Cheat Sheet - AWS Certified AI Practitioner Test

Service/Term	Definition/Usecase
SageMaker Data Wrangler	Data Preparation, Transformation and feature engineering Tool Augment the data - Generate New instances of data for underrepresented groups to fix Bias by balancing data sheet Single Interface for Data Selection, Cleaning, exploration, visualization and proccessing It has SQL Support for data Query It has Data Quality tool to analyze Quality of the data
Amazon SageMaker Model Cards	SageMaker Model Cards are a feature of SageMaker that you can use to record information about ML models. SageMaker Model Cards include information such as training details, evaluation metrics, and model performance.
SageMaker Canvas	You can use SageMaker Canvas to build ML models without needing to write any code. SageMaker Canvas does not have any models that can perform content moderation of creative content types.
Amazon SageMaker Ground Truth	SageMaker Ground Truth is a service that uses a human workforce to create accurate labels for data that you can use to train models. SageMaker Ground Truth does not store information about model training and performance for audit purposes.
Amazon SageMaker Model Monitor	SageMaker Model Monitor establishes an automated alert system that alerts when there are variations in the model's quality, such as data drift and anomalies. You can use SageMaker Model Monitor to monitor deployed models for performance issues, data drift, and operational inconsistencies. You would primarily use SageMaker Model Monitor to ensure that the model's performance remains stable over time
SageMaker Studio	SageMaker Studio offers a suite of integrated development environments (IDEs), including JupyterLab, RStudio, and Visual Studio Code - Open Source (Code-OSS). You can use SageMaker Studio to build content moderation models that can handle creative content types. However, this solution requires additional operational overhead.
Guardrails for Amazon Bedrock	Amazon Bedrock Guardrails evaluates user inputs and FM responses based on use case specific policies, and provides an additional layer of safeguards regardless of the underlying FM.
Amazon Rekognition	Amazon Rekognition is a fully managed AI service for image and video analysis. You can use Amazon Rekognition to identify inappropriate content in images, including drawings, paintings, and animations. Amazon Rekognition is designed specifically for performing content moderation of the creative content types. Additionally, you can access

	Amazon Rekognition directly through an API. Therefore, Amazon Rekognition requires the least operational overhead.
Low bias and low variance (Ideal)	**Low bias** indicates that the model is not making erroneous assumptions about the training data. **Low variance** indicates that the model is not paying attention to noise in the training data. This is an ideal outcome for model training and would not result in model overfitting / underfitting.
Overfitting (low bias and high variance)	**Low bias and high variance:** Low bias indicates that the model is not making erroneous assumptions about the training data. High variance indicates that the model is paying attention to noise in the training data and is overfitting.
Underfitting (High bias and low variance)	**High bias** indicates that the model is making erroneous assumptions about the training data. **Low variance** indicates that the model is not paying attention to noise in the training data, which will lead to underfitting
AWS Artifact	AWS Artifact is an audit resource that provides on-demand access to security and compliance documentation for the AWS Cloud.
AWS CloudTrail	CloudTrail is a service that tracks user activity and API usage on AWS. You can use CloudTrail for audit purposes to record actions taken by users, roles, and services in your AWS account.
Amazon CloudWatch	CloudWatch is a centralized logging service that monitors AWS resources and stores application logs and performance metrics. You can use CloudWatch to monitor and observe resources
AWS Trusted Advisor	Trusted Advisor provides resources and recommendations for cost optimization, security, and resilience. Trusted Advisor evaluates your AWS environment, compares environment settings with best practices, and recommends actions to remediate any deviation from best practices.
Amazon Macie	Macie uses ML to discover, monitor, and protect sensitive data that is stored in Amazon S3. You can use Macie to identify and protect PII. You can use Macie to comply with data governance and privacy regulations
AWS Config	AWS Config provides an overview of your AWS resource configurations. You can use AWS Config to identify how resources were configured in the past. AWS Config can identify settings that do not meet compliance standards, such as if an S3 bucket is publicly accessible.
Amazon DocumentDB	Amazon DocumentDB is a fully managed, native JSON document database. You can use Amazon DocumentDB to operate critical document workloads at scale without the need to manage infrastructure. Amazon DocumentDB supports vector search. You can use vector search to store, index, and search millions of vectors with millisecond response times. Amazon DocumentDB can perform real-time similarity queries with low latency.

Amazon OpenSearch Service	OpenSearch Service is a fully managed service that you can use to deploy, scale, and operate OpenSearch on AWS. **You can use OpenSearch Service vector database capabilities for many purposes.** For example, you can implement semantic search, retrieval augmented generation (RAG) with large language models (LLMs), recommendation engines, and multimedia searches. OpenSearch Service supports storing vector embeddings for similarity search capabilities with low latency. OpenSearch Service can also scale to store millions of embeddings and can support high query throughput.
Amazon SageMaker Clarify	SageMaker Clarify is a feature of **SageMaker that helps you explain how a model makes predictions and whether datasets or models reflect bias.** SageMaker Clarify also includes a library to evaluate FM performance. The foundation model evaluation (FMEval) library includes tools to compare FM quality and responsibility metrics, including bias and toxicity scores. FMEval can use built-in test datasets, or you can provide a test dataset that is specific to your use case.
	It can detect biases in training data and model predictions. You can use SageMaker Clarify to provide insights into model decisions. Therefore, SageMaker Clarify is a suitable solution to develop responsible and fair AI systems.
SageMaker JumpStart	SageMaker JumpStart is a hub that consists of hundreds of **open source pre-trained models** for a wide range of problem types. However, a company cannot insert its models into SageMaker JumpStart.
SageMaker Model Registry	SageMaker Model Registry is a fully managed catalog for ML models. You can use SageMaker Model Registry to manage model versions, associate metadata with models, and manage model approval status. **You can use SageMaker Canvas to push built models to SageMaker Model Registry. SageMaker Studio users can then access the same SageMaker Model Registry and the models in the registry.** This solution requires the least operational overhead because the company needs only to register the models to implement the workflow.
Embeddings	Embeddings are vector representations of content that captures semantic relationships. Embeddings provide content with similar meanings to have close vector representations. Embeddings are a crucial component of text generation models. Embeddings give the model the ability to understand and generate coherent and meaningful text.
Amazon Inspector	Amazon Inspector is a vulnerability management service that continuously scans workloads for software vulnerabilities and unintended network exposure. Amazon Inspector assesses the security and compliance of your AWS resources by performing automated security checks based on best practices and common vulnerabilities. Amazon Inspector can assess EC2 instances and Amazon ECR repositories to provide detailed findings and recommendations for remediation. You can

	use Amazon Inspector to maintain a secure and compliant AWS environment.
Amazon Comprehend	Amazon Comprehend is a service that uses **natural language processing (NLP) to extract insights from documents.** Comprehend can use built-in or custom models to analyze text in real-time. You can recognize entities, extract key phrases, detect dominant languages, detect and redact PII, determine sentiment, detect targeted sentiment, or analyze syntax.
Amazon Textract	Amazon Textract is a service that you can use to add document text detection and analysis to applications. You can use Amazon Textract to identify handwritten text, to extract text from documents, and to extract specific information from documents. Amazon Textract does not provide access to FMs.
Amazon Kendra	Amazon Kendra is an intelligent search service that provides answers to questions based on the data that is provided. Amazon Kendra uses semantic and contextual understanding to provide specific answers. Amazon Kendra does not provide access to FMs.
Amazon Q Business	Amazon Q Business is a generative AI virtual assistant that can answer questions, summarize content, generate content, and complete tasks based on the data that is provided. Amazon Q Business does not provide access to FMs.
Retrieval Augmented Generation(RAG)	Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data (external) sources before generating a response.
Fine Tuning	Fine-tuning refers to the process of taking a pre-trained langauge model and further training it on a specific tasks or domain-specific dataset. Fine-tuning allows the model to adapt its knowledge and capabilities to better suit the requirments of the business use case. There are 2 ways of fine-tuning a model: 1/ Instruction fine-tuning uses examples of how the model should respond to a specific instruction. Prompt tuning is a type of instruction fine-tuning. 2/ Reinforcement learning from human feedback (RLHF) provides human feedback data resulting in a model that is better aligned with human preferences. During fine-tuning, the model's parameters are updated to better capture the patterns and nuances in the task-specific data. Fine-tuning FMs is medium overhead Benefits of fine-tuning: Increase specificity, Improve accuracy, Reduce bias, Boost efficiency
Continued Pre-Training	Using unlabeled data - industry specific unlabeled data

Transfer learning (A method of fine tuning)	This approach is a method where a model developed for one task is reused as the starting point for a model on a second task(different but related task).
Chain of Thought	Chain of thought is a prompt engineering technique that breaks down a complex question into smaller parts. Chain-of-thought prompting is the recommended technique when you have arithmetic and logical tasks that require reasoning.
Amazon Lex	Used for **conversational voice and text**. Amazon Lex is a fully managed artificial intelligence (AI) service with advanced natural language models to design, build, test, and deploy conversational interfaces in applications.
Amazon Transcribe	Converts speech to text
Amazon Polly	Converts text to speech
Amazon Personalize	Personalized Product Information
Amazon Translate	Translates between 75 languages
Amazon Forecast	Predicts future points in time series data
Amazon fraud Detector	* Detects fraud and fradulant activities * Checks online transactions, product reviews, checkouts and payments
Bias	Unfair prejudice or preference that favors or disfavors a person or group.
Fairness	Impartial and just treatment without discrimination
Overfitting	When a model performs well on training data but fails to generalize to new data; **Low bias and high variance:** Low bias indicates that the model is not making erroneous assumptions about the training data. High variance indicates that the model is paying attention to noise in the training data and is overfitting.
Underfitting	When a model is too simple to capture the underlying patterns in the data; **High bias and low variance**: High bias indicates that the model is making erroneous assumptions about the training data. Low variance indicates that the model is not paying attention to noise in the training data, which will lead to underfitting.
Explainability	The ability to understand how a model arrives at a prediction Explainability is how to take an ML model and explain the behavior in human terms. Through model agnostic methods (partial dependence plots, SHAP dependence plots, or surrogate models) you can discover meaning b/t input data attributions and model outputs. With that

	understanding you can explain the nature and behavior of the AI/ML model.
Interpretability	Interpretability is a feature of model transparency. **Interpretability is the degree to which a human can understand the cause of a decision**. Interpretability is access into a system so that a human can interpret the model's output based on the weights and features.
Traditional ML model	Predict the customer turnover rate for a telecommunication companyCreate a text sentiment analysis application - it doesn't generate new content
Generative AI model	Develop a large patent repository of English-to-French translations that includes image processingBuild unique, realistic images or videos from text prompts and descriptions for advertising and marketing campaigns
ROUGE (Model Evaluation)	Recall-Oriented Understudy for Gisting Evaluation - a set of metrics used to **evaluate automatic summarization** of texts in addition to **machine translation quality** in NLP. ROUGE is widely used because it is not complex. It is interpretable, and correlates reasonably well with human judgment, especially when evaluating the **recall** aspect of summaries.
BLEU (Model Evaluation)	Bilingual Evaluation Understudy - a metric used to **evaluate the quality of text that has been machine-translated from one natural language to another**. BLEU is fundamentally a **precision** metric. It checks how many words or phrases in the machine translation appear in the reference translations.
Top K	Top-K is a parameter used in language models to limit the selection of tokens to the K most probable options during text generation, controlling the balance between diversity and predictability in the output. It allows data scientists to fine-tune the model's creativity and coherence when deploying and using language models through SageMaker's infrastructure, often in combination with other sampling techniques.
Top P	Top P is a setting that controls the diversity of the text by limiting the number of words that the model can choose from based on their probabilities. Top P is set on a scale of 0-1. **Low Top P** (like 0.25), the model will only consider words that make up the top 25% of the total probability distribution. This can help the output to be more focused and coherent because the model is limited to choosing from the most probable words given the context. **High top P** (0.99) - the model will consider a broad range of possible words for the next word in the sequence because it will include words that make up the top 99% of the total probability distribution. This can lead to more diverse and creative outputs because the model has a wider pool of words to choose from.
F1	F1 score balances precision and recall by combining them in a single metric.

	The F1 score is a metric that you can use to **evaluate classification models**. F1 = 2 * P * R / P + R (P = Precision , R = Recall)
BERT (Model Evaluation)	Bidirectional encoder representation tranformers. BERTScore uses pretrained contextual embeddings from models like BERT to evaluate the quality of text-gen tasks. BERTScore computes the cosine similarity between the contextual embeddings of words in the candidate and reference text. BERTScore is sensitive to minor paraphrasing and synonym usage that does not affect the overall meaning conveyed by the text. BERTScore is used in cases where capturing the deeper semantic meaning of the text is important. **BERTScore is a metric that you can use to evaluate the quality of text that is generated by a text-to-text language model**. BERTScore measures the **semantic similarity** between the generated text and the reference text. Therefore, you can use BERTScore to assess the similarity between chatbot and human responses.
Semantic Robustness	Evaluates how much your model output changes as the result of small, semantic-preserving changes in the input. Foundation Model Evaluations (FMEval) measure how your model output changes as a result of keyboard typos, random changes to uppercase, and random additions or deletions of white spaces.
Perplexity	Perplexity is a metric that you can use to evaluate language models. Perplexity measures the probability of a model to generate a given sequence of words.
Accuracy	Correct predictions / All predictions. the percentage of correct predictions on a 0-1 scale. Accuracy is not a good measure when the data is skewed (if 90% of the data is the same, the model can score a 90% by predicting that answer all the time)
Precision	True positives/(true positives + false positives)
Recall	True positives/(true positives + false **negatives**). a.k.a. TPR (True Positive Rate) / Sensitivity. False Postive **is NOT** Taken into Account. Only True Positives are considered in Recall.
False Postive Rate (FPR)	False Positives / (False Positives + True Negatives) a.k.a specificity
AUC-ROC	Area under the curve - Receiver Operating Characteristics — taken from graph of **True Positive Rate (TPR or Recall)** over **False Positive Rate (FPR)**. Increasing the threshold results in few false positives, but more false negatives. A score of 1 indicates perfect accuracy. A score of .5 indicates 50/50.
Mean Squared Error (MSE)	Mean squared error, or the average of the squared differences between the predicted and actual values. MSE values are always positive. The

	better a model is at predicting the actual values, the smaller the MSE value is.
R squared or R2	The percentage of the difference in the target column that can be explained by the input column. Quantifies how much a model can **explain the variance** of a dependent variable. Values range from one (1) to negative one (-1). Higher numbers indicate a higher fraction of explained variability. Values close to zero (0) indicate that very little of the dependent variable can be explained by the model. Negative values indicate a poor fit and that the model is outperformed by a constant function (or a horizontal line).
Root Mean Squared Error (RMSE)	Root Mean Squared Error, or the standard deviation of the errors. Measures the square root of the squared difference between predicted and actual values, and is averaged over all values. **It is used to understand model prediction error**, and it's an important metric to indicate the presence of large model errors and outliers. Values range from zero (0) to infinity, with smaller numbers indicating a better model fit to the data. RMSE is dependent on scale, and should not be used to compare datasets of different types.
Temperature	The temperature parameter in generative models is a scaling factor that controls the randomness or diversity of the generated outputs. A higher temperature value increases the probability of sampling from less likely or lower-probability output tokens, resulting in a more diverse and unpredictable response. A lower temperature value favors the most probable outputs, leading to more deterministic and repetitive respones. **Higher Temperature Value generates Most Creative, Random Output.**
Amazon SageMaker	A fully managed service that data scientists and developers use to quickly build, train, and deploy ML models.
Amazon Bedrock	A fully managed service that makes FMs from Amazon and leading AI companies available through an API. Amazon Bedrock has a broad set of capabilities to quickly build and scale genAI applications with security, privacy and responsible AI. With Bedrock serverless experience, you can quickly get started using FMs without the need to manage any infrastructure. You can also privately customize FMs with your own data and seamlessly integrate and deploy them into your apps using AWS tools and capabilities. Bedrock's RAG Implementation is “Knowledge Base”
Amazon Nimble Studio	Accelerate visual content creation in the cloud
Amazon Sumerian	3D Content creation
Gen AI Architectures	Generative adversarial networks (GANs), Variational autoencoders (VAEs), Transformers, Diffusion Model

AI Project Lifecycle stages	Identify use case -> Feature Engineering -> Experiment and Select Model -> Adapt, Align, Augment -> Evaluate -> Deploy and Integrate → Monitor
AWS Audit Manager	Continually audit your AWS usage to simplify risk and compliance assessment
SageMaker inference	**Real-time** inference allows you to deploy your model to SageMaker hosting services and get a fully managed, autoscaling endpoint that can be used for real-time inference. **Serverless** inference lets you deploy and scale without managing any underlying architecture. **Asynchronous** inference queues incoming, large requests and processes them asynchronously. **Batch transform is for batch inference**
AWS AI Service Cards	Resource to help customers better understand our AWS AI services.
Amazon SageMaker Debugger	Amazon SageMaker Debugger helps debug and optimize machine learning models by monitoring and profiling training jobs in real-time. It does not address label inconsistencies directly.
Amazon Augmented AI (Amazon A2I)	Amazon A2I is a service to build human review systems for ML solutions. You can use Amazon A2I to create a workflow for human reviewers to audit individual predictions. Amazon A2I is not a reporting tool designed to support system-level compliance audits.
Amazon SageMaker Autopilot	Amazon SageMaker Autopilot uses tools provided by SageMaker Clarify to help provide insights into how ML models make predictions. These tools can help ML engineers, product managers, and other internal stakeholders understand model characteristics.
Epoch	One training cycle through the entire dataset. It is common to have multiple iterations per an epoch. The number of epochs you use in training is unique on your model and use case.
AWS PrivateLink	AWS PrivateLink provides private connectivity between virtual private clouds (VPCs), supported AWS services, and your on-premises networks without exposing your traffic to the public internet
Learning Rate	The Learning rate hyperparameter controls the step size at which a model's parameters are updated during training. It determines how quickly or slowly the model's parameters are updated during training. - A **high learning rate** means that the parameters are updated by a **large step size**, which can lead to **faster convergence** but may also cause the optimization process to overshoot the optimal solution and become unstable. - A **low learning rate** means that the parameters are updated by a **small step size**, which can lead to **more stable convergence** but at the cost of slower learning.
Supervised Learning	Supervised learning is a type of machine learning where the algorithm is trained on a **labeled dataset**, meaning the input data is paired with the correct output. The goal is for the algorithm to learn the mapping

	between input and output so it can accurately predict outcomes for new, unseen data.
Unsupervised Learning	Unsupervised learning involves training algorithms on **unlabeled data, without predefined outputs or correct answers.** The goal is for the algorithm to **discover hidden patterns, structures, or relationships** within the data on its own, often used for **clustering, dimensionality reduction, or anomaly detection**.
Semi-Supervised Learning	Semi-supervised learning is a hybrid approach that **combines elements of both supervised and unsupervised learning**, using a small amount of labeled data along with a larger amount of unlabeled data. This method aims to leverage the benefits of both approaches, improving model performance when fully labeled datasets are scarce or expensive to obtain.
Stop Sequences	Stop sequences are **specific tokens or phrases that instruct an AI model to cease generating text at a designated point,** such as the end of a sentence or list. They can enhance control over output by ensuring that the generated content does not exceed the desired length or format, allowing for more structured and concise responses.
Intelligent Document Processing (IDP)	Intelligent Document Processing (IDP) involves automating the process of manually entering data from paper-based documents or document images to integrate with other digital business processes.
Feature Extraction	Feature extraction is the technique of **creating new features by transforming or combining the original input features**. It aims to capture essential information in a lower-dimensional space, uncover hidden patterns, and improve model performance, particularly useful for complex, unstructured data like images or text.
Feature Selection	Feature selection is the process of **choosing a subset of the most relevant original features** from a dataset. It aims to reduce dimensionality, improve model interpretability, and decrease computational complexity while maintaining or improving model performance.
Multi-class Classification	Multiclass classification is a task where **each instance is assigned to one and only one class from three or more possible classes.** It deals with problems where the classes are mutually exclusive, and the goal is to predict a single class label for each input.
Multi-label Classification	Multilabel classification allows **each instance to be associated with multiple classes or labels simultaneously**. This approach is used when classes are not mutually exclusive, and the objective is to predict multiple class labels for each input, making it suitable for complex, real-world scenarios where items can belong to multiple categories.
Amazon Bedrock Custom Model Import (preview)	With the Custom Model Import feature, you can now bring your own custom models and use them seamlessly on Amazon Bedrock. Whether you've fine-tuned Meta Llama or Mistral AI models to suit your specific

	needs, or developed a proprietary model based on popular open architectures, you can now import those custom models and use them alongside the foundation models (FMs).
Amazon Bedrock Knowledge Bases	With Amazon Bedrock Knowledge Bases, you can give FMs and agents contextual information from your company's private data sources for RAG to deliver more relevant, accurate, and customized response

Sage Maker Performance Metrics by Category