Sample Exam – Answers

Sample Exam set A
Version 2.0

ISTQB® AI Testing Syllabus

Compatible with Syllabus version 2.0

International Software Testing Qualifications Board

Version 2.0 Released 2026/04/17

Copyright Notice

Document Responsibility

The ISTQB® Examination Working Group is responsible for this document.

This document is maintained by a core team from ISTQB® consisting of the Syllabus Working Group and Exam Working Group.

Acknowledgements

This document was produced by a core team from the ISTQB®: Klaudia Dussa-Zieger, Stuart Reid, Vipul Koch, Kyle Siemens, Qin Liu, Werner Henschelchen, Jarosław Hryszko

The core team thanks the Exam Working Group review team, the Syllabus Working Group and Member Boards for their suggestions and input.

Revision History

Sample Exam – Answers Layout Template used:	Version 2.12 Date: November 27, 2024

Version	Date	Remarks
1.0	2021/10/01	Release for GA
2.0 Beta	2026/01/05	Beta Review
2.0	2026/04/17	Release for GA

Table of Contents

Introduction

Purpose of this document

The example questions and answers and associated justifications in this sample exam have been created by a team of subject matter experts and experienced question writers with the aim of:

- Assisting ISTQB® Member Boards and Exam Boards in their question writing activities
- Providing training providers and exam candidates with examples of exam questions

These questions cannot be used as-is in any official examination.

Note, that real exams may include a wide variety of questions, and this sample exam ***is not*** intended to include examples of all possible question types, styles or lengths, also this sample exam may both be more difficult or less difficult than any official exam.

Instructions

In this document you may find:

- Answer Key table, including for each correct answer:
  - K-level, Learning Objective, and Point value
- Answer sets, including for all questions:
  - Correct answer
  - Justification for each response (answer) option
  - K-level, Learning Objective, and Point value
- Additional answer sets, including for all questions [does not apply to all sample exams]:
  - Correct answer
  - Justification for each response (answer) option
  - K-level, Learning Objective, and Point value

- *Questions are contained in a separate document*

Answer Key

Question Number (#)	Correct Answer	LO	K-Level	Points
1	a	AI-1.1.1	K2	1
2	c, e	AI-1.1.2	K2	1
3	a	AI-1.1.3	K2	1
4	b	AI-1.1.4	K2	1
5	a	AI-1.1.5	K2	1
6	b	AI-1.1.6	K2	1
7	c	AI-2.1.1	K2	1
8	b	AI-2.1.2	K2	1
9	c	AI-2.2.1	K2	1
10	a	AI-3.1.1	K2	1
11	d	AI-3.1.2	K2	1
12	d	AI-3.1.4	K2	1
13	c	AI-3.2.1	K2	1
14	b	AI-3.2.3	K2	1
15	c	AI-3.3.1	K3	2
16	a	AI-3.4.1	K2	1
17	a	AI-4.1.1	K2	1
18	b	AI-4.1.2	K2	1
19	b	AI-4.1.3	K2	1
20	b	AI-4.2.1	K2	1

Question Number (#)	Correct Answer	LO	K-Level	Points
21	d	AI-4.2.2	K3	2
22	b	AI-4.3.1	K2	1
23	a	AI-4.3.2	K2	1
24	a	AI-5.1.1	K2	1
25	c	AI-5.1.2	K2	1
26	d	AI-5.1.3	K2	1
27	b	AI-5.1.4	K2	1
28	c	AI-5.1.5	K3	2
29	b	AI-5.1.6	K2	1
30	b	AI-6.1.1	K2	1
31	a	AI-6.1.2	K2	1
32	a	AI-6.1.3	K2	1
33	c	AI-6.1.4	K2	1
34	b	AI-6.1.5	K3	2
35	b	AI-6.1.7	K2	1
36	c	AI-6.1.8	K2	1
37	c	AI-6.1.9	K2	1
38	d	AI-6.1.10	K2	1
39	b, d	AI-7.1.1	K2	1
40	b	AI-7.1.2	K2	1

Answers

Question Number (#)	Correct Answer	Explanation / Rationale	Learning Objective (LO)	K-Level	Number of Points
1	a	a) Is correct. This correctly differentiates AI-based systems from conventional systems. AI-based/ML systems learn patterns in the data, and many can adapt through self-learning, while conventional systems use predefined logic and produce predictable outputs. Note: this does not apply to rule-based expert systems. b) Is not correct. While AI-based systems can sometimes process data faster, this is not always the case, and their advantage lies more in adaptability and handling complexity than in speed. c) Is not correct. This reverses the characteristics of the two types of systems. AI-based systems are probabilistic and less explainable, while conventional systems are deterministic and more straightforward to interpret. d) Is not correct. Conventional systems are more suitable for critical tasks where being able to explain results is important, AI-based systems typically have poor transparency and explainability and so are more appropriate for use in unregulated areas. Also, AI-based systems are often more appropriate for more complex problems.	AI-1.1.1	K2	1
2	c, e	a) Is not correct. Narrow AI is not self-learning in all cases; it is task specific. General AI is not limited to specialized problems, and super AI is not limited to pre-defined tasks. b) Is not correct. Narrow AI does not operate independently of human input, general AI is not exclusive to robotics, and super AI does not merely enhance human decision-making but hypothetically surpasses human intelligence entirely.	AI-1.1.2	K2	1

		c) Is correct. This accurately differentiates the three types of AI, highlighting narrow AI's task-specific nature, general AI's human-like versatility, and super AI's hypothetical nature and superior capabilities. d) Is not correct. While narrow AI is correctly identified as task specific, general AI, if achieved, would have many real-world applications. Additionally, it is unclear what relationship super AI would have with generative AI models. e) Is correct. All AI today is narrow AI. When general AI will be achieved is not clear. Once we have general AI then given the availability of knowledge on the internet, it is almost certain that super AI will be achieved.			
3	a	a) Is correct. AI is the broadest term. It refers to the general concept of creating machines that can perform tasks that typically require human intelligence. This includes a wide range of approaches and techniques, not just machine learning. ML is a subset of AI. It's a specific approach to achieving AI where machines are given the ability to learn from data without being explicitly programmed. Instead of writing specific rules for every situation, ML algorithms learn patterns and make predictions based on the data. DL is a subset of machine learning. It's a specialized type of ML that utilizes artificial neural networks with multiple layers (hence "deep"). b) Is not correct. AI and ML existed as concepts and fields of study long before deep learning became prominent. Deep learning is a relatively recent advancement within the broader fields of machine learning and AI. c) Is not correct. DL and ML are not interchangeable. DL is a specific type of ML. Also, AI doesn't represent a "separate approach" - ML and DL are approaches to implementing AI.	AI-1.1.3	K2	1

		d) Is not correct. While it's true AI encompasses ML and DL, "working in parallel" is not a valid way to describe their relationship. DL is a type of ML, not a parallel methodology at the same level that can run in parallel.			
4	b	a) Is not correct. It describes analysis and understanding, which are not capabilities of GenAI, and does not focus on the generative aspect of generative AI. b) Is correct. This accurately describes generative AI, highlighting its ability to create new content based on learned patterns in training data. c) Is not correct. Generative AI focuses on creating new content rather than improving existing content as a precursor to more traditional AI tasks, such as classification and prediction tasks. d) Is not correct. Generative AI is not limited to text and images; it has applications in healthcare, drug discovery, data simulation, and other fields.	AI-1.1.4	K2	1
5	a	a) Is correct. This comparison accurately highlights the advantages of GPUs over CPUs for machine learning tasks, due to their parallel processing capabilities, while CPUs excel in general-purpose computing. b) Is not correct. While CPUs have faster clock speeds, they are less efficient than GPUs for ML tasks because they lack the massive parallel processing capabilities required for data handling tasks such as matrix multiplication. c) Is not correct. AI-specific hardware like ASICs is better suited for edge computing rather than training ML models, which is typically done in the cloud. d) Is not correct. Neuromorphic processors are designed to mimic brain structure and do not rely on the von Neumann architecture; therefore, this statement is inaccurate.	AI-1.1.5	K2	1
6	b	Given the following statements about AI model development and hosting:	AI-1.1.6	K2	1

		i. It achieves lower development costs by using public cloud resources, eliminating the need for local hardware investment ⇨ This is not a hybrid approach as it only talks about development on the cloud. ii. It uses local development of data preparation components for sensitive data for increased security before moving to the cloud for training of the full system ⇨ This is hybrid as we have local development for data preparation and training on the cloud. iii. It results in lower costs because laptops are used for local development and there are low upfront hardware costs by hosting AI models on public clouds ⇨ This is hybrid as there is local development on laptops and hosting on public clouds. iv. It simplifies development and hosting by standardizing processes on local servers, removing the need for complex cloud-based configurations ⇨ This is not hybrid as everything is on local servers. v. It guarantees the highest security by hosting AI models on private clouds, thereby avoiding the risks associated with local hardware vulnerabilities ⇨ This is not hybrid as everything is on private clouds. Hence, the correct option is b) ii and iii.ii			
7	c	Considering each of the examples in turn: A. The success rate of a remote operator in forcing a drone into the safe-landing protocol when its AI navigation system exhibits hazardous	AI-2.1.1	K2	1

		behavior. This measure assesses the effectiveness of human intervention in the system's operation during hazardous AI behavior, aligning with the definition of intervenability: the ability to permit external intervention in automated processes. This matches with intervenability (4). B. The average time required to successfully override a fraud management system's automated decision to block a customer's transaction. This measure is concerned with how quickly a user can override an automated system's decision and the degree to which a user can control or influence the system's actions. This is a direct indicator of user controllability (2). C. The F1-score of an object detection model in an autonomous car in heavy rain. The F1-score in adverse conditions (heavy rain) tests how well the model performs under challenging scenarios, which is a measure of AI robustness (1). D. The time required for an e-commerce recommendation engine to update its suggestions to reflect a new, rapidly emerging fashion trend. The speed at which a system adapts its recommendations to new trends reflects its ability to adapt its functionality to changing requirements, which is functional adaptability (3). In which case, the correct combination is: 1C – 2B – 3D – 4A, which makes c) correct.			
8	b	a) Is not correct. The actual problem is the opposite: requirements are generally too vague and are often only implicitly provided by data, leading to poor traceability. b) Is correct. This unpredictability is a core safety challenge when deploying AI in safety-related systems, as it complicates verification, validation, and ongoing assurance of safe operation.	AI-2.1.2	K2	1

		c) Is not correct. The problem is that self-learning systems do not stop adapting after deployment and so move away from their original tested behavior. d) Is not correct. Mature safety-related standards currently lack any provisions for AI, and some explicitly restrict its use.			
9	c	a) Is not correct. This relates to intervenability, as defined in ISO 25059, which describes the user's ability to control the autonomous patrol robot when they notice it might run into a sculpture. It also provides a specific, measurable timeframe for the intervention. b) Is not correct. This relates to functional correctness, which is defined in ISO 25059. It describes how the system adapts to changes in humidity to maintain the desired environment. It provides specific, measurable humidity and a timeframe for the control system to react. c) Is correct. It is the LEAST likely option because it focuses on some of the sub-characteristics of usability, such as learnability and operability, which are generic quality characteristics applicable to most systems, and are not directly associated with the AI-specific characteristics in ISO 25059 (i.e. user controllability and transparency, which are sub-characteristics of usability for AI-based systems). The acceptance criterion is also more subjective than the other options and so more challenging to test. d) Is not correct. This relates to transparency, which is defined in ISO 25059. It describes the system's provision of extra information to explain its decision-making.	AI-2.2.1	K2	1
10	a	Considering the descriptions of example systems: 1. Reinforcement learning - The amount spent can be considered the reward function for this system, with the system changing its behavior to increase the amount spent. (B)	AI-3.1.1	K2	1

		2. Classification - The app utilizes text in what can be considered a source language and a corresponding 'correct' translation of this source, employing a form of supervised learning with no explicit reward function mentioned. Classification uses parallel training data—pairs of matching sentences in two languages (e.g., "Hello" in English and "Bonjour" in French)—where each pair has a label linking source to target text. The model learns mappings via supervised training, encoding input text features and predicting the correct translation output, minimizing errors across many such pairs to generalize to new sentences. (C) 3. Regression is used in this scenario to predict a continuous outcome (in this case, the time until equipment failure) based on a set of input variables (sensor data and historical maintenance records). By analyzing the relationship between these variables, the ML model can identify patterns and trends that indicate when equipment is at risk of failure. (D) 4. Clustering - By analyzing user interactions (e.g., likes, comments, shared content) and stated interests, the social network platform can identify patterns and similarities among calibera users. These patterns can then be used to group users into communities or "clusters" that share common interests and behaviors. (A) Hence, 1B – 2C – 3D – 4A, is correct and so option a) is the correct option.			
11	d	1. Model performance is tested using validation data This is part of the 'Evaluate the Model' activity. (D) 2. The origin of the test data used to test the model is identified This is part of the 'Prepare & Test Data' activity. (B) 3. Test data are used to determine that the agreed performance criteria are met This is part of the 'Test the Model' activity. (C)	AI-3.1.2	K2	1

		4. The model is tested on the target platform This is part of the 'Deploy the Model' activity. (A) Hence, the correct option is: 1D – 2B – 3C – 4A, and so d) is correct.			
12	d	a) Is not correct. The RAG approach does not add new layers to the neural network to store documentation; instead, it retrieves relevant external data at inference time and augments the model's input with this information. b) Is not correct. While fine-tuning with high-quality data can reduce bias, it does not guarantee the prevention of unfair outputs, as bias can persist or even be introduced during the fine-tuning process. c) Is not correct. Fine-tuning can be performed on all layers or just a subset. Often, only some layers are updated while others are kept frozen, depending on the task and available data. d) Is correct. RAG relies on curating and indexing relevant external data before use, but it does not require modifying the architecture or parameters of the pre-trained model itself; instead, it augments the model's inputs with retrieved information at runtime.	AI-3.1.4	K2	1
13	c	a) Is not correct. This statement highlights a fundamental step in data preparation: collecting and consolidating data from different sources before any processing or analysis begins. However, data preparation typically also involves cleaning, transforming, and organizing data, in addition to gathering raw data. b) Is not correct. This statement is inaccurate because feature engineering is generally performed before (or during) model training, not after. The purpose of feature engineering is to create or modify features to improve ML model performance during the training phase. c) Is correct. This correctly identifies two common data pre-processing techniques: augmentation (increasing data by creating modified versions) and sampling (reducing data by selecting subsets).	AI-3.2.1	K2	1

		d) Is not correct. This statement partially captures the purpose of EDA, which is to understand data characteristics and detect issues through visualization and summary statistics. However, EDA is not "exploratory testing" but rather is an analytical process that informs subsequent data preparation steps.			
14	b	a) Is not correct. The training dataset is primarily used to fit the model's parameters, not to optimize hyperparameters. The validation dataset is typically used to tune hyperparameters, and the test dataset is never used to generate training data. b) Is correct. The training dataset is used to fit or create the model, the validation dataset is used to tune the model's hyperparameters and prevent overfitting, and the test dataset is reserved for evaluating the model's performance on unseen data. It accurately describes the distinct and sequential roles of each dataset in the ML workflow. c) Is not correct. The training dataset is not used for final model evaluation - that is the role of the test dataset. Additionally, the test dataset is not used for tuning hyperparameters, as that would risk overfitting to the test dataset. d) Is not correct. The training dataset's primary role is not to ensure generalization but to fit the model. The validation dataset is not used for deployment, and the test dataset is not used for initial evaluation but for final assessment after model development	AI-3.2.3	K2	1
15	c	a) Is not correct. See option c for the correct formula and calculation. b) Is not correct. See option c for the correct formula and calculation. c) Is correct. The formula for Precision = TP/ (TP+FP) *100 = 78/(78+22) = 78/100 *100 d) Is not correct. See option c for the correct formula and calculation.	AI-3.3.1	K3	2
16	a	a) Is correct. This is the next step in the ML training loop. The loss is fed back through the network to adjust the values of the weights and biases.	AI-3.4.1	K2	1

		b) Is not correct. The training proceeds to the next batch of data or completes the epoch. It doesn't rerun the same data just to confirm the loss value. c) Is not correct. Activation functions are chosen during model design and are not usually changed during a training run. d) Is not correct. Resetting the weights happens at the start of training (initialization), not after calculating the loss value for a batch of data.			
17	a	Considering the provided example AI-based systems: i. A system that learns from real-time data to improve its failure predictions and automatically updates maintenance schedules. This is a form of adaptive AI-based system that changes the schedules based on real-time data. ii. A spam filter in an email app, which identifies spam based on predefined rules. This is a form of locked AI-based system, which generates predictable test results, as it is based on following predefined spam identification rules that will not be changed until the rule-based model is rebuilt. iii. A recommendation engine on a streaming service that updates its suggestions based on a user's changing viewing habits and preferences. This is a form of adaptive system that adapts based on the user's viewing habits and preferences. iv. A personal assistant that learns from its user. This is an adaptive AI-based system because it adapts its parameters based on users' interactions with the system. v. A rule-based system for medical diagnosis. This is a form of a locked AI-based system, which will not change the domain rules until the rule-based model is rebuilt. Thus, the spam filter and medical diagnosis systems would be more straightforward to test because they are both forms of locked systems and generate predictable test results.	AI-4.1.1	K2	1

		And, so, option a) is correct			
18	b	a) Is not correct. While AI-based systems can be large and complex, the rationale for using a statistical approach is primarily driven by their non-deterministic nature, not the impracticality of automation or size. b) Is correct. AI-based systems are non-deterministic and require extensive, representative test datasets to achieve statistical significance, especially in assessing functional correctness statistically. c) Is not correct. This option misrepresents the concept. A single test case is not sufficient to assess the quality of the AI-based system. Statistical evaluation over many predictions is necessary. d) Is not correct. This contradicts best practices. A separate test dataset (from training and validation) is essential, not optional, and is part of the statistical rigor required.	AI-4.1.2	K2	1
19	b	a) Is not correct. Setting a seed is an implementation detail that is independent of the quality or completeness of the system's specifications. b) Is correct. The setting of a 'seed' is bound to a specific test execution and does not solve the fundamental probabilistic nature of the model when it operates in a real-world environment. c) Is not correct. Choosing a seed is a simple programming step and does not involve domain experts. d) Is not correct. Setting a seed does not make the behavior subjective. For a given test run, it should make the behavior objective and repeatable.	AI-4.1.3	K2	1
20	b	a) Is not correct. Generative AI models do not produce deterministic outputs, so exact matching with predefined expected results is not a suitable approach for this testing. b) Is correct. Because of the variability and complexity in GenAI outputs, testing focuses on coherence, rules compliance, and plausibility, rather	AI-4.2.1	K2	1

		than matching fixed expected outputs. Diverse inputs, optional prompts, and parameters all influence the test results. c) Is not correct. While generative AI models are probabilistic, testing (e.g., exploratory testing, metamorphic testing and adversarial testing) can still be used. d) Is not correct. Manual review is possible, but automated methods, including other GenAI tools or image recognition systems, can also effectively evaluate generated content. Therefore, asserting that automated testing 'does not apply' is inaccurate.			
21	d	a) Is not correct. Red teaming is typically performed before deployment. b) Is not correct. There is no guidance given in the question that security is more or less important than testing for bias. c) Is not correct. Red teaming is done proactively before deployment. d) Is correct. Red teaming should be conducted before deployment, not after and multiple vulnerability types (security and bias) should be tested together.	AI-4.2.2	K3	2
22	b	a) Is not correct. While system testing may have detected this defect, input data testing is the most effective test level for detecting it, as it would have been caught at an earlier phase. b) Is correct. Input data testing focuses on data quality and on the representativeness of data. It appears that areas in the UK with an altitude greater than 1,250 meters were not adequately represented in the training data. c) Is not correct. Component integration testing checks for defects in the interfaces and the interactions between components and is not focused on specific data values. d) Is not correct. ML model testing may have identified this defect, but input data testing would be the most effective test level to detect it, as it would have been caught earlier.	AI-4.3.1	K2	1

23	a	a) Is correct. Security and usability are generic quality characteristics and can apply to any system. Data bias and model performance are potential risks specific to ML systems. b) Is not correct. Both data bias and algorithmic bias are risks associated explicitly with an ML system and, therefore, do not apply to conventional systems. c) Is not correct. Risk management for both non-AI systems and AI systems (including self-learning ML systems) should be dynamic in nature. d) Is not correct.. Functional correctness may be the primary risk factor for conventional systems and for ML systems, but this is context dependent.	AI-4.3.2	K2	1
24	a	a) Is correct. The scenario describes a need to verify the data's origin and to determine that it hasn't been tampered with. This maps to data provenance testing. b) Is not correct. While the team would likely perform data representativeness testing on a new dataset, this activity addresses whether the data's characteristics match those of the real world, not whether its documented origin is accurate, or whether it is valid raw data. c) Is not correct. Feature testing evaluates the predictive power of the features, which is a different concern from verifying the source and integrity of the dataset itself. d) Is not correct. Dataset constraint testing checks for internal consistency of the dataset (e.g., ranges, types), but it cannot verify where the data came from or if it was illegitimately altered.	AI-5.1.1	K2	1
25	c	a) Is not correct. This approach focuses on the model logic, not the characteristics of the training data. It does not directly uncover bias patterns present within the data.	AI-5.1.2	K2	1

		b) Is not correct. This approach examines model outputs rather than the training data itself. It is valuable for identifying bias in predictions, but not for initial detection of bias in the data. c) Is correct. This approach could help highlight where bias might be introduced, but it relies on process checks rather than actual evidence from the data. d) Is not correct. This approach evaluates the outcomes or predictions under hypothetical changes, targeting outcome bias rather than directly assessing the training dataset.			
26	d	a) Is not correct. Both pipeline types would benefit from a full layered approach. An operational pipeline would require extensive end-to-end system testing, not just validation of individual scripts. b) Is not correct. Both pipeline types would use a range of tests. Limiting one to only system level test approach and the other to only component and integration level test approach is not following a layered approach. c) Is not correct. Configuration management verifies correct versions are used 'across training, testing, and production', implying it is important for both types of pipelines to provide consistency and reproducibility, not just operational ones. d) Is correct. The purpose of the pipeline dictates the test strategy. Training pipelines prioritize generating high quality data, whereas live operational pipelines focus on non-functional aspects such as performance efficiency, scalability, and AI robustness.	AI-5.1.3	K2	1
27	b	a) Is not correct. A reference dataset does not need to be universally applicable or solely derived from industry benchmarks. It is typically tailored to match the operational and contextual characteristics of the AI system's target population. b) Is correct. The reference dataset's core purpose is to provide a statistical baseline, allowing practitioners to objectively compare the distributions,	AI-5.1.4	K2	1

		feature correlations, and coverage of the training data against what is expected in real-world operational environments. c) Is not correct. Stratified sampling is typically applied to the reference dataset to create a comprehensive baseline, rather than directly to the training data itself. The reference dataset aims that all relevant subgroups are represented, serving as a standard for evaluating the representativeness of the other datasets. d) Is not correct. The reference dataset is not intended as the main source for final ML model testing or testing with high independence. It exists to evaluate the representativeness of data before model training.			
28	c	a) Is not correct. A range constraint only validates that a single attribute's values fall within specified bounds, but doesn't verify the mathematical relationship between monthly_payment and loan_amount. b) Is not correct. A count constraint checks the quantity of non-null values, not the mathematical correctness of calculated fields. c) Is correct. This scenario requires validating a mathematical relationship between two attributes (monthly_payment and loan_amount). The correlate constraint is specifically designed to check that values for one attribute correlate with values for another attribute according to a defined rule - in this case, that monthly_payment equals loan_amount/324. d) Is not correct. A duplicate constraint identifies identical values but having duplicate monthly payments could be legitimate (different applicants might have the same payment amount).	AI-5.1.5	K3	2
29	b	a) Is not correct. Multiple annotations are not normally automated. Multiple annotation typically relies on human annotators and does not ensure consistency; instead, it highlights inconsistency. b) Is correct. Multiple annotation involves data points being independently labeled by multiple annotators. Comparison of these labels leads to the	AI-5.1.7	K2	1

		discovery of disagreements among annotators, highlighting potential defects. c) Is not correct. Comparing label distributions is a part of data distribution analysis, not multiple annotations. d) Is not correct. Flagging high-loss data points is the role of model loss analysis, not multiple annotations.			
30	b	Considering each of the risks in turn: A. The ML model might perform differently for different demographic groups. This is best matched with Testing for Bias, which explicitly examines fairness across different populations. (2) B. Slightly modified inputs to the ML model might cause quite different and unexpected responses. This pairs correctly with Adversarial Testing, which deliberately tests an ML model's ability to handle small input perturbations. (3) C. Predictions made by the ML model might be inaccurate in some cases. This is best addressed by ML Functional Performance Testing, which directly evaluates whether the ML model produces accurate outputs. (1) D. ML model accuracy might have significantly decreased since it was deployed. This is appropriately matched with Drift Testing, which monitors performance changes over time in production. (4) Thus, the correct match between risks and test approaches is: 1C – 2A – 3B – 4D, and so b) is the correct option.	AI-6.1.1	K2	1
31	a	a) Is correct. This is a part of the non-functional requirement/documentation. b) Is not correct. It is not in the scope of model documentation. c) Is not correct. Model documentation should contain the source of the training data.	AI-6.1.2	K2	1

		d) Is not correct. Documentation of changes made by the self-learning system could only be made by the system itself at the time of the update.			
32	a	a) Is correct. When more tests are run for the same measured accuracy (83%), the margin of error typically decreases if the confidence level stays the same. This is because a larger sample size can reduce the margin of error and/or increase our statistical confidence in the measured value. If the ML model previously achieved 83% accuracy with a margin of error of ± 4% at a 94% confidence level, and additional tests were run while maintaining the same 83% accuracy measurement and maintaining the confidence level at 94%, this would result in a lower margin of error (e.g. ± 2%). This indicates we can be more certain that the true accuracy is indeed around 83%. b) Is not correct. When more tests are run, the margin of error typically decreases and the confidence level typically increases rather than decreases, assuming the measured accuracy remains the same. Going from 94% to 92% confidence would indicate fewer tests were run or that the test results were more variable, which contradicts the scenario where additional tests were performed. c) Is not correct. If additional tests were run and the accuracy remained at 83%, the margin of error would typically decrease and not increase, if the confidence level was maintained at 94%. The increased margin of error of ± 6% makes this option less likely than the option where it decreases (a). d) Is not correct. The scenario states that the ML model's accuracy for the additional tests was measured to be 83%, not 85%. An increase in the accuracy percentage would indicate that the ML model performed better in the additional tests, which contradicts the given information that the ML model "provided correct predictions fewer times than expected" while maintaining the same measured accuracy of 83%.	AI-6.1.3	K2	1

33	c	a) Is not correct. It incorrectly implies that internal knowledge of the ML model is irrelevant, whereas such knowledge can enhance the testing process by crafting more targeted adversarial examples. b) Is not correct. It incorrectly suggests that manual methods are the only approach, ignoring automated techniques, which are also crucial in adversarial testing. c) Is correct. In adversarial testing, the adversarial examples are often created to identify vulnerabilities by perturbing working inputs. d) Is not correct. This misses the point of adversarial testing, which deliberately uses updated input values (i.e. tests) to create adversarial examples.	AI-6.1.4	K2	1
34	b	Given the following follow-up test cases: Test case T1 differs from the source test case by the change in requirements for a 3D camera; it is now more specific. A 3D camera must be included. So, that means the follow-up expected results can only include the original test results at most (the previously recommended phones but with a 3D camera). Test case T2 also differs from the source test case by the change in requirements for a 3D camera; it is also more specific. No 3D camera should be included. So, that means the follow-up expected results can only include the original test results at most (the previously recommended phones but without a 3D camera). As T1 lists phones with a 3D camera, the remaining phones from the source test case must be those without a 3D camera, and therefore, they should be in T2. Therefore, T1 and T2 combined should contain all the cameras from the source test case, but with no overlap between the two. a) Is not correct. The camera SnapHappy M3 is missing in the combined outputs of T1 and T2	AI-6.1.5	K3	2

		b) Is correct. There is no overlap between the outputs of T1 and T2, and no camera is missing. c) Is not correct. The SnapHappy cameras are listed for both test cases. d) Is not correct. The output of the two test cases is identical, despite the differing requirements for the 3D camera.			
35	b	a) Is not correct. While dynamic testing is appropriate for when ground truth is available, it does not analyze the input data's statistical properties. b) Is correct. Because ground truth is unavailable, static drift testing, which doesn't require it, is the only viable option. c) Is not correct. The organization "has no mechanism to track if a customer actually leaves." Therefore, they have no actual results to compare against, making dynamic testing impossible to perform. d) Is not correct. Static testing looks at data distributions, whereas comparing performance metrics against ground truth is the mechanism of dynamic testing.	AI-6.1.7	K2	1
36	c	a) Is not correct. The ML model performs well on validation data, so it is unlikely to be a case of underfitting. b) Is not correct. Concept drift refers to changes that occur after the training of the ML model and the validation stages. c) Is correct. The poor performance on test data and good performance on validation data suggest overfitting. d) Is not correct. Acceptance criteria should be consistent with different sets of data, so low acceptance criteria are unlikely to lead to a difference between the test results with validation data and independent test data.	AI-6.1.8	K2	1
37	c	a) Is not correct. According to the syllabus, "A/B testing does not generate test cases and provides no guidance on how the tests should be designed, although operational inputs are often used in tests." A/B	AI-6.1.9	K2	1

		testing compares responses of two variants rather than generating test cases. b) Is not correct. A/B testing is not focused on component interactions within a system. It compares two different versions of a system, not components within a single system. c) Is correct. As stated in the syllabus, “Whenever the system is updated, A/B testing is used to test that the updated variant performs as well as, or better than, the previous variant.” This accurately describes how A/B testing is used in the context of ML systems. d) Is not correct. A/B testing does not analyze the internal algorithm structure. It is a “statistical testing approach that typically requires comparing test results from several test runs to determine the difference between the programs.” It focuses on comparing the outputs of two variants, not examining their internal structures.			
38	d	a) Is not correct. Varying hyperparameters is a minor adjustment to the same ML model. It doesn't fundamentally change the model. A pseudo-oracle should ideally be significantly different to detect defects effectively. This option is too similar to the SUT. b) Is not correct. Fine-tuning is a modification of the same ML model, typically involving further training on a specific dataset. It's not a substantial enough change to create a truly independent pseudo-oracle. It still relies on the same core model and framework, increasing the risk of shared defects. c) Is not correct. Retrieval-augmented generation (RAG) enhances a language model by integrating it with a retrieval mechanism. While RAG adds complexity, it still builds upon the core ML model being tested. The fundamental architecture and training process of the base model will be the same. While RAG introduces a difference, it is still an augmentation of the original model rather than an independent alternative.	AI-6.1.10	K2	1

		d) Is correct. Using a different ML development framework means using different libraries, potentially different underlying algorithms or implementations of algorithms, and a different ML development environment. This option maximizes the independence of the pseudo-oracle from the SUT, reducing the risk of shared defects and increasing the potential effectiveness of back-to-back testing.			
39	b, d	a) Is not correct. This describes a risk of poor performance efficiency. The correct mitigation is performance (efficiency) testing, which is distinct from ML functional performance testing (which evaluates functional correctness). b) Is correct. This describes a potential used library defect. ML functional performance testing can evaluate model behavior and expose anomalies stemming from the use of a new library that is defective and causing a change in test results. c) Is not correct. This describes the need to mitigate the risk of a defective ML framework installation. The appropriate mitigation for this is smoke testing, not full ML functional performance testing. d) Is correct. This describes the risk of poor interpretation of test results due to the stochastic nature of the learning process. ML functional performance testing, using multiple runs and statistical analysis, is the mitigation for this. e) Is not correct. This describes sub-optimal algorithm selection. The appropriate mitigation is an algorithm suitability review or A/B testing, not ML functional performance testing of a single algorithm that has not been selected yet.	AI-7.1.1	K2	1
40	b	a) This is not correct: It reverses the roles of both test types. Canary testing directly serves a small group of real users with the new ML model, while shadow testing processes live traffic in the background without ever affecting user responses.	AI-7.1.2	K2	1

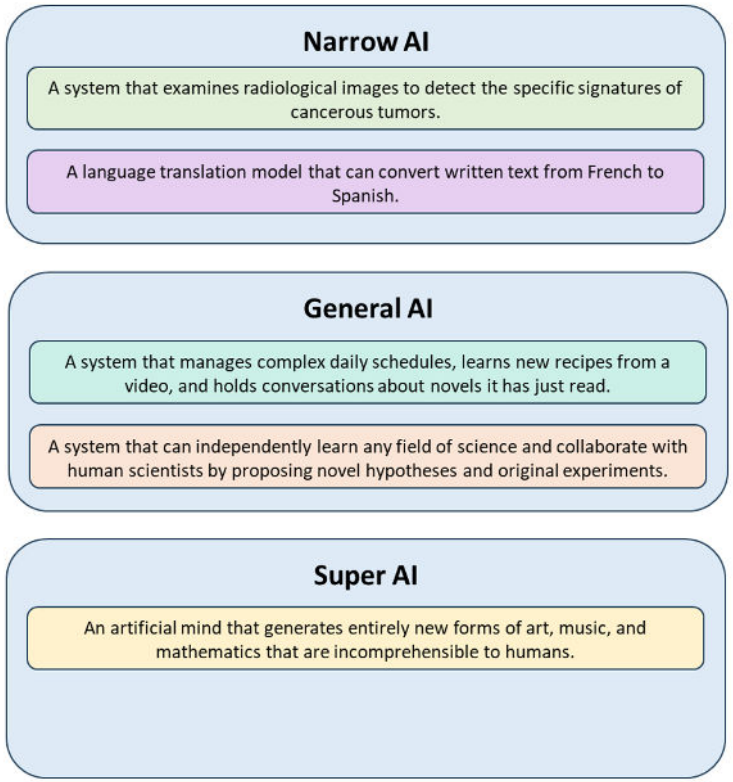
		b) This is correct: Shadow testing provides a risk-free way to see how a new ML model behaves with production data, whereas canary testing deliberately exposes a small subset of users to the new model to observe its real-world performance and impact. c) This is not correct: Both test types are used to validate a model's performance and functional correctness on live traffic. The primary difference is the risk profile, not whether one is for performance testing and the other is for component integration testing. d) This is not correct: It fundamentally misrepresents canary testing as an offline activity. Both canary testing and shadow testing are online methods that use live user data. The key distinction is whether the new ML model's response is actually sent to the user or is only logged for later analysis.			

Appendix: Answers to Additional Questions

Answer Key

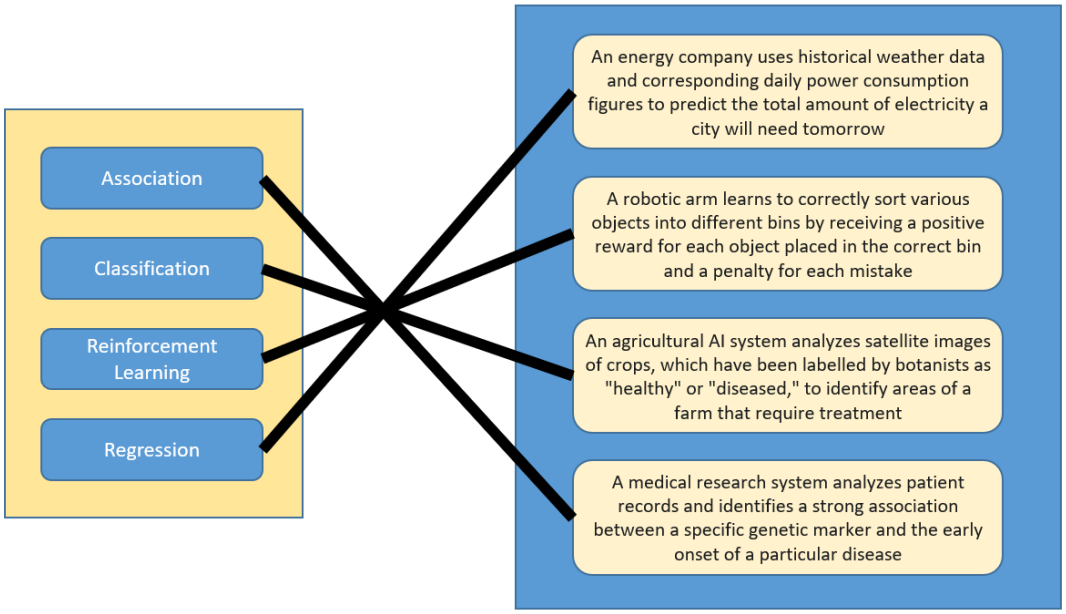
Question Number (#)	Correct Answer	LO	K-Level	Points
A1	New question type	AI-1.1.2	K2	1
A2	a	AI-1.1.7	K2	1
A3	c	AI-1.1.8	K2	1
A4	New question type	AI-3.1.1	K2	1
A5	New question type	AI-3.1.2	K2	1
A6	c	AI-3.4.3	K2	1

Answers

Question Number (#)	Correct Answer	Explanation / Rationale	Learning Objective (LO)	K-Level	Number of Points
A1		 **Explanation / Rationale**	AI-1.1.2	K2	1

		Considering each of the example systems in turn: • An artificial mind that generates entirely new forms of art, music, and mathematics that are incomprehensible to humans. This system demonstrates creativity and intelligence far beyond human capability, producing outputs not understandable by people. Such abilities surpass both human-level general intelligence and any current artificial general intelligence, and so this is an example of super AI. • A system that manages complex daily schedules, learns new recipes from a video, and holds conversations about novels it has just read. This AI performs a wide range of complex tasks that require learning, perception, planning, and language understanding, comparable to a human's versatility. It shows adaptability and competency across novel domains. Thus, it is an example of AGI. • A system that can independently learn any field of science and collaborate with human scientists by proposing novel hypotheses and original experiments. This system has the ability to autonomously and flexibly master multiple scientific disciplines and engage in creative research. Its competence covers learning, reasoning, and innovative thinking at a human expert's level. Thus, it is an example of AGI. • A system that examines radiological images to detect the specific signatures of cancerous tumors. This AI excels at a well-defined, specialized task (medical image analysis) without broader understanding or abilities outside of its domain. Its application is confined to pattern recognition within radiological data and does not extend to generalized cognition. Thus, it is an example of narrow AI. • A language translation model that can convert written text from French to Spanish. This model is built to perform a single function - translating between two languages. It demonstrates expertise in one bounded task and does not			

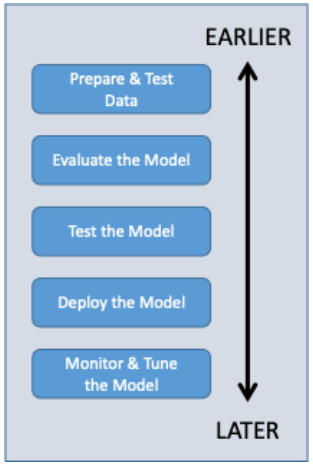
		possess comprehensive understanding or adaptability beyond language conversion. Thus, it is an example of narrow AI.			
A2	a	a) Is correct. Model building is a key function, and ML development frameworks provide “tools for defining the architecture of the ML model,” including the ability to specify the structure of the ML model, such as a decision tree. b) Is not correct. The word "eliminate" is an overstatement. ML development frameworks assist with, but do not remove the need for data preprocessing. Data handling is a support function that does not perform complete automation, eliminating the task. c) Is not correct. ML development frameworks exist at different abstraction levels, including “a higher-level API, simplifying model creation”. This suggests that programming is not a universal requirement, with the selection of the framework depending on the “expertise of the users.” d) Is not correct. It is not correct that ML development frameworks lock in developed ML models to that framework. Generally, any developed ML models can be deployed anywhere (within reason).	AI-1.1.7	K2	1
A3	c	a) Is not correct. ISO/IEC/IEEE TR 29119-11 is not about testing AI, but, as with all standards, it provides guidelines and does not impose penalties for non-compliance. b) Is not correct. OECD is a non-mandatory set of guidelines. c) Is correct. If the system is covered by the EU AI Act (i.e., it will be used in Europe), failing to follow a risk-based approach can result in severe penalties. d) Is not correct. ISO/IEC 25059 describes the unique quality characteristics of AI systems, but, like all standards, it provides guidelines rather than imposing penalties for non-compliance.	AI-1.1.8	K2	1
A4			AI-3.1.1	K2	1

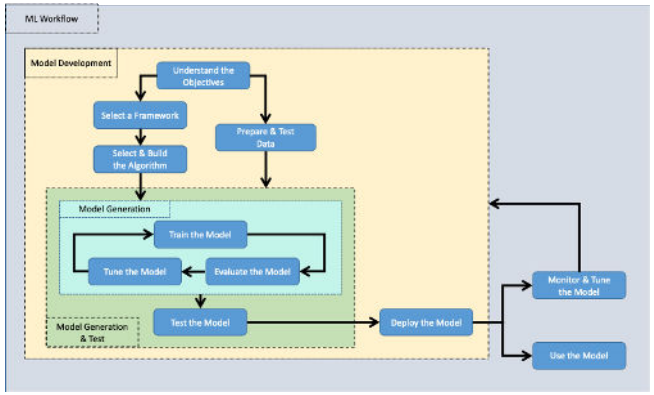


Explanation / Rationale

Considering each of the descriptions of example systems from top to bottom:

- Association - This system uses association by mining unlabeled patient records to identify dependencies between different attributes, such as a gene and a disease. It does not predict a class but instead reveals the strength of the relationship between variables within the data.
- Classification - It learns from a dataset of images that have been pre-labeled with discrete categories ('healthy' or 'diseased'). Its purpose is to categorize new images into one of these specific, non-numerical classes, distinguishing it from regression.

		• Reinforcement learning - The system learns through direct interaction with its physical environment, not from a preexisting dataset. The robot's behavior is shaped over time by a system of rewards (for correct placement) and penalties (for incorrect placement). • Regression - The ML model is trained on labeled data to predict a specific, continuous numerical value: the amount of electricity. The model's output is not a category but a quantity on a scale, which is the key characteristic of regression.			
A5		**Correct answer**  **Explanation/Rationale**	AI-3.1.2	K2	1

					
A6	c	a) Is not correct. The core limitation is the model's internal reasoning, a general problem applicable to many tasks, not specifically those involving human feedback. b) Is not correct. Structural coverage is a method for evaluating the test thoroughness of a single model, not for comparing different models. While model size may affect the raw coverage numbers, calculated values are typically presented as percentages, and therefore, this is not a limitation for its intended purpose. c) Is correct. It directly addresses the key limitations of coverage measures used in isolation. Neural networks can learn spurious correlations, leading to correct activations for incorrect reasons, meaning the model could achieve high structural coverage while making decisions based on irrelevant features. d) Is not correct. Structural coverage is a valuable supplement that should be combined with other methods. It is intended to enhance, not replace, other test activities, such as evaluating the functional correctness and usefulness of the model's outputs.	AI-3.4.3	K2	1