirements through standard quality assurance procedures

34 / 40
A company deploys a generative AI assistant and wants to define AI security tests that cover more than just conventional pen testing. To do that, the testing team starts identifying the threats that should be tested for.

Which threats **specifically** associated with generative AI systems should be included in the AI security testing?

A) Evasion attacks, insecure output handling, and model poisoning
B) Model exfiltration, prompt injection, and evasion attacks
C) Prompt injection, sensitive data disclosure, and insecure output handling
D) Sensitive data output from model, model drift, and unsafe tool invocation

35 / 40
An AI security team is testing a generative AI chatbot for prompt injection resistance. During the test, the system has blocked several crafted attack inputs.

What is the recommended **next** step to maintain security oversight of the generative AI chatbot?

A) Add input variation such as synonym replacement, encoding changes, or formatting shifts to attempt to circumvent detection
B) Conclude the current testing cycle and deem the defenses adequate since most prompt injection attempts were blocked
C) Run the same test several times to increase sample size and obtain more statistically relevant defense performance data
D) Send the prompt injection attack inputs directly to the model to better train it against similar threats in future deployments

36 / 40
The AI system of a company relies on large amounts of personal data for training and decision-making.

From an ethical perspective, which risk is **most** important for the company to address?

A) Data being collected, retained, or reused beyond justified purposes
B) Interface complexity making automated decisions difficult to review
C) Model accuracy dropping when resources are unevenly allocated
D) Software updates introducing compatibility issues across platforms

37 / 40
An online bookshop uses an AI system to analyze browsing history, purchase trends, and preorders so it can predict demand and make sure popular books are in stock on time. Later, the company reuses that same customer data to train a separate AI model that builds detailed marketing profiles and sends targeted ads to individual shoppers.

Which privacy principle is **most directly** violated here?

A) Data minimization
B) Data retention limitation
C) Transparency and consent
D) Use limitation and purpose specification

38 / 40
An organization wants to set up a framework to govern the responsible use of AI across its operations.

They use the ISO/IEC 27001 standard for information security management. The security team wants to set up their AI governance with the ISO/IEC standard that serves the same structural purpose for AI governance as the ISO/IEC 27001 standard does for information security.

Which standard is this?

A) ISO/IEC 23894
B) ISO/IEC 27005
C) ISO/IEC 42001
D) ISO/IEC 5338

39 / 40
A company wants to use an AI system to automatically screen and rank job applicants, based on their résumés.

According to the AI Act, in which category should the use of this AI system be classified?

A) Unacceptable risk
B) High risk
C) Limited risk
D) Minimal or no risk

40 / 40
An AI development team is preparing training data for a new machine learning model. The team lead wants to ensure the data sourcing approach supports copyright compliance.

What is the **most ethical** data sourcing practice?

A) Collecting data from publicly available domains only
B) Creating all training data within the organization
C) Downloading data from sources with implied permission
D) Using only data that appears widely reposted online

Answer key

1 / 40
An IT security manager works at a logistics company. The company has recently decided to adopt AI tools to optimize its delivery routes. The IT security manager leads the effort to organize AI security. The first two steps of the G.U.A.R.D. framework are already completed:

- The Govern step is done by setting up an AI security policy.
- The Understand step was done by mapping all AI assets and their risks.

What is the **next** step the IT security manager should take?

A) Adapt security practices to include AI-specific threat modeling, testing, and supply chain controls
B) Analyze the identified AI risks and establish a risk prioritization strategy based on their potential impact
C) Apply security controls across all AI systems to establish responsible AI security throughout the AI lifecycle
D) Assess the AI security policy created during the Govern step to include the mapped AI assets and their risks

A) Correct. After Govern and Understand, the next G.U.A.R.D step is Adapt, which is about updating security practices to incorporate AI-specific threats, testing, and supply chain measures. (Literature: A, Chapter 0)
B) Incorrect. This belongs to the Understand step, which is already completed.
C) Incorrect. Blanket control application without context does not align with the G.U.A.R.D. framework steps. The next step is to adapt practices, not indiscriminately apply controls. Controls will be applied at a later stage.
D) Incorrect. Revisiting policy is part of Govern. The next step after Understand is Adapt, not reworking governance.

2 / 40
A company reviews its AI strategy to ensure the AI strategy benefits the business while being ethical and operationally sound. The AI security officer categorizes the two main dimensions of responsible and trustworthy AI.

What is the **main** difference between responsible AI and trustworthy AI?

A) - Responsible AI focuses on ethics, society, and governance.
- Trustworthy AI focuses on technical and operational aspects, transparency, and explainability.

B) - Responsible AI focuses on regulatory compliance with relevant laws and legislation.
- Trustworthy AI focuses on data protection and cybersecurity to align with privacy principles.

C) - Responsible AI focuses on system accuracy and technical performance metrics.
- Trustworthy AI focuses on ethical guidelines and social impact to create consumer trust.

D) - Responsible AI focuses on vendor selection, risks, and cost management.
- Trustworthy AI focuses on safe and reliable user training and AI systems adoption.

A) Correct. Responsible AI emphasizes ethics, society, and governance (principles, policies, oversight), whereas trustworthy AI emphasises the more technical and operational aspects of AI use in an organization, such as robustness, reliability, transparency, and explainability. (Literature: A, Chapter 0)

B) Incorrect. This narrows responsible AI to regulatory compliance and trustworthy AI to privacy and cybersecurity. These are subsets and do not capture the broader difference.

C) Incorrect. This reverses the focus: accuracy and performance are part of technical trustworthiness, while ethics and social impact are central to responsible AI.

D) Incorrect. Vendor selection, cost, and user training are change-management topics, not the core conceptual distinction between responsible and trustworthy AI.

3 / 40
A cybersecurity team already has a threat detection framework. Now they must also protect a new large language model (LLM) used in customer support for a chatbot.

The team wants to use the same tools they use for web applications:
- Signature-based detection rules
- Fixed input validation controls

Why is this approach **not** enough to secure the LLM?

A) Because AI systems have different attack surfaces and require specific controls for the LLM's behavior and data pipeline
B) Because the threat detection framework does not include incident response procedures, which makes it unsuitable for LLMs
C) Because LLMs generate outputs that are too variable for rule-based systems to evaluate and must be reviewed by a human
D) Because LLMs process requests faster than web applications, meaning signature-based rules cannot keep up with them
E) Because web application firewalls are not licensed for use with LLMs and require separate vendor agreements

A) Correct. Conventional cybersecurity tools such as web application firewalls and signature-based detection systems are designed to identify known attack patterns against fixed software interfaces. AI systems introduce entirely new attack surfaces: the training data can be poisoned, the model itself can be inverted to extract sensitive information, and outputs can be manipulated through adversarial inputs that do not match any known signature. These threats require AI-specific controls that operate on the model's behavior and data pipeline, not only on its network interface. (Literature: A, Chapter 0 and 1)
B) Incorrect. Conventional cybersecurity frameworks do include incident response procedures. This option mischaracterizes the limitation of conventional security for AI systems, which is about attack surface coverage, not procedural gaps.
C) Incorrect. While LLM outputs are highly variable, conventional rule-based controls may still contribute to layered defenses in AI environments. The limitation is that they do not adequately address AI-specific attack surfaces such as model manipulation, adversarial inputs, prompt injection, and training data attacks.
D) Incorrect. Processing speed is a performance consideration and is not the reason signature-based detection is insufficient for AI systems. The limitation here is not timing, but the nature of the attack surface.
E) Incorrect. There is no general licensing restriction preventing web application firewalls from being applied to AI model endpoints. This option describes a procurement consideration rather than the substantive security limitation of the proposed approach.

4 / 40
When using AI, a company will have AI-specific assets. Each AI-specific asset also has its own key threats.

What is a **correct** combination of AI-specific asset and its key threat?

A) - Asset: audit logs generated by network firewalls
- Threat: AI resource exhaustion

B) - Asset: end-user authentication credentials
- Threat: output contains conventional injection

C) - Asset: software source code stored in a version control system
- Threat: direct runtime model poisoning

D) - Asset: training data used to build a model
- Threat: sensitive data disclosure in output

A) Incorrect. Audit logs from network firewalls are a general IT security asset, not an AI-specific asset described in the OWASP AI Exchange. AI resource exhaustion typically relates to the overuse of computational resources needed for AI tasks, which does not directly connect to audit logs.

B) Incorrect. End-user authentication credentials are a common IT security concern, not an AI-specific asset described in the OWASP AI Exchange. Conventional injection threats like SQL or command injection are more relevant to application security rather than AI outputs. This pairing does not correctly align.

C) Incorrect. Software source code can be an AI-specific asset if it relates to AI models. However, direct runtime model poisoning pertains to the manipulation of an AI model during its execution, rather than its source code in version control.

D) Correct. Training data is a critical AI-specific asset. Disclosure of data in output is a legitimate threat, especially if sensitive or proprietary data from the training set inadvertently appears in the AI-generated outputs. This combination correctly identifies the asset and its associated threat in an AI context. (Literature: A, Chapter 0)

5 / 40
A law firm uses an AI system to do document review and analysis. The security team applies a risk management approach. They have completed an inventory of all potential threats to the AI system.

What is the **next** step the team should take?

A) Apply a threat modeling process to bridge the threats to a set of concrete and prioritized risks.

B) Assign a risk owner to each identified threat to ensure protection by design since the beginning.

C) Deploy security controls for every threat to mitigate further risks across the AI system lifecycle.

D) Share the full threat list with relevant stakeholders to raise awareness and establish transparency.

A) Correct. After listing threats, the team should apply a threat modeling process, which serves as the bridge between a list of threats and a set of concrete, prioritized risks. This allows the organization to focus on applying controls to the most significant risks. (Literature: A, Chapter 0)

B) Incorrect. Assigning risk owners before analyzing and prioritizing threats skips a critical step. Without understanding the concrete impact and likelihood of each threat, ownership assignments lack meaningful context.

C) Incorrect. Deploying controls for every threat without prioritization is inefficient and may misallocate resources. Prioritization through threat modeling is needed before selecting and applying controls.

D) Incorrect. Sharing a raw threat list without prioritization does not help stakeholders make informed decisions. The threat modeling process must first convert the list into concrete, prioritized risks before communication is meaningful.

6 / 40
An AI security engineer reviews the access control design for an AI multi-agent system. Each agent in the system is provisioned with a shared service account. This account has read-and-write access to the production database, the internal file system, and the external API (application programming interface) gateway.

The different AI agents autonomously invoke tools and execute multi-step workflows across internal systems. The provisioning was done on a shared account to simplify credential management across the AI agent pipeline.

Why is this shared account design **particularly risky** in an agentic AI context?

A) Because AI agents generate higher volumes of API calls than human users, so it is more likely that rate-limiting controls are triggered and the shared account gets suspended
B) Because AI agents do not use multi-factor authentication, so the shared credential provides a single point of failure with no other verification layers, which an attacker can exploit
C) Because AI agents operate autonomously across multiple reasoning steps within a single session, so a compromised agent can chain actions across all accessible systems
D) Because regulatory frameworks for AI systems require that each AI agent is provisioned with a unique credential, so shared accounts are a direct compliance violation
E) Because sharing service accounts across multiple AI agents creates too much operational complexity during mandated credential rotation and activities in access management

A) Incorrect. API call volume is a performance and availability concern, not a security risk specific to the shared credential design. Rate limiting affects all account types and does not explain why a shared account is disproportionately risky for agents compared to human users sharing the same account.
B) Incorrect. The absence of multi-factor authentication for service accounts is a general security concern applicable to both human user shared accounts and agent shared accounts. It does not explain why the risk is disproportionately higher in an agentic AI context specifically.
C) Correct. The security risk addressed in the question is related to excessive agency. Agentic AI systems may autonomously chain actions across multiple systems, tools, and reasoning steps within a single session. A compromised agent using a shared account may therefore perform broad unauthorized actions before human intervention or review occurs. (Literature: A, Chapter 0 and 1)
D) Incorrect. Regulatory requirements for AI agent credential provisioning vary by jurisdiction and sector and are not universally established at the level of specificity described. This answer does not address the security risk the question presents.
E) Incorrect. This addresses an operational challenge, not the specific security risk created by the shared account design in an agentic AI context.

7 / 40
A machine learning engineer at a financial institution is investigating two separate incidents involving the institution's fraud detection model.

- **Incident 1**: An external attacker, with no access to the model's architecture, training data, or parameters, submits thousands of crafted transactions. These are iteratively refined based on whether the fraud alert was triggered or not.
- **Incident 2**: A researcher uses a locally trained surrogate model with a similar architecture, to generate adversarial transactions. These successfully evade the production model without prior interaction.

Which two evasion strategies are these?

A) - Incident 1: evasion after poisoning
- Incident 2: perfect-knowledge evasion

B) - Incident 1: partial-knowledge evasion
- Incident 2: transfer attack

C) - Incident 1: perfect-knowledge evasion
- Incident 2: zero-knowledge evasion

D) - Incident 1: transfer attack
- Incident 2: partial-knowledge evasion

E) - Incident 1: zero-knowledge evasion
- Incident 2: transfer attack

A) Incorrect. Incident 1 is incorrectly identified as evasion after poisoning. There is no evidence that the attacker changed the training data. The attacker only tested many transactions and learned from the model's answers. Incident 2 is incorrectly identified as perfect-knowledge evasion. The researcher did not have full access to the real model. They used a surrogate model.

B) Incorrect. Incident 1 is incorrectly identified as partial-knowledge evasion. The attacker had no access to the model's design, data, or parameters. They only used the model's outputs. Incident 2 is correctly identified as transfer attack, but the full answer is still wrong because Incident 1 is not partial-knowledge evasion.

C) Incorrect. Incident 1 is incorrectly identified as perfect-knowledge evasion. This attack does not give the attacker full access to the model. Incident 2 is incorrectly identified as zero-knowledge evasion. The researcher used a surrogate model to build the attack, so this is a transfer attack.

D) Incorrect. Incident 1 is incorrectly identified as transfer attack. In Incident 1, the attacker tested the real model many times and changed the inputs from the answers. Incident 2 is incorrectly identified as partial-knowledge evasion. The researcher used a separate model, so this is a transfer attack.

E) Correct. Incident 1 is correctly identified as zero-knowledge evasion. The attacker had no internal access and used only the model's outputs. Incident 2 is correctly identified as transfer attack. The researcher built the attack on a surrogate model, and it also fooled the real model. (Literature: A, Chapter 2)

8 / 40
Erick is a user of a social media platform which automatically moderates comments using an AI system. He tries to get a policy-violating comment past the platform's automated moderation because he wants to boost engagement on his post. Erick does not have information about the model's architecture, parameters, or training data, and he only sees a binary message when he tries to post his comment, either "blocked" or "posted". By repeatedly making small changes to the wording of his comment, Erick eventually finds a version that gets posted.

What type of evasion input threat is described here?

A) Zero-knowledge evasion
B) Partial-knowledge evasion
C) Perfect-knowledge evasion
D) Evasion after poisoning
E) Transfer attack

A) Correct. Erick has no information about the AI model and uses a query-based attack strategy to find a successful attack input, being able to post his comment. This matches the definition of a zero-knowledge evasion. (Literature: A, Chapter 2)
B) Incorrect. Partial-knowledge evasion would involve some side information guiding the attack, which Erick does not have.
C) Incorrect. Perfect-knowledge evasion would require full access to the AI model's architecture, parameters, and training data to craft inputs precisely.
D) Incorrect. Evasion after poisoning refers to evasion enabled by having previously poisoned the AI model's training data. This is not what happened here.
E) Incorrect. Transfer attack would mean creating adversarial examples using a surrogate model and then applying these adversarial examples to the target model.

9 / 40
A legal services company develops a chatbot that answers user's questions and can retrieve information from internal and external resources, like websites and company documents. During a demo, the chatbot's behavior triggers an investigation by the AI security engineer. She identifies two separate threats:

- **Threat 1:** A message crafted to override the chatbot's instructions and safety rules can cause the chatbot to follow the tester's commands instead of its configured policy.
- **Threat 2:** A vendor webpage retrieved by the chatbot contains embedded text that acts like instructions. When the chatbot ingests that page, it treats the content as commands and deviates from its policy.

What types of threat are these?

A) - Threat 1: evasion
- Threat 2: data poisoning

B) - Threat 1: data poisoning
- Threat 2: direct prompt injection

C) - Threat 1: direct prompt injection
- Threat 2: indirect prompt injection

D) - Threat 1: indirect prompt injection
- Threat 2: evasion

A) Incorrect. Threat 1 is not evasion, as no adversarial input misled the model. Threat 2 is not data poisoning, as the model's training data was not altered.

B) Incorrect. Threat 1 is not data poisoning, as the model's training data was not altered. Threat 2 is not direct prompt injection, as the instructions came indirectly from web content.

C) Correct. Threat 1 is direct prompt injection, where user prompts trick the AI. Threat 2 is indirect prompt injection, where web content contains hidden instructions. (Literature: A, Chapter 2)

D) Incorrect. Threat 1 is not indirect prompt injection, as the instructions came directly from the user. Threat 2 is not evasion, as the model was not tricked by input perturbations.

10 / 40
A consulting company has an AI system in place to assist with day-by-day tasks. After a meeting with a client, the AI assistant summarizes the call. The AI assistant starts to insert a confidential internal links into the draft e-mail to be sent to the client. However, before sending, a self-operating anomaly detector pauses the sending and quarantines the draft.

The AI security architect wants to limit the AI meeting's assistant's access in advance only to the rights of the individuals being served.

Which prompt injection protection layers are illustrated here?

A) Automated oversight and user-based least privilege
B) Just-in-time authorization and human oversight
C) Model alignment and intent-based least privilege
D) Prompt injection input/output handling and rate limit

A) Correct. A self-operating anomaly detector that pauses and quarantines the draft is automated oversight. Granting the assistant only the permissions of the individuals it serves, assigned in advance, is user-based least privilege. (Literature: A, Chapter 2)
B) Incorrect. There is no step-up or per-action approval that justifies just-in-time authorization, and the oversight is automated, not human.
C) Incorrect. The scenario does not mention alignment of the model, and the least-privilege described is based on user identity pre-assigned rights, not dynamic, intent-scoped privileges.
D) Incorrect. No rate limiting is described, and the control shown is a self-operating anomaly detector (automated oversight), not specific I/O labeling measures.

11 / 40
An auditor is assessing the privacy risks of a deployed healthcare large language model (LLM) and its compliance with data protection laws. To do that, the auditor probes the model with targeted queries about a specific patient and carefully matched synthetic non-patients. By observing differences in the model's response patterns and confidence scores, the auditor can reliably determine whether that patient's record was included in the training data, even though no actual medical details are revealed.

What type of attack is simulated here?

A) Data disclosure in model output
B) Indirect prompt injection
C) Membership inference
D) Model inversion

A) Incorrect. Data disclosure in model output occurs when the model directly reveals sensitive information contained in the training data or runtime context. In this scenario, no actual medical details are disclosed.
B) Incorrect. Indirect prompt injection occurs when a third party embeds adversarial instructions in content that is retrieved and inserted into a prompt, manipulating the model's behavior. The scenario describes no such injection.
C) Correct. Membership inference occurs when an attacker presents the model with input identifying a specific individual and uses the model's confidence signals to determine whether that individual appears in the training set. (Literature: A, Chapter 2)
D) Incorrect. Model inversion occurs when an attacker reconstructs part of the training set by intensive experimentation, optimizing inputs to maximize confidence signals in the model output. This type of attack reconstructs sensitive attributes or records from the model, not just whether a record was in the training set.

12 / 40
What is a **main** consequence of a successful model exfiltration attack?

A) The attacker can extract the raw training dataset used to build the original model, potentially including sensitive records.
B) The attacker gains direct control over the deployed model and can modify its weights and parameters without prior authorization.
C) The attacker obtains a replica of the model that can be used to craft evasion attacks without triggering the original's defenses.
D) The original model becomes unavailable to legitimate users due to resource exhaustion, causing service degradation and repeated timeouts.

A) Incorrect. Model exfiltration replicates the model's behavior through input-output pairs, but it does not extract the original training dataset. Extracting training data is a different threat (model inversion or membership inference).
B) Incorrect. Model exfiltration does not grant the ability to modify the original deployed model. It results in a copy of the model's functionality, not control over the original.
C) Correct. One key consequence of model exfiltration is that the attacker can use the replica of the model to perform evasion attacks, circumventing rate limiting, access control, and detection mechanisms present in the original system. (Literature: A, Chapter 2)
D) Incorrect. Denial-of-service (DoS) is a consequence of resource exhaustion, which is a separate threat. Model exfiltration does not directly impact availability of the original model.

13 / 40
A cloud-hosted large language model (LLM) is hit with a sponge attack. Attackers send very long or tricky inputs that force extra computation, driving up costs and causing slowdowns or outages. In order to address this situation, the AI security engineer wants to put two controls in place:

- **Control 1** will validate and sanitize inputs to reject or correct malicious content, such as abnormally large inputs.
- **Control 2** will restrict the computational resources consumed per model input to prevent overuse.

What are these two controls?

A) - Control 1: anomalous input handling
- Control 2: obscure confidence limits

B) - Control 1: denial-of-service (DoS) input validation
- Control 2: limit resources

C) - Control 1: model access control
- Control 2: evasion robust model

D) - Control 1: monitor use
- Control 2: rate limit

A) Incorrect. Anomalous input handling detects inputs that deviate from the training data distribution and is primarily relevant to evasion and model exfiltration threats. Obscure confidence limits output information to reduce the attacker's ability to optimize adversarial samples.

B) Correct. These are the two controls presented in the question and specifically designed to address input-driven resource exhaustion, complementing the general controls. (Literature: A, Chapter 2)

C) Incorrect. Model access control restricts who can interact with the model and is a general control applicable to multiple input threats. Evasion robust model is a development-time control designed to increase the model's resilience against evasion attacks.

D) Incorrect. Monitor use is about observing, correlating, and logging model usage, inputs, outputs, and system behavior to identify events or patterns that may indicate a (cyber)security incident. Rate limit is about limiting the rate (frequency) of access to the model.

14 / 40
An AI security engineer identifies the following threats for a company:

- **Adversarial prompts** are crafted by attackers to cause misclassifications.
- A **compromised ingestion pipeline** introduces backdoored records into the corpus before the model is built.
- Attackers tamper with **labeling workflows** changing ground-truth annotations to bias model behavior.
- **Training data** is obtained from external or third-party sources without proper validation.

According to the OWASP AI Exchange, which threat does **not** involve risks of data poisoning during development-time?

A) Adversarial inputs
B) Compromised ingestion pipeline
C) Labeled workflows
D) Training data

A) Correct. This describes adversarial examples at deployment time, not development-time data poisoning. (Literature: A, Chapter 3)
B) Incorrect. Slipping rigged items into the collected corpus before model building implants a backdoor during development, which configures a development-time poisoning risk.
C) Incorrect. Tampering with labels during annotation poisons the dataset before the model is built, which is a development-time attack.
D) Incorrect. Using unverified external datasets introduces the risk of poisoned samples embedded in data sources. The OWASP AI Exchange highlights supply chain risks and the need for data quality control.

15 / 40
According to the OWASP AI Exchange, what is an example of **direct development-time model poisoning**?

A) Attackers craft adversarial inputs at runtime to mislead the model
B) Attackers manipulate weights and a training pipeline to alter the model's behavior.
C) Sensitive training data is unintentionally exposed through model outputs.
D) Surge of traffic overwhelms the AI service causing it to slow down or have downtime.

A) Incorrect. This is a runtime adversarial example attack, not development-time poisoning.
B) Correct. Direct development-time model poisoning involves compromising the development environment, training pipeline, or model parameters. The attacker may inject logic, modify weights, or alter training procedures to embed malicious behavior. (Literature: A, Chapter 3)
C) Incorrect. This refers to confidentiality threats such as model inversion or data disclosure, not model poisoning.
D) Incorrect. That is a denial-of-service (DoS) issue, not model poisoning.

16 / 40
According to the OWASP AI Exchange, what is an example of **supply-chain model poisoning**?

A) Attackers compromise supply-chain data, models, or components before integration.
B) Attackers manipulate inputs at runtime to trick the model into producing incorrect outputs.
C) The model produces incorrect results due to insufficient data or training quality.
D) The model unintentionally leaks sensitive training data through its responses.

A) Correct. The attack occurs before the model is deployed, via third-party sources such as pre-trained models, external datasets, and Machine Learning (ML) components or pipelines. The OWASP AI Exchange emphasizes the need for supply chain management to ensure trust and integrity. (Literature: A, Chapter 3)
B) Incorrect. This describes evasion or prompt injection attacks, not supply-chain model poisoning. These attacks happen during model use, not during development or acquisition.
C) Incorrect. Incorrect results due to lack of data or training quality are not attacks. Supply-chain poisoning specifically involves intentional malicious tampering.
D) Incorrect. This refers to confidentiality threats like model inversion or sensitive data disclosure, not model poisoning.

17 / 40
An AI development team is conducting a risk assessment of their development environment. They identify two distinct threats:

Threat 1 involves unauthorized access to the datasets used to train and test their model.
Threat 2 involves unauthorized access to the model's parameters and weights stored in a compromised repository.

What types of threat are these?

A) - Threat 1: development-time data leak
- Threat 2: direct development-time model leak
B) - Threat 1: direct augmentation data leak
- Threat 2: repository leak
C) - Threat 1: direct runtime data leak
- Threat 2: direct runtime model leak
D) - Threat 1: input data leak
- Threat 2: source code/configuration leak

A) Correct. Development-time data leak involves unauthorized access to training and test data, breaching the confidentiality of sensitive data or intellectual property contained within those datasets. Direct development-time model leak involves unauthorized access to model attributes (parameters, weights, architecture), breaching confidentiality of model intellectual property. (Literature: A, Chapter 3)
B) Incorrect. Direct augmentation data leak does not match training and testing datasets, and repository leak is too vague. The issue is specifically a model artifact leak, not merely a generic repository issue.
C) Incorrect. The threats described are not runtime leaks. They occur in the development environment.
D) Incorrect. Input data leak generally refers to leakage of user-provided inputs, not training and test datasets. Model weights are not source code/configuration, but model artifacts.

18 / 40
A security team reviews the runtime threat landscape of an AI system. The AI system is integrated into a web application. The team identifies two conventional security threats:

- **Threat 1:** An attacker can gain access to the production server's memory and extracts the model's weights and parameters.
- **Threat 2:** an attacker with access to the model registry replaces the deployed model artifact with a trojaned version.

What consequences do these conventional security threats have that are **specific** to AI systems?

A) - Threat 1 causes the system to auto-retrain on malicious data.
- Threat 2 forces the model to reject valid prompts.

B) - Threat 1 changes the model's architecture at runtime.
- Threat 2 silently compresses weights to lower precision.

C) - Threat 1 disables safety filters making the system more permissive by design.
- Threat 2 makes outputs deterministic and predictable.

D) - Threat 1 enables model theft and potential training data inference.
- Threat 2 allows undetected model tampering.

A) Incorrect. Threat 1 is not about poisoning via auto-retraining, but about model theft and data inference. Threat 2 is about integrity tampering and backdoors, not rejecting prompts.

B) Incorrect. Runtime architecture changes and silent precision compression are not AI-specific consequences of the threats. They describe implausible or non-malicious behavior instead of theft or covert tampering.

C) Incorrect. Disabling safety filters maps to tampering, not to the exfiltration risk of threat 1. Making outputs deterministic misses the integrity nature of threat 2.

D) Correct. Threat 1 leads to model theft and potential training data inference. Threat 2 leads to covert integrity attacks (undetected model tampering). (Literature: A, Chapter 4)

19 / 40
A financial services company deploys a machine learning (ML) model to detect fraudulent transactions. The ML model is hosted in their production environment and continuously processes live transaction data. Two security incidents occur:

- **Incident 1**: An attacker gains unauthorized access to the deployed model environment and alters the model's parameters so that certain fraudulent transactions are no longer flagged. The manipulation is not detected, and the attacker benefits from this.
- **Incident 2**: An attacker exploits a vulnerability in the production system. This allows the attacker to replicate the model and analyze its behavior for further attacks.

Based on the OWASP AI Exchange, what type of security incidents are described?

A) - Incident A: data poisoning
- Incident B: direct runtime model poisoning

B) - Incident A: data poisoning
- Incident B: evasion attack

C) - Incident A: direct runtime model poisoning
- Incident B: direct runtime model leak

D) - Incident A: direct runtime model leak
- Incident B: evasion attack

A) Incorrect. Direct runtime model leak correctly identifies security incident B. The attacker extracts model parameters and architecture, compromising model confidentiality. This enables replication and further exploitation. However, data poisoning occurs during development time via training data manipulation. The scenario describes runtime compromise, not training-phase tampering.

B) Incorrect. Data poisoning occurs during development time via training data manipulation. The scenario describes runtime compromise, not training-phase tampering. An evasion attack involves manipulating inputs, not modifying the model itself. In the scenario, the model is directly altered.

C) Correct. Direct runtime model poisoning correctly identifies security incident A. The attacker modifies the deployed model, altering its behavior. This is a model integrity attack known as runtime poisoning (reprogramming). Direct runtime model leak correctly identifies security incident B. The attacker extracts model parameters and architecture, compromising model confidentiality. (Literature: A, Chapter 4)

D) Incorrect. Direct runtime model leak correctly identifies security incident B. The attacker extracts model parameters and architecture, compromising model confidentiality. This enables replication and further exploitation. However, an evasion attack involves manipulating inputs, not modifying the model itself. In the scenario, the model is directly altered. (Literature: A, Chapter 4)

20 / 40
A healthcare organization deploys an AI assistant to help clinicians summarize patient records and provide recommendations. The system integrates with internal databases and external tools, and clinicians rely on its generated outputs during daily operations. Two threats are observed:

- **Threat 1**: The AI assistant generates a response that includes a hidden malicious script embedded in HTML content. When the output is viewed, the script executes and sends session data to the attacker.
- **Threat 2**: An attacker exploits weak access controls in the system to intercept and retrieve sensitive patient data.

Based on the OWASP AI Exchange, what types of threat are described?

A) - Threat 1: data poisoning
- Threat 2: input data leak

B) - Threat 1: data poisoning
- Threat 2: model inversion attack

C) - Threat 1: output containing conventional injection
- Threat 2: input data leak

D) - Threat 1: output containing conventional injection
- Threat 2: model inversion attack

A) Incorrect. Threat 2 is correctly identified as an input data leak. Sensitive data provided to the model is exposed or intercepted, violating confidentiality. However, data poisoning occurs during development-time by altering training data. The scenario describes a runtime conventional injection, not training manipulation.

B) Incorrect. Data poisoning occurs during development-time by altering training data. The scenario describes a runtime conventional injection, not training manipulation. Model inversion involves inferring training data from outputs. The scenario does not describe a situation whereby model inference has taken place.

C) Correct. Threat 1 is correctly identified as output containing conventional injection. The AI output includes malicious executable content, in this case script injection. Threat 2 is correctly identified as an input data leak. Sensitive data provided to the model is exposed or intercepted, violating confidentiality. (Literature: A, Chapter 4)

D) Incorrect. Threat 1 is correctly identified as output containing conventional injection. The AI output includes malicious executable content, in this case script injection. However, model inversion involves inferring training data from outputs. The scenario does not describe a situation whereby model inference has taken place.

21 / 40
A company uses a retrieval-augmented generation (RAG) system to answer employee questions about internal policies. The AI system retrieves documents from a vector database and adds the documents to the prompt before sending them to the model.

A security review finds that an attacker gained unauthorized access to the vector database. The attacker viewed stored augmentation documents and replaced a legitimate policy document with another one, containing false information. Employees who ask the AI system about that policy now receive incorrect guidance.

Which security threats does this scenario illustrate?

A) - Threat 1: augmentation data leak
- Threat 2: output containing conventional injection

B) - Threat 1: augmentation data manipulation
- Threat 2: augmentation data leak

C) - Threat 1: membership inference
- Threat 2: augmentation data manipulation

D) - Threat 1: output containing conventional injection
- Threat 2: membership inference

A) Incorrect. Output containing conventional injection involves malicious executable content embedded in model output, which is not present in this scenario. Although augmentation data leak is correctly identified, the complete answer is incorrect.

B) Correct. Augmentation data manipulation occurs when augmentation data stored in a vector database is modified by an attacker, inserting false information that changes the behavior of the model. Augmentation data leak occurs because an attacker gained unauthorized access to augmentation data stored in the vector database. The vector database represents an additional location where augmentation data can be accessed or modified, creating a new attack surface and expanding exposure risks beyond the original document repository. (Literature: A, Chapter 4)

C) Incorrect. Membership inference is an attack where an adversary determines whether specific data appears in the training set by analyzing confidence signals. This is not described in the scenario. While augmentation data manipulation is correctly identified, the pairing with membership inference is incorrect.

D) Incorrect. Output containing conventional injection refers to malicious code embedded in model output. Membership inference is an attack where an adversary determines whether specific data appears in the training set by analyzing confidence signals. Neither of these security threats are described in the scenario.

22 / 40
General governance controls should be implemented to ensure effective security oversight of AI systems.

Which approach demonstrates this **best**?

A) Defining clear policies, roles, and risk management practices to guide the secure development, deployment, and monitoring of AI systems
B) Including security controls in the AI systems to apply security-by-design principles in the design stage of the AI system
C) Meeting minimum security requirements as AI systems are more secure by default and do not require multiple security layers
D) Separating governance controls for AI from information security controls to make sure these two different concepts are equally covered

A) Correct. Governance controls are about defining and enforcing policies, roles, accountability, and risk management within an Information Security Management System (ISMS), ensuring end-to-end oversight across the AI lifecycle. (Literature: A, Chapter 1)
B) Incorrect. The application of security by design principles in the design stage of AI systems is an important part of the general governance controls, but it is not sufficient. Governance involves policies, ownership, risk, monitoring, and continuous improvement.
C) Incorrect. AI systems are not secure by default. Relying on minimum requirements ignores defense-in-depth and continuous oversight needed for evolving AI risks.
D) Incorrect. Separating AI governance from information security creates silos and gaps. A general best practice is integration with enterprise risk and Information Security Management System (ISMS) frameworks to maintain coherent controls and accountability.

23 / 40
A mid-size marketing agency has little AI experience and no formal AI governance yet. They want to use a generative AI tool to draft customer e-mails. The AI toll needs access to customer names and e-mails. The company should establish AI governance and establish security oversight before they can use the AI tool.

What is the **bare minimum** the company should do for security oversight?

A) Conduct red team testing of the generative AI e-mail system to mitigate security risks before launch
B) Hire an AI governance specialist to do a threat assessment and define policies and controls for deployment
C) Inventory the planned AI use and perform a risk analysis to identify threats, controls, and responsibilities
D) Request legal advice from a company specialized in AI and the related AI Act requirements
E) Start with a pilot and monitor the AI e-mail system for performance, misuses, and incidents with logs

A) Incorrect. Red teaming is not a bare minimum governance step. Red teaming is a later-stage assurance activity that depends on a defined scope, baseline controls, and ownership established by initial governance.
B) Incorrect. Hiring an AI governance specialist is a resourcing decision, not a minimum governance step.
C) Correct. This option represents the bare minimum governance step for AI security oversight by identifying risks, threats, controls, and responsibilities before deployment. (Literature: A, Chapter 1)
D) Incorrect. Legal review is important, but it is part of compliance and cannot replace security governance. Delaying governance leaves personal data and misuse risks unmanaged.
E) Incorrect. Piloting and monitoring are operational and reactive, and piloting on a small customer segment is an implementation strategy. Without prior governance, piloting and monitoring lack defined guardrails, roles, and controls to protect CRM data.

24 / 40
How should the coverage of general governance controls in AI security be **best** understood?

A) As a collection of encryption methods used to secure AI outputs after generation
B) As a set of overarching controls that apply across all AI threats and lifecycle stages
C) As controls limited to the model training phase, focusing on dataset quality and tuning techniques
D) As optional measures that organizations may adopt if the AI systems do not process sensitive data
E) As technical safeguards applied exclusively at runtime to detect malicious inputs

A) Incorrect. Encryption is just one possible technical control and does not represent governance coverage. General controls are not limited to cryptography or outputs.
B) Correct. Governance is cross-cutting: policies, roles, accountability, and risk management that apply across threats and all lifecycle stages. (Literature: A, Chapter 1)
C) Incorrect. Governance applies to the entire lifecycle, not just the training phase.
D) Incorrect. Governance controls are not optional and are not limited to sensitive data processing.
E) Incorrect. General controls are not limited to runtime or technical defenses. They span organizational, procedural, and governance layers.

25 / 40
A company in healthcare decides to use a ready-made AI model, provided by a third-party, for an application. The application processes user input and retrieves external data.

The AI security engineer is worried about the model's security. The security strategy that defines the division of security responsibilities between the AI model provider and the healthcare company is not defined.

What is **correct** about these responsibilities?

A) Most security controls can be delegated to the provider when using ready-made AI models, since the provider's training and hosting cover security risks.
B) The deployer is responsible for all security controls when they integrate the AI model into their systems and consider security before deployment.
C) The provider has trained the AI model, so they manage model-level controls, while the deployer should manage application-level controls.
D) The provider is responsible for all security controls because it trained the AI model and hosts it, so all risks and attacks should be handled by the provider.

A) Incorrect. Prompt injection and data leakage in the context of external tools or retrieval-augmented generation (RAG) are primarily application-layer risks and cannot be fully delegated to the provider. Hosting and training alone do not address these risks.
B) Incorrect. The provider trains and finetunes the model and defines base safety controls. Therefore, they are responsible for part of security, which means there is a shared responsibility.
C) Correct. The provider trains and finetunes the model, owning model-level controls. The deployer must still be aware of threats when using ready-made models, owning application-level controls. (Literature: A, Chapter 0)
D) Incorrect. Although the provider trains and hosts the model, they are not responsible for all the security controls. They are responsible for part of security, which means there is a shared responsibility.

26 / 40
An organization has deployed a customer-facing AI application using a large language model (LLM) supplied by a third-party provider. The model is accessed through a managed API (application programming interface) service and has not been further trained or modified by the organization.

The security team discovers that carefully crafted user inputs can cause the model to ignore its configured behavioral boundaries and produce outputs that violate organizational content restrictions. The team considers which control to implement that falls within the organization's responsibility as a deployer rather than the third-party provider's.

Which control should the security team implement to address the issue?

A) Insert an output validation layer that scans all model responses before they reach the end users and block outputs that violate the organization's content policy
B) Modify the provider’s inference pipeline to enforce stricter generation constraints that limit response generation and prevent disallowed content from emerging
C) Request that the provider retrain and safety-tune the supplied model using additional refusal and policy-alignment data to improve resistance against jailbreak inputs
D) Require the provider to revise and apply strengthened, service-wide content moderation rules and enforcement so that unsafe outputs are filtered globally before delivery

A) Correct. Deployers are responsible for controls implemented around the AI application and its outputs, while third-party providers manage the underlying model training and inference behavior. An output validation layer is a deployer-side control that can block policy-violating responses before they reach end users. (Literature: A, Chapter 0)
B) Incorrect. Modifying the provider’s inference pipeline is not within the deployer’s control when using a third-party managed API. At best, they can adjust exposed parameters, not change the pipeline.
C) Incorrect. Retraining or safety-tuning the model is the provider’s responsibility. The deployer can request it but cannot implement it themselves for a managed API model.
D) Incorrect. Global moderation policies and provider-wide filtering controls are managed by the third-party provider, not by the deployer organization.

27 / 40
A team is preparing training data for a generative AI assistant. The team finds that they do not need all the fields and data they currently use for the AI model's performance. As a next step, they are considering controls to limit sensitive data and improve confidentiality and integrity.

What is the **best** way to improve confidentiality and integrity?

A) Keep all the fields in the training data and encrypt the data at rest and in backups
B) Preserve identifiers in training data to simplify future retraining tasks and processes
C) Remove fields and data that are unnecessary for model training and performance
D) Use only synthetic data for all generative AI training and evaluation phases

A) Incorrect. Encryption at rest protects stored data, but it does not reduce unnecessary sensitive data exposure or minimize the attack surface. It improves confidentiality only at rest and does little for integrity.
B) Incorrect. Keeping identifiers in training data increases exposure risk. It undermines confidentiality and can harm integrity.
C) Correct. Data minimization removes unnecessary fields and records, reducing the data attack surface. It strengthens integrity by giving attackers fewer chances to tamper with the data and by stopping the model from relying on irrelevant patterns. (Literature: A, Chapter 1)
D) Incorrect. Using synthetic data is not always feasible or required and it does not guarantee confidentiality and integrity improvements. Synthetic data may still preserve sensitive patterns or introduce other risks.

28 / 40
A product team is training an AI model for customer support. The current training set includes full customer addresses, phone numbers, and legacy records that are not needed to meet performance goals but are likely useful in the future. After a red team exercise highlights a risk of data leakage and manipulation via memorized details, the team decides to reduce exposure. They do this by removing unnecessary fields and entire records so that sensitive information is not present in the dataset.

Which **general** control was implemented to limit sensitive data in this situation?

A) Allowed data
B) Data minimization
C) Obfuscate training data
D) Short retain

A) Incorrect. Allowed data defines what categories are permitted, but the scenario requires deleting unnecessary records, which is data minimization.
B) Correct. Data minimization is about removing unneeded records from the training set so that sensitive data is not present and therefore cannot be leaked. (Literature: A, Chapter 1)
C) Incorrect. Obfuscate training data transforms data but still retains it. Therefore, risks can remain, unlike removal.
D) Incorrect. Short retain limits how long data is kept, not which records are included in the training set.

29 / 40
A retail company decides to stop storing customer payment details after each transaction is completed.

How does this action **reduce** the risk of data exposure?

A) By eliminating the need for encryption of data in transit or archived for auditability purposes
B) By ensuring compliance with all relevant data protection regulations, reducing complexity
C) By guaranteeing that no unauthorized user can access the company network and database
D) By preventing retained data from being leaked, reconstructed or inferred from the system

A) Incorrect. Data minimization does not remove the need for encryption. Data that is transmitted still needs to be protected, regardless of how much data is stored.
B) Incorrect. Data minimization supports compliance efforts, but it does not automatically ensure full compliance with all data protection regulations, which have many additional requirements.
C) Incorrect. Limiting stored data does not guarantee that unauthorized users cannot access the network or database. Network access control is a separate security measure unrelated to data minimization.
D) Correct. When data is not collected or retained, it cannot be leaked, reconstructed, or inferred from the system. This directly lowers the impact of dataset theft or unauthorized access. (Literature: A, Chapter 1)

30 / 40
A regional hospital network uses an AI agent to read inbound vendor e-mails, draft replies, and automatically send responses. During a security testing following a recent phishing incident, testers found that prompt injection could make the AI agent ignore its rules and send unauthorized responses.

Which design decision works **best** to limit the effects of this unwanted behavior?

A) Disable all AI agent integrations and revert to fully manual human workflows for most processes
B) Expand user training for prompt safety, and emphasize oversight during agent-assisted e-mailing
C) Keep broad autonomous send permissions, and upgrade to a more advanced generative model
D) Restrict the AI agent permissions to task-specific scopes, and require approval for high-risk actions

A) Incorrect. This removes automation rather than implementing controls to safely retain functionality.
B) Incorrect. User awareness helps but does not enforce AI runtime controls.
C) Incorrect. Improving model quality does not sufficiently reduce the blast radius if the generative AI system still has broad permissions.
D) Correct. Least-privilege and human-in-the-loop controls are recommended to limit impact if prompt injection occurs. Task-specific least privilege limits what the agent can do, while approval of high-risk actions adds oversight, preventing unauthorized responses even when the model is fooled. (Literature: A, Chapter 1)

31 / 40
What is the **best** way to minimize the effects of unwanted model behavior?

A) Emphasizing prompt hardening, role-based access, and scheduled audits of model outputs
B) Relying on automated output filters, exception logging, and occasional fairness spot-checks
C) Testing for unwanted bias and applying controls like oversight, least privilege, and continuous validation
D) Using strong input validation, constrained tool permissions, and periodic human-in-the-loop reviews

A) Incorrect. These are useful but incomplete. Prompt hardening is brittle, role-based access control (RBAC) without a broader least-privilege program is limited, and scheduled audits are periodic rather than continuous, leaving gaps and lacking systematic bias testing.
B) Incorrect. Automated output filters can be bypassed, exception logging does not prevent harm, and occasional fairness spot-checks are insufficient compared to continuous validation and structured bias testing.
C) Correct. Controls to limit the effects of unwanted model behavior are: oversight, least model privilege, model alignment, AI transparency, explainability, continuous validation, and unwanted bias testing. (Literature: A, Chapter 1)
D) Incorrect. Strong input validation, constrained tool permissions, and periodic human review help, but this list omits continuous validation and dedicated bias testing. Periodic reviews leave gaps and do not provide comprehensive, ongoing risk reduction.

32 / 40
Limiting unwanted model behavior is a critical strategy for impact limitation and blast radius control.

How does this strategy **improve** performance and **reduce** risk?

A) By focusing on attacks that can affect AI systems and limiting oversight over system performance factors and accuracy
B) By implementing controls that remove the need for continuous testing and validation over the time to reduce complexity
C) By improving operational reliability and resource efficiency while minimizing incidents and wasted compute requests
D) By increasing model autonomy and allowing it to generate more outputs to continuously improve system performance

A) Incorrect. AI transparency is one of the controls for limiting unwanted behavior. It is directly relevant to accuracy because attacks or drifts that cause unwanted behavior are accuracy problems.
B) Incorrect. Regular testing and performance validation ensure that the model remains consistent with its intended use over time and limits the effect of unwanted behavior. Unwanted bias testing is one of the controls to limit unwanted behavior in AI systems. Continuous testing and validation over the time are highly recommended.
C) Correct. Limiting unwanted behaviors reduces harmful outputs and operational incidents, improves reliability, and increases resource efficiency by reducing retries, unnecessary tool calls, and wasted compute resources. (Literature: A, Chapter 1)
D) Incorrect. Increasing the model autonomy does not lead to improved performance. The model's autonomy level should be under control so that unwanted model behavior can be limited.

33 / 40
Within the broader scope of AI red teaming, what is the **primary** purpose of AI security testing?

A) Assessing whether an AI model is resilient to specific attacks by reproducing those attacks in a controlled environment
B) Evaluating whether an AI model produces accurate and traceable outputs when exposed to expected input data
C) Measuring whether an AI model demonstrates the expected computational efficiency when processing large volumes of requests
D) Verifying whether an AI model meets its defined business requirements through standard quality assurance procedures

A) Correct. AI security testing is the part of AI red teaming that tests if the AI model can withstand certain attacks by simulating these attacks. Its primary purpose is therefore to verify the model's resilience through controlled attack reproduction. (Literature: A, Chapter 5)
B) Incorrect. Assessing output accuracy under normal input conditions describes functional or performance testing, not AI security testing, which focuses on attack simulation.
C) Incorrect. Computational efficiency under routine workloads is a concern of performance or load testing, not the attack-focused scope of AI security testing.
D) Incorrect. Checking compliance with business requirements through quality assurance is outside the scope of AI security testing, which specifically targets the model's ability to withstand attacks.

34 / 40
A company deploys a generative AI assistant and wants to define AI security tests that cover more than just conventional pen testing. To do that, the testing team starts identifying the threats that should be tested for.

Which threats **specifically** associated with generative AI systems should be included in the AI security testing?

A) Evasion attacks, insecure output handling, and model poisoning
B) Model exfiltration, prompt injection, and evasion attacks
C) Prompt injection, sensitive data disclosure, and insecure output handling
D) Sensitive data output from model, model drift, and unsafe tool invocation

A) Incorrect. Evasion attacks are more commonly associated with predictive AI systems, while insecure output handling can affect generative AI systems. This combination does not best represent the core generative AI-specific threats that should be included in the AI security testing.
B) Incorrect. Although these are valid AI security threats, evasion attacks are more commonly associated with predictive AI systems than with generative AI assistants.
C) Correct. These three threats affect generative AI and, therefore, should be tested for. (Literature: A, Chapter 5)
D) Incorrect. Model drift is a performance concern, not a security threat.

35 / 40
An AI security team is testing a generative AI chatbot for prompt injection resistance. During the test, the system has blocked several crafted attack inputs.

What is the recommended **next** step to maintain security oversight of the generative AI chatbot?

A) Add input variation such as synonym replacement, encoding changes, or formatting shifts to attempt to circumvent detection
B) Conclude the current testing cycle and deem the defenses adequate since most prompt injection attempts were blocked
C) Run the same test several times to increase sample size and obtain more statistically relevant defense performance data
D) Send the prompt injection attack inputs directly to the model to better train it against similar threats in future deployments

A) Correct. Adding variation algorithms such as synonyms, encoding, and formatting to try to bypass detection adds detection robustness. Expanding input diversity improves coverage, exposes bypasses, and maintains security oversight. (Literature: A, Chapter 5)
B) Incorrect. The test should not be concluded yet. The system must still be resilient against deliberate bypass attempts.
C) Incorrect. Rerunning the same tests without introducing variation is unlikely to expose new bypass techniques or improve coverage.
D) Incorrect. Sending raw attack inputs directly to the model bypasses production filters and detections mechanisms. Additionally, feeding attack inputs directly into training is a model improvement activity, not the next step in testing.

36 / 40
The AI system of a company relies on large amounts of personal data for training and decision-making.

From an ethical perspective, which risk is **most** important for the company to address?

A) Data being collected, retained, or reused beyond justified purposes
B) Interface complexity making automated decisions difficult to review
C) Model accuracy dropping when resources are unevenly allocated
D) Software updates introducing compatibility issues across platforms

A) Correct. AI systems are data-intensive and raise concerns about how personal data is collected, retained, and reused. Privacy principles and legislation often require a legal basis or consent for such use, along with safeguards such as use limitation and data minimization. (Literature: A, Chapter 6)
B) Incorrect. Interface complexity is mainly a usability concern. It does not directly address the risks that arise from AI systems relying on large amounts of personal data.
C) Incorrect. Hardware resource allocation concerns technical performance and model operation. It is not the main concern associated with the use of personal data in AI systems.
D) Incorrect. Software update and compatibility issues concern system maintenance and interoperability. They are not the main concern that arises from AI systems using large amounts of personal data.

37 / 40
An online bookshop uses an AI system to analyze browsing history, purchase trends, and preorders so it can predict demand and make sure popular books are in stock on time. Later, the company reuses that same customer data to train a separate AI model that builds detailed marketing profiles and sends targeted ads to individual shoppers.

Which privacy principle is **most directly** violated here?

A) Data minimization
B) Data retention limitation
C) Transparency and consent
D) Use limitation and purpose specification

A) Incorrect. Data minimization refers to collecting only the data that is strictly necessary for a defined purpose, not to the restriction on repurposing already collected data for a different goal.
B) Incorrect. Data retention limitation concerns how long data is stored before it must be deleted, rather than whether data collected for one purpose can be reused for a different purpose.
C) Incorrect. Transparency and consent relate to informing individuals about data collection and obtaining their agreement, not specifically to the restriction on using data collected for one purpose to train a model for a different purpose.
D) Correct. The scenario illustrates a violation of use limitation and purpose specification, because data collected for safety and security should not be repurposed as a training dataset for profiling or personalized marketing. (Literature: A, Chapter 6)

38 / 40
An organization wants to set up a framework to govern the responsible use of AI across its operations.

They use the ISO/IEC 27001 standard for information security management. The security team wants to set up their AI governance with the ISO/IEC standard that serves the same structural purpose for AI governance as the ISO/IEC 27001 standard does for information security.

Which standard is this?

A) ISO/IEC 23894
B) ISO/IEC 27005
C) ISO/IEC 42001
D) ISO/IEC 5338

A) Incorrect. ISO/IEC 23894 focuses on AI risk management guidance, not on providing a management system for the governance of responsible AI.
B) Incorrect. ISO/IEC 27005 addresses information security risk management and is not designed to serve as a governance management system for responsible AI.
C) Correct. ISO/IEC 42001 can be seen as a management system standard for the governance of responsible AI in an organization, similar to how ISO/IEC 27001 is a management system standard for information security. (Literature: A, Chapter 1)
D) Incorrect. ISO/IEC 5338 focuses on extending software lifecycle practices and engineering, not on providing a governance management system for responsible AI.

39 / 40
A company wants to use an AI system to automatically screen and rank job applicants, based on their résumés.

According to the AI Act, in which category should the use of this AI system be classified?

A) Unacceptable risk
B) High risk
C) Limited risk
D) Minimal or no risk

A) Incorrect. The AI Act does not impose a ban on AI-based hiring tools. It classifies them as high-risk, which means they are subject to specific requirements and restrictions, but they are not outright prohibited.
B) Correct. Employment screening may affect access to work and fundamental rights and must be categorized as high risk. Organizations must ensure compliance before deploying AI systems for such purposes. (Literature: A, Chapter 6)
C) Incorrect. Limited risk covers primarily transparency-only obligations. Employment screening may affect access to work and fundamental rights and must be categorized as high risk.
D) Incorrect. Minimal risk is for non-consequential uses, like spam filters or video games. Employment screening may affect access to work and fundamental rights and must be categorized as high risk.

40 / 40
An AI development team is preparing training data for a new machine learning model. The team lead wants to ensure the data sourcing approach supports copyright compliance.

What is the **most ethical** data sourcing practice?

A) Collecting data from publicly available domains only
B) Creating all training data within the organization
C) Downloading data from sources with implied permission
D) Using only data that appears widely reposted online

A) Incorrect. Public availability alone does not guarantee that data is licensed for the organization's intended use. License terms must be reviewed to confirm they are sufficient for the planned application.
B) Correct. Producing data in-house is one of the recognized ethical data sourcing strategies, as it ensures the organization has full rights over the data used to train AI models. (Literature: A, Chapter 0)
C) Incorrect. Implied permission is not a valid basis for data use. Explicit permissions must be obtained from the data owner before using third-party data for AI training.
D) Incorrect. Content that is widely reposted online may still be protected by copyright. Repetition or broad circulation does not make material free to reuse for AI training. The focus must be on whether the data is created in-house, properly licensed, or obtained with the necessary permissions.

Evaluation

The table below shows the correct answers to the questions in this sample exam.

Question	Answer	Question	Answer
1	A	**21**	B
2	A	**22**	A
3	A	**23**	C
4	D	**24**	B
5	A	**25**	C
6	C	**26**	A
7	E	**27**	C
8	A	**28**	B
9	C	**29**	D
10	A	**30**	D
11	C	**31**	C
12	C	**32**	C
13	B	**33**	A
14	A	**34**	C
15	B	**35**	A
16	A	**36**	A
17	A	**37**	D
18	D	**38**	C
19	C	**39**	B
20	C	**40**	B

Certified for what's next

Contact EXIN

www.exin.com