WHIZLABS

AWS Certified AI Practitioner

Cheat Sheet

Quick Bytes for you before the exam!

*The information provided in the Cheat Sheet is for educational purposes only; created in our efforts to help aspirants prepare for the **AWS AI Practitioner—Beta Exam**. Though references have been taken from **AWS documentation**, it's not intended to be a substitute for the official docs. The document can be reused, reproduced, and printed in any form; ensure that appropriate sources are credited and required permissions are received.*

Index

—Back to Index—

Fundamentals of AI and ML

Artificial Intelligence

What is Artificial Intelligence?

- Artificial Intelligence (AI) is a field of computer science focused on developing systems that can exhibit intelligent behavior, such as reasoning, learning, and autonomous action.
- AI functions by combining large amounts of data with intelligent algorithms. These algorithms are trained on the data to learn patterns and make informed decisions. This process enables AI to perform tasks that require human-like intelligence, such as understanding language and driving vehicles.
- AWS offers a range of pre-built AI services, as well as customizable infrastructure options, designed to streamline AI development and reduce costs.

How does AI process information and make decisions?

Data Collection: AI systems require vast amounts of data to learn from. This data can be anything from images and text to numerical values.

Algorithm Selection: The appropriate algorithm is chosen based on the specific task the AI is designed to perform. Common algorithms include:

- **Machine Learning:** It involves training algorithms on large datasets to identify patterns and make predictions.
- **Deep Learning:** A subset of machine learning that uses artificial neural networks to learn from complex patterns.
- **Natural Language Processing (NLP):** AI can understand and respond to human language.
- **Training:** The algorithm is trained on the collected data. This involves feeding the data into the algorithm and allowing it to learn from it.
- **Testing:** The trained algorithm is tested on new data to evaluate its performance.
- **Deployment:** Once the algorithm has been tested and refined, it can be deployed in real-world applications.

Key components of AI application architecture:

Artificial intelligence architecture consists of three core layers. All the layers run on IT infrastructure that provides the necessary compute and memory resources for the AI to run.

Layer 1: Data layer

The data layer is the cornerstone of AI. It involves preparing and organizing data for use in machine learning, natural language processing, and image recognition technologies.

Layer 2: Model layer

The model layer focuses on the AI system's decision-making capabilities. Organizations often select pre-existing foundation models or large language models and customize them using techniques that incorporate relevant data.

Layer 3: Application layer

The application layer is the user-facing interface of the AI system. It allows users to interact with the AI, request specific tasks, generate information, or make data-driven decisions.

Applications of Artificial Intelligence:

Chatbots and smart assistants-

- Chatbots and smart assistants powered by AI are increasingly capable of conducting conversations that resemble human interactions.

Intelligent document processing-

- Intelligent document processing (IDP) transforms business documents such as emails, images, and PDFs into organized, structured information. IDP leverages AI technologies, including natural language processing (NLP), deep learning, and computer vision, to extract, categorize, and validate data.

Application performance monitoring-

- AI-driven APM tools leverage historical data to anticipate potential issues before they arise. Additionally, these tools can address problems in real time by offering developers practical solutions, ensuring smooth application operations, and resolving performance bottlenecks.

Predictive maintenance-

- AI-driven predictive maintenance leverages vast amounts of data to detect potential issues that could cause operational, system, or service interruptions. By anticipating problems before they arise, predictive maintenance helps businesses minimize downtime and avoid disruptions.

Medical research-

- AI technology in medical research streamlines automates repetitive tasks, and handles large datasets. It can be employed to enhance the entire pharmaceutical discovery and development process, transcribe medical records, and accelerate the time-to-market for new products.

Business Analytics-

- Business analytics leverages AI to gather, process, and examine intricate datasets. With AI-driven analytics, you can predict future trends, identify underlying causes within the data, and streamline time-consuming tasks.

Limitations of Artificial Intelligence in AWS:

While AWS offers a robust suite of AI and machine learning services, there are inherent limitations that users should be aware of:

1. **Data Quality and Quantity:**
   - **Bias:** If AI models are trained on biased data, it can lead to unfair or inaccurate results.
   - **Noise:** Incomplete or noisy data can hamper model accuracy and performance.
2. **Computational Resources:**
   - **Cost:** Training and running AI models can become expensive, especially for large-scale applications.
   - **Infrastructure:** Access to high-performance computing infrastructure may be required for complex models.
3. **Ethical Considerations:**
   - **Privacy:** Handling sensitive data raises privacy concerns, especially when using AI for tasks like facial recognition or natural language processing.
   - **Fairness:** AI models can perpetuate existing biases or discrimination if not designed and trained carefully.
4. **Human Oversight:**
   - **Dependency:** AI systems-related tools still require human oversight to ensure they are used ethically and effectively.
   - **Error Correction:** AI models can make mistakes, and human intervention may be necessary to correct errors or biases.

Basic AI Terminologies

What is Machine Learning?

Machine Learning (ML) is a subset of artificial intelligence (AI) focused on developing algorithms that improve automatically through experience and data. Simply put, machine learning allows computers to learn from data and make decisions or predictions without explicit programming.

What is Deep Learning?

Deep learning, a subset of artificial intelligence, enables computers to learn from data like human cognition. By analyzing complex patterns within images, text, audio, and other forms of data, deep learning models can provide accurate insights and predictions. These models can automate tasks that traditionally necessitate human intelligence, such as image description or transcription of audio files into text.

What is the Large Language Model [LLM])?

Large language models (LLMs) are sophisticated deep learning models trained on extensive text datasets. These models utilize a transformer architecture, a neural network framework composed of an encoder and decoder. The encoder and decoder leverage self-attention mechanisms to extract contextual meaning from text sequences, comprehending the relationships between words and phrases.

What is Responsible AI?

- ➔ **Broader Implications:** AI systems have substantial effects on individuals, communities, and the environment.
- ➔ **Ethical Priorities:** Organizations should emphasize fairness, transparency, and ethical considerations in AI practices.
- ➔ **Balancing Act:** Businesses must balance ethical AI practices with the pursuit of competitive advantage in a fast-evolving field.

What is Neural networks?

Artificial neural networks serve as the foundation for many artificial intelligence systems. Inspired by the human brain's structure and function, these networks employ interconnected computational units, often referred to as artificial neurons or nodes. Similar to biological neurons, these nodes process information through mathematical calculations. By working together in a network, these nodes can collectively solve complex problems and learn from data.

What is Natural Language Processing (NLP)?

Natural language processing (NLP) employs neural networks to extract meaning from textual data. It leverages computational methods designed to comprehend human language, enabling machines to process words, grammar, and sentence structure. This technology facilitates tasks such as document summarization, chatbot interactions, and sentiment analysis.

What is Computer vision?

Computer vision, powered by deep learning, enables computers to interpret and understand visual information from images and videos. This technology can be applied in various domains, such as content moderation to identify inappropriate images, facial recognition, and image classification. It is also crucial in autonomous vehicles, where it helps to perceive the surrounding environment and make timely decisions.

What is Speech Recognition?

Speech recognition technology leverages deep learning algorithms to decipher human speech, discern individual words, and comprehend meaning. Neural networks are employed to transcribe spoken language into written text and to gauge the emotional tone of the speaker. Speech recognition finds applications in various domains, such as virtual assistants and call center systems, where it is used to interpret spoken commands and execute corresponding actions.

What is Generative AI?

Generative AI is a type of artificial intelligence that can produce original content, including images, videos, text, and audio, based on textual instructions. Unlike traditional AI, which primarily analyzes data, generative AI utilizes deep learning techniques and vast datasets to create high-quality, human-like creative outputs.

While this technology opens up new possibilities for creative expression, there are also concerns regarding bias, harmful content, and intellectual property rights. In conclusion, generative AI marks a significant advancement in AI's ability to generate human-quality language and content.

Differences between AI, ML, Deep Learning & Gen AI

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)	Deep Learning	Generative AI
Definition	AI is a broad field of creating machines capable of performing tasks that typically require human intelligence.	ML is a subset of AI focused on systems that learn from data to make decisions.	DL is a subset of ML that uses neural networks with multiple layers (deep neural networks) to learn from large amounts of data.	Gen AI is a type of AI focused on generating new content, such as images, text, or music, based on learned patterns.
Types	Rule-based systems, expert systems, decision trees, ML, and DL.	Supervised, unsupervised, and reinforcement learning.	Convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers.	Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Large Language Models (LLMs).
Data Dependency	It can include rule-based systems with no learning component.	It relies on labeled or unlabeled data for learning patterns.	It requires large amounts of data to train deep neural networks effectively.	It requires large datasets for training to generate realistic outputs.
Applications	Robotics, natural language processing, expert systems, etc.	Predictive analytics, recommendation systems, fraud detection.	Image recognition, NLP, autonomous vehicles.	Image and video generation, text generation (e.g., chatbots), music composition.
Examples	Siri, autonomous robots, chess-playing computers.	Spam filters, personalized recommendations, weather forecasting.	Facial recognition systems, self-driving cars, and language translation services.	DALL-E for image generation, ChatGPT for text generation, and DeepDream for image creation.

—Back to Index—

Understanding Foundation Model

What Are Foundation Models?

Foundation models (FMs) are large-scale deep-learning neural networks trained on extensive datasets. They have revolutionized the approach data scientists take to machine learning (ML) by providing a starting point that speeds up the development of new AI applications. These models are designed to perform a broad range of general tasks, such as language understanding, text and image generation, and natural language processing (NLP).

Features of Foundation Models

Adaptability:

Foundation models are uniquely versatile and capable of executing a variety of tasks with high accuracy based on input prompts. This makes them significantly different from traditional ML models, which are typically designed for specific tasks like sentiment analysis, image classification, or trend forecasting.

General-Purpose Nature:

Due to their large size and broad training, foundation models can serve as base models for more specialized applications. Over the years, these models have grown in complexity and size, with models like BERT and GPT-4 showcasing this evolution.

Applications of Foundation Models

Language Processing:

Foundation models excel in natural language tasks, including answering questions, writing scripts, and translating languages.

Visual Comprehension:

These models are highly effective in computer vision, identifying images, generating images from text, and editing photos and videos.

Code Generation:

Foundation models can write and debug code in various programming languages based on natural language instructions.

Human-Centered Engagement:

They support decision-making processes, such as clinical diagnoses and analytics, by continuously learning from human inputs during inference.

Examples of Foundation Models

BERT (2018):

A bidirectional model trained on a vast dataset, capable of analyzing text and predicting sentences. It laid the groundwork for future models like GPT.

GPT (Generative Pre-trained Transformer):

Released by OpenAI, GPT models have evolved from GPT-1 with 117 million parameters to

GPT-4, which boasts 170 trillion parameters. These models are capable of tasks ranging from text generation to question answering.

Amazon Titan:

A foundation model from Amazon offers generative and embedding models for tasks like text summarization, information extraction, and personalization.

Challenges with Foundation Models

High Resource Demands:

Developing foundation models requires substantial infrastructure, making it costly and time-intensive.

Integration Complexity:

For practical use, these models must be integrated into software systems, which involves additional development for prompt engineering and fine-tuning.

Comprehension and Reliability Issues:

While foundation models can generate coherent responses, they may struggle with understanding context and can produce unreliable or biased answers.

AWS Support for Foundation Models

Amazon Bedrock:

This service simplifies the development and scaling of generative AI applications by offering access to foundation models via an API, allowing users to choose the most suitable model for their needs.

Amazon SageMaker JumpStart:

A hub for ML models and solutions, SageMaker JumpStart provides access to a wide range of foundation models, including popular ones like Llama 2 and Falcon, supporting the development of diverse AI applications.

Reference:

https://aws.amazon.com/what-is/foundation-models/

AI Models: Types

1. **Computer Vision Models**
   - **Amazon Rekognition**: This service provides pre-trained models for image and video analysis, including capabilities like object detection, facial recognition, and scene detection.
2. **Natural Language Processing (NLP) Models**
   - **Amazon Comprehend**: Used for analyzing text, Amazon Comprehend can perform sentiment analysis, entity recognition, and language detection.
   - **Amazon Translate**: Provides real-time translation services between different languages.
   - **Amazon Lex**: Powers conversational interfaces, enabling the creation of chatbots that can interact through voice and text.
   - **Amazon Polly**: Converts written text into lifelike speech in multiple languages.
3. **Speech Recognition Models**
   - **Amazon Transcribe**: This service converts speech into text, making it useful for transcriptions, subtitles, and more.
4. **Document Processing Models**
   - **Amazon Textract**: Extracts text, tables, and other data from scanned documents, making it easier to process and analyze paper-based information.
5. **Recommendation and Forecasting Models**
   - **Amazon Personalize**: Delivers personalized recommendations by analyzing user behavior and preferences.
   - **Amazon Forecast**: Utilizes time-series data to predict future trends, such as sales forecasts or inventory needs.
6. **Search and Information Retrieval Models**
   - **Amazon Kendra**: An enterprise search service that uses machine learning to provide relevant search results across documents and data sources.
7. **Custom Machine Learning Models**
   - **Amazon SageMaker**: A comprehensive platform for building, training, and deploying custom machine learning models. It supports a wide range of algorithms and frameworks.
8. **Generative AI Models**
   - **Amazon Bedrock**: A service that provides access to foundational models for generative AI, allowing users to create custom applications like text generation or image creation.
   - **SageMaker JumpStart**: Offers pre-trained models and solutions for generative AI tasks, which can be fine-tuned for specific needs.

9. Edge AI Models

- **AWS IoT Greengrass ML Inference**: Enables machine learning inference on edge devices, allowing models to be deployed in environments where real-time processing is critical.

10. Hybrid AI Models

- **Amazon Neptune ML**: Integrates machine learning with graph databases, enabling advanced data analysis and knowledge graph applications.

Machine Learning

What is Machine Learning?

- **Core Concept:** Machine learning revolves around creating algorithms that facilitate decision-making and predictions. These algorithms enhance their performance over time by processing more data.
- **Traditional vs. ML Programming:** Unlike traditional programming, where a computer follows predefined instructions, machine learning involves providing a set of examples (data) and a task. The computer then figures out how to accomplish the task based on these examples.
- **Example:** To teach a computer to recognize images of cats, we don't give it specific instructions. Instead, we provide thousands of cat images and let the machine learning algorithm identify common patterns and features. Over time, the algorithm improves and can recognize cats in new images it hasn't seen before.

Types of Machine Learning

Machine learning can be broadly classified into three types:

1. **Supervised Learning:** The algorithm is trained on labeled data, allowing it to make predictions based on input-output pairs.
2. **Unsupervised Learning:** The algorithm discovers patterns and relationships within unlabeled data.
3. **Reinforcement Learning:** The algorithm learns by trial and error, receiving feedback based on its actions.

Applications of Machine Learning

Machine learning powers many of today's technological advancements:

- **Voice Assistants:** Personal assistants like Siri and Alexa rely on ML to understand and respond to user queries.
- **Recommendation Systems:** Platforms like Netflix and Amazon use ML to suggest content and products based on user behaviour.
- **Self-Driving Cars:** Autonomous vehicles use ML to navigate and make real-time decisions.
- **Predictive Analytics:** Businesses use ML to forecast trends and make data-driven decisions.

—Back to Index—

ML Pipeline: Components with AWS Services

A Machine Learning (ML) pipeline in AWS refers to a structured workflow that automates the various stages involved in developing, training, and deploying machine learning models.

1. Data Collection

- **Amazon S3 (Simple Storage Service)**: Used to store large datasets. AWS provides secure and scalable storage for structured and unstructured data.
- **AWS Glue**: A data integration service that helps to discover, prepare, and combine data across multiple sources for analysis.
- **Amazon RDS (Relational Database Service)**: For storing and managing relational data that can be used for training ML models.

2. Exploratory Data Analysis (EDA)

- **Amazon SageMaker Studio**: Provides an integrated environment where data scientists can perform EDA using Jupyter notebooks. It supports visualization libraries like Matplotlib, Seaborn, and Pandas for statistical analysis and data exploration.
- **Amazon Athena**: An interactive query service that allows you to analyze data in Amazon S3 using SQL. Useful for quick analysis without the need to move data.

3. Data Pre-processing

- **AWS Glue and AWS Data Wrangler**: These tools help in cleaning, normalizing, and transforming raw data into a format suitable for modeling. This may involve handling missing values, normalization, and data scaling.
- **Amazon SageMaker Processing**: Allows running pre-processing jobs that can scale to handle large datasets.

4. Feature Engineering

- **Amazon SageMaker Feature Store**: A fully managed repository for storing, retrieving, and sharing features across different models and teams. It helps in automating the process of feature extraction and management.
- **Amazon SageMaker Data Wrangler**: Simplifies the process of feature transformation, enabling users to create new features by combining existing ones.

5. Model Training

- **Amazon SageMaker**: Supports training custom models using a wide variety of built-in algorithms or your own code. It also offers distributed training, enabling you to scale training jobs across multiple instances.
- **AWS Deep Learning AMIs**: Provides pre-configured environments with popular deep learning frameworks like TensorFlow, PyTorch, and Apache MXNet.

6. Hyperparameter Tuning

- **Amazon SageMaker Automatic Model Tuning**: Also known as hyperparameter optimization (HPO), this service automatically tunes the model's hyperparameters to improve performance, using techniques like Bayesian optimization.

7. Model Evaluation

- **Amazon SageMaker Debugger**: Offers insights into the training process by monitoring and profiling training jobs. It helps in identifying issues like overfitting and underfitting by analyzing training metrics.
- **Amazon SageMaker Model Monitor**: Used post-deployment to track model performance and detect data drift over time, ensuring that the model remains accurate.

8. Model Deployment

- **Amazon SageMaker Endpoint**: Allows you to deploy your trained models in real-time, making them accessible via API for inference.
- **Amazon Elastic Kubernetes Service (EKS)**: Supports deploying models in a Kubernetes-managed environment for larger, more complex applications.

9. Monitoring

- **Amazon CloudWatch**: Monitors deployed models in real-time, collecting and tracking metrics, logging, and triggering alerts for model performance or infrastructure issues.
- **Amazon SageMaker Model Monitor**: Continuously monitors deployed models for concept drift, data quality issues, and other anomalies that might affect the model's accuracy over time.

Fundamentals of ML Operations (MLOps)

MLOps in AWS is a set of practices that combine Machine Learning (ML) and DevOps to streamline the development, deployment, and management of ML models in the Amazon Web Services (AWS) cloud environment.

1. Experimentation

- **Rapid Prototyping**: AWS services like Amazon SageMaker allow data scientists to quickly build, test, and iterate on machine learning models using Jupyter notebooks and pre-built algorithms.
- **Experiment Tracking**: SageMaker Experiments helps in tracking and comparing different model runs, capturing parameters, configurations, and outcomes for better reproducibility and collaboration.

2. Repeatable Processes

- **Pipeline Automation**: SageMaker Pipelines automates the entire machine learning workflow, from data preparation to model deployment, ensuring that each step is repeatable and consistent.
- **Infrastructure as Code (IaC)**: Using AWS CloudFormation or Terraform, you can define and deploy infrastructure in a consistent and repeatable manner, ensuring that environments are identical across different stages.

3. Scalable Systems

- **Elastic Resources**: AWS provides scalable compute resources like EC2 instances and SageMaker-managed instances that can automatically scale up or down based on the workload, ensuring efficient use of resources.
- **Distributed Training**: SageMaker supports distributed training, allowing large-scale models to be trained faster across multiple GPUs or instances.

4. Managing Technical Debt

- **Version Control**: Versioning models, datasets, and code ensures that you can track changes, reproduce results, and avoid issues caused by outdated or inconsistent components.
- **Model Registry**: SageMaker Model Registry helps in managing different versions of models, storing metadata, and promoting models through various stages of development and production.

5. Achieving Production Readiness

- **Continuous Integration/Continuous Deployment (CI/CD)**: Implementing CI/CD pipelines with AWS CodePipeline or Jenkins integrates code changes, tests, and deployments seamlessly, ensuring models are always production-ready.

- **Security and Compliance**: AWS provides tools like AWS Identity and Access Management (IAM) and AWS Key Management Service (KMS) to secure data, models, and pipelines, ensuring compliance with industry standards.

6. Model Monitoring

- **Performance Monitoring**: SageMaker Model Monitor automatically monitors deployed models for accuracy and performance drift, alerting teams to any issues that may require attention.
- **Logging and Analytics**: AWS CloudWatch and AWS X-Ray can be used to log model predictions, track performance metrics, and diagnose issues in real time.

7. Model Re-training

- **Automated Retraining**: SageMaker Pipelines and Step Functions can automate the retraining process when a model's performance drops or new data becomes available.
- **Data Drift Detection**: Monitoring tools like SageMaker Model Monitor can detect when input data distribution shifts, triggering a model retraining pipeline to ensure the model remains accurate.

8. Scalability and Flexibility

- **Scalable Deployment**: SageMaker endpoints can automatically scale to handle increasing traffic, ensuring that the model can serve predictions efficiently regardless of load.
- **Multi-Model Endpoints**: Allows deploying multiple models on a single endpoint, optimizing resource utilization and reducing costs.

9. Collaboration and Governance

- **Collaboration Tools**: SageMaker Studio provides a unified interface where data scientists and engineers can collaborate, share experiments, and work on models together.
- **Governance and Auditing**: AWS provides tools to maintain governance, such as SageMaker Clarify for bias detection and SageMaker Model Monitor for ensuring model compliance with business rules.

10. Technical Debt Management

- **Artifact Management**: Using services like S3 for storing datasets, models, and logs helps in managing and organizing artifacts efficiently, reducing the technical debt associated with disorganized resources.
- **Code Reusability**: Utilizing modular code and standardized practices across teams minimizes redundant work and accelerates future projects.

Amazon SageMaker

What is Amazon SageMaker?

- Amazon SageMaker is a comprehensive platform that empowers users to develop, train, and deploy machine learning models efficiently. This fully managed service offers a wide range of tools, including notebooks, debuggers, profilers, pipelines, and MLOps capabilities, to streamline the entire ML lifecycle.
- Amazon SageMaker offers a range of pre-built tools, including algorithms, pre-trained models, and solution templates, to expedite the development and deployment of machine learning models to help data scientists and practitioners.

Algorithm Selection:

Problem Type	Appropriate Algorithm
Binary Classification	Logistic Regression, XGBoost, etc.
Multiclass Classification	XGBoost, Linear Learner, etc.
Regression	Linear Learner, XGBoost, etc.
Object Detection	Faster R-CNN, SSD, etc.
Anomaly Detection	Random Cut Forest, etc.
Clustering	K-Means, DBSCAN, etc.
Topic Modeling	Latent Dirichlet Allocation (LDA)
Recommender Systems	Factorization Machines, etc.

Features:

Prepare Data -

- **SageMaker Feature Store:-** Amazon SageMaker Feature Store is a centralized platform designed to store, share, and manage features used in machine learning models. Features are the data inputs that models rely on during both training and inference.
- **SageMaker Data Wrangler:-** Amazon SageMaker Data Wrangler selects, understands, and transforms data to prepare it for machine learning (ML) in minutes. reduces data prep time for tabular, image, and text data from weeks to minutes. It enables a rapid assessment of ML model accuracy and helps identify potential problems before deployment.

—Back to Index—

- Geospatial ML with Amazon SageMaker:- Amazon SageMaker empowers data scientists and ML engineers to build, train, and deploy ML models using geospatial data such as satellite imagery, maps, and location data.

Build -

- **SageMaker Notebooks:-** Amazon SageMaker Notebooks offer a fully managed Jupyter environment, enabling data scientists and ML engineers to explore, analyze, and develop machine learning models efficiently.
- **SageMaker Jumpstart:-** Amazon SageMaker JumpStart is a machine learning (ML) hub that can help you quickly evaluate, compare, and select Foundation models based on pre-defined quality and responsibility metrics to perform tasks like article summarization and image generation.
- **SageMaker Studio Lab:-** Amazon SageMaker Studio Lab is a free service based on open-source JupyterLab that allows customers to use AWS compute resources to create and run their Jupyter notebooks.

Train -

- **SageMaker Model Training:-** Amazon SageMaker Model Training streamlines the process of training and tuning machine learning models, significantly reducing time and costs while eliminating the need for infrastructure management.
- **SageMaker Experiments:-** Amazon SageMaker offers a managed MLflow capability that simplifies machine learning and generative AI experimentation. Data scientists can easily use MLflow within SageMaker for model training, registration, and deployment. Administrators can quickly establish secure and scalable MLflow environments on AWS.
- **SageMaker HyperPod:-** Amazon SageMaker HyperPod simplifies the process of building and optimizing ML infrastructure for training foundation models, significantly reducing training time by 40%. By automatically distributing training workloads across thousands of accelerators, HyperPod enables parallel processing and accelerates model performance.

Deploy -

SageMaker Model Deployment:- Amazon SageMaker simplifies the deployment of machine learning models, including foundation models, offering optimal cost-effectiveness for inference requests across various applications.

- **SageMaker Pipelines:-** Amazon SageMaker Pipelines is a serverless workflow orchestration service that automates machine learning (ML) and large language model (LLM) workflows.

End-to-End ML -

- **SageMaker MLOps:-** Amazon SageMaker offers specialized tools for managing machine learning operations (MLOps), streamlining and standardizing processes throughout the machine learning lifecycle.
- **SageMaker Canvas:-** Amazon SageMaker Canvas offers a visual interface that simplifies the machine learning process. It allows you to prepare data, build, and deploy ML models efficiently.
- **SageMaker Studio:-** Amazon SageMaker Studio provides a comprehensive suite of tools for the entire machine learning development steps, including data preparation, model building, training, deployment, and management.

Feature	SageMaker Canvas	SageMaker Studio
Target Audience	Data scientists and ML engineers with limited coding experience	Data scientists and ML engineers with advanced coding skills
Interface	Visual, no-code interface	Integrated development environment (IDE)
Model Building	Automated model selection and training	Manual model selection and training using various algorithms
MLOps	Basic MLOps features (monitoring, versioning)	Advanced MLOps capabilities (pipeline creation, experiment tracking)
Use Cases	Rapid prototyping, exploratory data analysis, simple ML models	Complex ML models, custom pipelines, research projects

SageMaker Ground Truth:- Amazon SageMaker Ground Truth provides a robust platform for incorporating human expertise into the machine learning process, enhancing model performance through continuous feedback.

ML Governance -

- **ML Governance with SageMaker:-** Amazon SageMaker offers specialized governance features to ensure responsible machine-learning practices. Amazon SageMaker Role Manager allows administrators to quickly establish necessary permissions.
- **SageMaker Clarify:-** SageMaker Clarify streamlines the process of identifying potential biases in your dataset. Specify the input features you're concerned about, like gender or age, and SageMaker Clarify will conduct a thorough analysis to uncover any potential biases present in those features.

Fundamentals of Generative AI

Generative AI

What is Generative AI?

- Generative AI is a form of artificial intelligence capable of generating original content, such as text, visuals, and audio.
- You can train it to understand and generate text in human and programming languages. It can also learn complex subjects like art, chemistry, biology, etc. By leveraging past training data, it can apply its knowledge to new and unfamiliar tasks.
- Using AWS, you can rapidly develop and expand generative AI applications tailored to your specific data, use cases, and customer needs. Benefit from enterprise-grade security and privacy, access to cutting-edge foundational models, and a data-centric approach.

How does Generative AI work?

Generative AI utilizes machine learning models, which are extensively trained on massive datasets, to produce new content.

Foundation models:- Foundation models are machine learning models trained on massive, diverse datasets without specific task labels. They are capable of performing a wide variety of general tasks.

Large language models:- LLMs are one class of FMs. These models are capable of performing a variety of linguistic functions, including summarization, text generation, classification, open-ended dialogue, and information extraction.

Benefits of Generative AI:

- Accelerates research
- Enhances customer experience
- Optimizes business processes
- Boosts employee productivity

Generative AI Models:

Diffusion models:- Diffusion models create new data by gradually introducing noise to existing data samples and then carefully removing it. This process involves adding controlled random changes to the original data over multiple iterations. The model ensures that the generated data remains coherent and realistic by carefully controlling the noise level.

Once the data has been sufficiently noised, the diffusion model reverses the process. It gradually removes the noise, step by step, until it produces a new data sample that closely resembles the original. This reverse denoising process allows the model to generate new data that is consistent with the underlying data distribution.

Generative Adversarial Networks (GAN):- GANs are widely used in generating realistic images, transforming visual styles, and data augmentation tasks.
Generative Adversarial Networks (GANs) employ a competitive training process between two neural networks. The generator network produces fake data by introducing random noise, while the discriminator network attempts to differentiate between this fake data and real data.
As training progresses, the generator refines its ability to create increasingly realistic data, while the discriminator becomes more adept at identifying fake data. This iterative process continues until the generator generates data that is indistinguishable from authentic data, even to the discriminator.
Variational Auto-encoders:- VAEs create a compressed representation of data, often referred to as a latent space. This latent space is a mathematical construct that captures the essence of the data. Imagine it as a unique code, summarizing the key features of the data.
For example, if studying faces, the latent space contains numbers representing eye shape, nose shape, cheekbones, and ears.
VAEs use two neural networks. The encoder network transforms input data into a mean and variance for each dimension of the latent space. A random sample is drawn from a Gaussian distribution using these parameters, resulting in a latent space representation. This compressed representation serves as a simplified version of the input. The decoder network takes this latent sample and attempts to reconstruct the original data. The quality of the reconstruction is evaluated using mathematical metrics that compare the reconstructed output to the original input.
Transformer-based models:- The transformer-based generative AI model builds upon the encoder and decoder concepts of VAEs.
Transformer-based models use a self-attention mechanism. Self-attention helps enable generative models to prioritize significant words during processing. Transformer-based models leverage multiple multiple encoder layers called attention heads to identify diverse interconnections between words. Each head focuses on distinct sections of the input sequence, facilitating a comprehensive analysis of the data.

Tools for building generative AI applications:

- Amazon Bedrock
- Amazon SageMaker
- AWS Trainium
- AWS Inferentia
- Amazon EC2 P5 Instances
- Amazon EC2 UltraClusters

Generative AI-powered applications:

- Amazon Q
- PartyRock - Amazon Bedrock Playground
- AWS HealthScribe

Use Cases of Generative AI:

- Chatbots and Virtual Assistants
- Conversational Analytics
- Employee Assistant
- Code Generation
- Personalization
- Productivity and Creativity solution templates

Limitations of Generative AI on AWS:

Security:- The use of proprietary data to train generative AI models raises concerns about data privacy and security. It's crucial to implement measures that protect sensitive information and prevent unauthorized access.

Creativity:- While generative AI can produce creative content, its outputs are often constrained by the data it has been trained on. Human creativity, which involves deeper understanding, emotional resonance, and original thought, remains a challenge for AI to fully replicate.

Cost:- Training and running generative AI models require significant computational resources. Cloud-based solutions offer a more accessible and cost-effective approach compared to building models from scratch.

Explainability:- Generative AI models are often considered "black boxes," making it difficult to understand their decision-making processes. Improving their interpretability and transparency is essential to build trust and promote wider adoption.

—Back to Index—

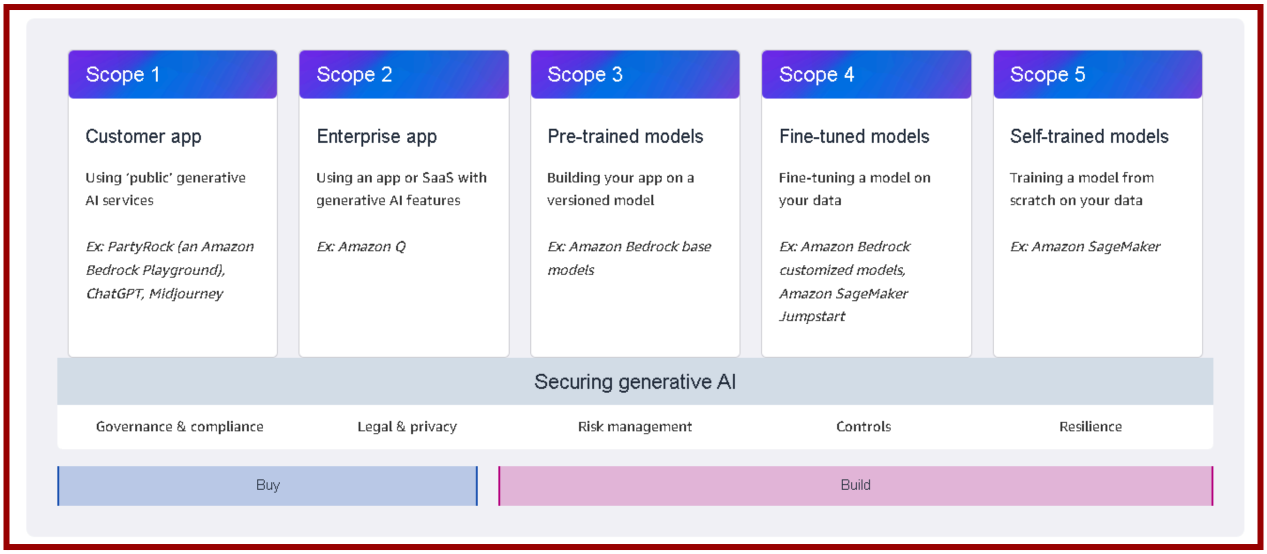
GenAI Security Scoping Matrix

A Generative AI Security Scoping Matrix offers a structured approach for organizations to evaluate and implement security measures across the entire lifecycle of AI applications. By categorizing security concerns, it provides a targeted framework for safeguarding AI systems.

- AWS provides multiple services to secure Generative AI workloads. AWS services vary significantly in their underlying infrastructure, software, access mechanisms, and data handling. To simplify security management, we've organized these services into logical categories called 'scopes'.

Scoping (Determine your scope):

- To begin, you'll need to determine which scope your use case fits into.
- The scopes are numbered 1–5, representing the least ownership to greatest ownership your organization has over the AI model and its associated data.



Scope 1	Scope 2	Scope 3	Scope 4	Scope 5
Customer app	Enterprise app	Pre-trained models	Fine-tuned models	Self-trained models
Using 'public' generative AI services	Using an app or SaaS with generative AI features	Building your app on a versioned model	Fine-tuning a model on your data	Training a model from scratch on your data
Ex: PartyRock (an Amazon Bedrock Playground), ChatGPT, Midjourney	*Ex: Amazon Q*	*Ex: Amazon Bedrock base models*	*Ex: Amazon Bedrock customized models, Amazon SageMaker Jumpstart*	*Ex: Amazon SageMaker*

Securing generative AI

Governance & compliance · Legal & privacy · Risk management · Controls · Resilience

Buy · Build

Figure 1: *Generative AI Security Scoping Matrix*

Buying generative AI:

- **Scope 1:** Consumer app – Your business consumes a public third-party generative AI service, that is either free or paid. At this scope, you do not have ownership or access to the underlying training data or model. You can only interact with the service through its provided APIs or applications, adhering to the provider's terms of use.
  Example: A worker uses a generative AI chatbot to brainstorm marketing campaign concepts.
- **Scope 2:** Enterprise app – Your business uses a third-party enterprise application that has generative AI capabilities, and a business relationship is established between your organization and the vendor.

Example: You use a third-party enterprise scheduling application that has a generative AI capability embedded within to help draft meeting agendas.

- **Scope 3:** Pre-trained models – Your business utilizes an existing third-party generative AI foundation model to power its application. This model is accessed and integrated into your operations via an application programming interface.
  Example: A customer support chatbot was developed utilizing the Anthropic Claude foundation model, accessed via the Amazon Bedrock API.
- **Scope 4:** Fine-tuned models – Your business refines an existing third-party generative AI foundation model by fine-tuning it with data specific to your business, generating a new, enhanced model that's specialized to your workload.
  Example: By leveraging a foundation model through an API, you can create a marketing application that tailors promotional materials specifically to your products and services.
- **Scope 5:** Self-trained models – Your business builds and trains a generative AI model from scratch using data that you own or acquire. You own every aspect of the model.
  Example: Your business wants to create a model trained exclusively on deep, industry-specific data to license companies in that industry, creating a completely novel LLM.

By identifying the specific applications of generative AI, security teams can prioritize their efforts and evaluate the potential risks within each security domain.

Let's examine how scoping influences security requirements within each security discipline.

- **Governance and compliance –** Implementing effective policies, procedures, and reporting mechanisms can enable businesses to operate efficiently while mitigating risks.
- **Legal and privacy –** The specific legal, regulatory, and privacy requirements for using or creating generative AI solutions.
- **Risk management –** Assessing risks associated with generative AI and proposing countermeasures.
- **Controls –** Implementing security measures to reduce risk.
- Resilience – Designing reliable generative AI systems that consistently meet business SLAs.

Amazon SageMaker JumpStart

Amazon SageMaker JumpStart is a machine learning hub that can speed up your ML development. Using SageMaker JumpStart, you can select, evaluate, and compare FMs quickly based on pre-defined quality and responsibility metrics to perform tasks like image generation and article summarization.

Features:

Foundation Models:- Discover a variety of foundational models from leading providers like AI21 Labs, Databricks, Hugging Face, Meta, Mistral AI, Stability AI, and Alexa. These models can be used to accomplish a wide range of tasks including summarizing articles and generating text, images, or videos.

Built-in algorithms:- You can utilize built-in solution templates through the SageMaker Python SDK. These algorithms address common ML tasks, including image, text, and tabular data classification, as well as sentiment analysis.

Prebuilt solutions:- SageMaker JumpStart offers pre-built, end-to-end solutions for common machine learning applications like demand forecasting, credit risk assessment, fraud detection, and computer vision.

Benefits of SageMaker JumpStart:

Publicly available foundation models

Built-in ML algorithms

Customizable solutions

Support collaboration

Use cases and Advantages:

1. Foundation Model Integration

- Deploy models like LLaMA 2 and Stable Diffusion in VPC mode, even without internet.
- Access pre-trained models for easy deployment and tuning.

2. Large Language Models (LLMs)

- Simplifies deploying and tuning LLMs, including 40M parameter models for NLP tasks.

3. Text Classification

- Pre-built models for text classification with customization options.

4. Image Generation

- Deploy Stable Diffusion XL for high-quality image generation.

5. No-Code Solutions

- Fast, no-code deployment for quick AI solutions, accessible to non-experts.

6. Learning Resources

- Video tutorials and guides for easy model deployment and tuning.

Amazon Bedrock

- Amazon Bedrock is a managed serverless service providing various high-performing foundation models (FMs) from top AI companies like AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon.
- These models are accessible via a unified API for creating generative AI applications focusing on security, privacy, and responsible AI practices.
- With Amazon Bedrock, you can quickly test and compare different foundation models to find the best fit for your use case.
- These models can be tailored to your unique data using techniques like fine-tuning and Retrieval Augmented Generation (RAG).
- Additionally, you can create agents that can perform tasks using your company's systems and information.

How does Amazon Bedrock help to build generative AI applications?

- **Model Choice -** Choose from a range of leading FMs: Amazon Bedrock's single API lets you easily switch between different foundation models and their updates.
- **Customization -** Privately adapt models with your data: Model customization lets you deliver differentiated and personalized user experiences. Fine-tune foundation models with your data to create unique, personalized experiences.
- **RAG -** Deliver more relevant FM responses: To provide FMs with relevant company data, organizations use RAG. This technique feeds data into prompts to improve responses.
- **Agents -** Execute complex tasks across company systems: Amazon Bedrock agents automate complex tasks using your company's systems and data. Agents analyze requests, execute relevant APIs, and provide secure, private responses.

Amazon Bedrock offers models in 3 states:

- **Active**: The model provider is actively developing this version, and it will continue to be updated with bug fixes and minor improvements.
- **Legacy:** A version is marked as a legacy when a more advanced version delivers superior results. Amazon Bedrock determines an EOL date for outdated versions.
- **EOL:** This version is outdated and inoperable. Requests made to it will fail.

Use cases:

- **Text generation -** Produce unique content for your blog, social media, and web pages
- **Virtual assistants -** Build assistants that understand user inquiries, automatically divide tasks, interact conversationally to gather necessary details, and execute actions to complete the requested task.
- **Text and image search -** Identify and compile relevant information to answer questions and provide recommendations based on a large body of textual and visual data.

—Back to Index—

- **Text summarization -** Acquire concise summaries of extensive documents, such as articles, reports, research papers, technical documentation, and even books, to effectively extract essential information.
- **Image generation -** Generate lifelike and visually engaging images for advertising campaigns, websites, presentations, and other applications.

Amazon Bedrock Agents:

Amazon Bedrock Agents enable you to develop and configure autonomous agents for your application.

Features:

- Amazon Bedrock Agents securely access your company's data, enhance user requests with relevant information, and provide accurate responses.
- Amazon Bedrock Agents orchestrate and analyze the tasks, dividing them into the appropriate logical order using the FM's reasoning capabilities.
- Amazon Bedrock Agents enable the dynamic generation and execution of code in a secure environment. This automates complex analytical queries previously difficult to address using model reasoning alone.
- Amazon Bedrock Agents enable you to incorporate business logic into your chosen backend service. Moreover, the return of control functionality empowers you to execute time-consuming actions in the background (asynchronously) while continuing the orchestration flow.
- Amazon Bedrock Agents possess the ability to maintain memory across interactions, allowing them to remember historical conversations and improve the accuracy of multi-step tasks.

Amazon Bedrock Guardrails:

Amazon Bedrock Guardrails helps you establish safeguards for generative AI applications aligned with your specific use cases and responsible AI policies.

Amazon Bedrock Guardrails offer additional customizable protections beyond the built-in safeguards of foundation models.

Amazon Bedrock Guardrails protects your generative AI applications by evaluating both user prompts and model responses.

- Blocking up to 85% more harmful content.
- Filtering out over 75% of inaccurate responses for RAG and summarization tasks.
- Customers can customize and implement safety, privacy, and truthfulness safeguards within a unified solution.

Features:

- Amazon Bedrock Guardrails can be combined with Amazon Bedrock Agents and Knowledge Bases to create generative AI applications that adhere to your responsible AI policies.
- Customers can establish multiple guardrails, each tailored with a different combination of controls, and apply these guardrails to various applications and use cases.
- Amazon Bedrock Guardrails offers customizable content filters to screen for harmful content, including hate speech, insults, sexually suggestive material, violence, misconduct (including criminal activity), and prompt attacks (prompt injection and jailbreaking).
- Amazon Bedrock Guardrails employs **contextual grounding checks** to identify and filter hallucinations when responses deviate from the provided information, such as being factually incorrect or introducing new data.

PartyRock - Amazon Bedrock Playground:

PartyRock is a powerful tool designed to let you explore and experiment with the various foundation models available on the Amazon Bedrock platform. **It** is designed specifically for entertainment and creativity.

Features:

- **Chat playground -** The chat playground allows you to interact with the conversational models available on Amazon Bedrock. When you enter a prompt into the model
- **Text playground -** The text playground offers a platform to explore Amazon Bedrock's text models. By inputting a text prompt, you can see the model's generated output.
- **Image playground -** The image playground allows you to explore the capabilities of Amazon Bedrock's image models. By entering a text description, you can see how the model transforms your words into a visual representation.

Amazon Q

Amazon Q is a generative AI-powered assistant that helps accelerate software development and uses companies' internal data.

Amazon Q Business is a generative AI–powered assistant that can provide answers, summaries, content generation, and secure task completion based on your enterprise data.

Amazon Q Developer supports developers and IT professionals in various tasks, including coding, testing, application updates, error diagnosis, security assessments, and AWS resource optimization.

Features:

Amazon Q offers advanced planning and reasoning abilities to transform and implement new code features as requested by developers.

Amazon Q is capable of understanding and respecting current governance identities, roles, and permissions, providing personalized interactions.

Amazon Q integrates with:

- **Amazon QuickSight**
- **Amazon Connect**
- **AWS Supply Chain**

—Back to Index—

Applications of Foundation Models

Prompt Engineering

- **Prompts** are a specific set of inputs provided by the user that direct the LLMs on Amazon Bedrock to produce relevant outputs.
- **Prompt engineering** is the process of designing text prompts to obtain desired responses from a Large Language Model (LLM). It is the art of communicating with an LLM.
- Prompt engineers use trial and error methods to generate input texts that guide an application's generative AI to work as expected.

Benefits of Prompt Engineering:

- **Greater developer control**
- **Improved user experience**
- **Increased flexibility**

Prompt Engineering Techniques:

- **Chain-of-thought prompting:-** Chain-of-thought prompting involves dividing complex questions into smaller, logical steps, similar to a thought process. This technique enhances the model's ability to reason and solve problems effectively.
- **Tree-of-thought prompting:-** The tree-of-thought technique extends chain-of-thought prompting by generating multiple potential next steps and evaluating each using a tree search approach.
- **Maieutic prompting:-** Maieutic prompting is similar to tree-of-thought prompting. The model is asked to answer a question with an explanation, followed by prompts to explain parts of the explanation. Inconsistent explanations are eliminated, improving performance in complex commonsense reasoning.
- **Complexity-based prompting:-** This prompt-engineering method utilizes multiple chain-of-thought rollouts, selecting those with the longest thought sequences and the most frequent conclusions.
- **Generated knowledge prompting:-** This technique involves generating relevant facts to support the prompt, and then completing the prompt. This strategy often results in higher-quality completions as the model is guided by relevant information.
- **Least-to-most prompting:-** This method involves prompting the model to identify and address subproblems sequentially. This allows subsequent subproblems to benefit from solutions to earlier ones.
- **Self-refine prompting:- I**n this approach, the model is prompted to solve a problem, evaluate its own solution, and then improve upon it by considering the original problem, its solution, and its evaluation. This cycle repeats until a predetermined reason to stop.

- **Directional-stimulus prompting:-** This prompt engineering method employs a hint or cue, like specified keywords, to direct the language model toward the intended outcome.

Use-Cases:

Subject matter expertise:

- **Example:-** Imagine a doctor using a language model to generate potential diagnoses for a complex patient. By inputting symptoms and patient details, the AI, guided by carefully crafted prompts, can list possible diseases and narrow down the options based on additional information.

Critical thinking:

- **Example:-** In decision-making, a model could be prompted to evaluate various options, weigh their pros and cons, and suggest the most suitable course of action.

Creativity:

- **Example:-** Writers can use prompt-engineered models to brainstorm ideas for stories, generating characters, settings, and plot points. Graphic designers might employ these models to generate color palettes that convey specific emotions and then create designs using those palettes.

—Back to Index—

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is a process to enhance the capabilities of large language models by referencing an authoritative knowledge base beyond its original training data, resulting in more accurate and informative responses.

- Large Language Models (LLMs) are trained on massive datasets with billions of parameters to generate original output for tasks such as answering questions, translating languages, and completing sentences. Retrieval Augmented Generation (RAG) enhances the capabilities of LLMs by integrating them with domain-specific or organizational knowledge bases.
- This approach avoids the need for model retraining and provides a cost-effective way to ensure that LLM outputs remain accurate, relevant, and valuable in diverse contexts.

Benefits of Retrieval Augmented Generation:

- **Cost-effective implementation**
- **Current information**
- **Enhanced user trust**
- **More developer control**

How does Retrieval Augmented Generation work?

LLM incorporates the new knowledge and its training data to create better responses.

- **Create external data-**
  - The new data outside the LLM's original training data set is called *external data*. This data can come from various sources like APIs, databases, or document repositories and may exist in different formats such as files, database records, or long-form text.
  - Another AI technique, embedding language models, transforms textual data into numerical representations stored in vector databases, creating a knowledge library that the generative AI models can understand.
- **Retrieve relevant information-**
  - The next step is to perform a relevancy search. User queries are also converted into vectors to retrieve relevant information. The relevancy was calculated and established using mathematical vector calculations and representations.
- **Augment the LLM prompt-**
  - The RAG model can generate more accurate and informative responses to user queries by integrating this retrieved data into the prompts provided by the large language model. This technique, known as prompt engineering, facilitates effective communication between the user and the AI.
- **Update external data-**

○ To ensure that retrieved information is up-to-date, it's essential to refresh the documents and update their corresponding embeddings asynchronously. This can be achieved through automated processes that happen in real time or by scheduling regular updates.

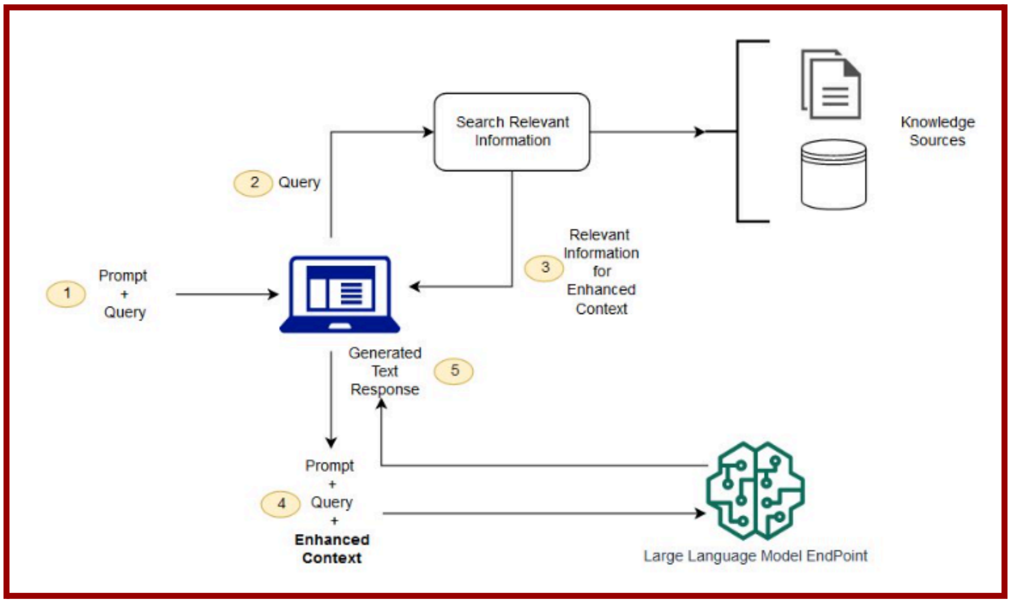

Search Relevant Information
Knowledge Sources
2 Query
1 Prompt + Query
3 Relevant Information for Enhanced Context
Generated Text Response 5
4 Prompt + Query + Enhanced Context
Large Language Model EndPoint

Figure: Retrieval Augmented Generation

RLHF - Reinforcement Learning from Human Feedback

- Reinforcement learning from human feedback (RLHF) is a machine-learning technique that enhances ML model performance by incorporating human feedback.
- Reinforcement learning (RL) techniques train software to make decisions that maximize rewards, making their outcomes more accurate.

How does RLHF work?

RLHF involves a four-step process to prepare a model ready for deployment.

- **Data collection:-** A collection of human-written prompts and their respective answers is established to serve as training material for the language model.
- **Supervised fine-tuning of a language model:-** You can use a commercial pre-trained model as the base model for the RLHF process. You can fine-tune the model to the company's internal knowledge base using the retrieval-augmented generation (RAG) technique. After fine-tuning, evaluate the model's output against predetermined prompts by comparing its responses to human-generated examples collected in the initial stage.
- **Building a separate reward model:-** From a set of multiple responses from the model answering the single prompt, human evaluators can express their preferences for each option. By analyzing these ratings, we construct a reward model capable of automatically predicting how a human would score any given response.
- **Optimize the language model with the reward-based model:-** The language model uses the reward model to iteratively refine its response generation strategy. By internally assessing various potential responses, the model selects the one that it predicts will yield the highest reward according to the reward model's criteria. This adaptive approach ensures that the model's outputs align more closely with human preferences.

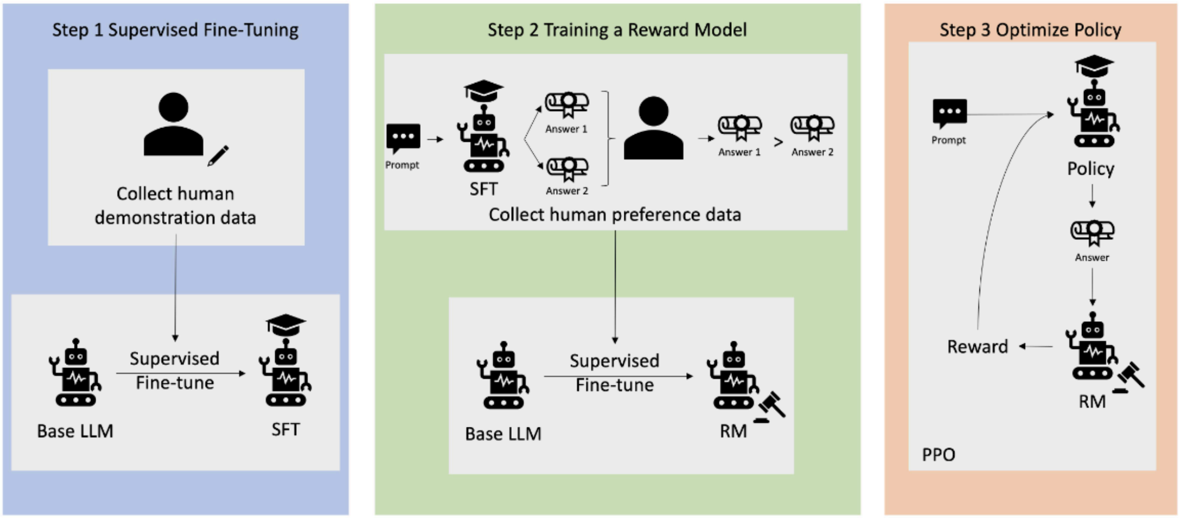


Figure: RLHF learning process

Applications of RLHF:

- RLHF can be used in AI image generation: for example, gauging the degree of realism, technicality, or mood of artwork
- RLHF can generate music that aligns with specific emotions or create soundtracks.
- RLHF can enhance a voice assistant's tone, making it more approachable, curious, and reliable.

How can AWS services be utilized to fulfill RLHF requirements?

Amazon SageMaker Ground Truth simplifies the task of labeling and annotating data for Reinforcement Learning from Human Feedback (RLHF), guaranteeing accurate and reliable human input.

Recall-Oriented Understudy for Gisting Evaluation [ROUGE]:

- ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a suite of metrics commonly used in natural language processing (NLP) to evaluate the quality of machine-generated text.
- These metrics primarily focus on comparing the generated text and the ground truth reference (human-written) text of a validation dataset.
- ROUGE measures are designed to assess various aspects of text similarity, such as the precision and recall of n-grams (contiguous sequences of words) in system-generated and reference texts.
- The goal is to determine how effectively the model can generate text that is similar to the original content.

Guidelines for Responsible AI

Responsible AI

Responsible AI refers to the development of AI systems that are fair, transparent, accountable, safe, and unbiased.

Components of responsible AI:

- **Fairness -** Considering the potential effects on diverse groups.
- **Explainability -** Understanding and assessing system outputs.
- **Privacy and security -** Obtaining, utilizing, and securing data and models.
- **Safety -** Safeguarding against system failures and malicious use.
- **Controllability -** Establishing mechanisms to monitor and control AI actions.
- **Veracity and robustness -** Achieving accurate system outputs, under both normal and adverse conditions.
- **Governance -** Promoting responsible AI by integrating best practices across the AI supply chain.
- **Transparency -** Ensuring stakeholders have the knowledge required to engage effectively with the AI system.

Services and tools to build Responsible AI:

Foundation model (FM) evaluations:-

- Model Evaluation on Amazon Bedrock
- Amazon SageMaker Clarify

Implementing safeguards in generative AI:-

- **Guardrails in Amazon Bedrock**

Detecting bias:-

Biases are imbalances in data or disparities in the performance of a model across different groups.

- **Amazon SageMaker Clarify**

Explaining model predictions:-

- **Amazon SageMaker Clarify**

Monitoring and human review

- Amazon SageMaker Model Monitor
- Amazon Augmented AI

Improving governance

- ML Governance from Amazon SageMaker

—Back to Index—

Amazon SageMaker Clarify

Machine learning offers opportunities to identify and measure bias throughout the ML lifecycle. **Amazon SageMaker Clarify** helps detect bias in data and models before, during, and after training:

1. **Pre-Training Bias:** Detect bias in the raw data before model training begins.
2. **Post-Training Bias:** Measure bias in the model's outputs after training.
3. **Monitoring Bias:** Continuously monitor bias in model predictions after deployment.

Benefits of SageMaker Clarify:

- Evaluate foundation models (FMs) in minutes:- Automate the evaluation of foundation models for your generative AI applications based on criteria such as accuracy, resilience, and bias to uphold responsible AI principles.
- Build trust in ML models:- Assess your FM's performance during customization using both automated and human-based methods.
- Accessible, science-based metrics and reports:- Generate user-friendly metrics, reports, and practical examples to support the FM customization and MLOps workflow.

SageMaker Clarify's Strategy for Addressing Bias

- **Bias Metrics**: SageMaker Clarify offers model-agnostic metrics to measure bias and fairness based on different fairness concepts.
- **Automation**: SageMaker Clarify automates bias detection and monitoring throughout the ML lifecycle.
- **Data Monitoring**: SageMaker Clarify tracks bias in model predictions after deployment, ensuring continuous oversight of model behavior.
- **Tools for Bias Detection**
- SageMaker Clarify **Sample Notebooks:** SageMaker Clarify provides a notebook for bias detection and explainability, helping users run bias detection jobs and interpret feature attributions.

The sample notebook for bias detection can run in **Amazon SageMaker Studio** using **Python 3 (Data Science)**. It walks through the process of detecting bias and explaining model predictions. **SageMaker Clarify** offers comprehensive tools to measure and monitor bias, helping ensure machine learning models are fair and unbiased across all development and deployment stages.

Amazon SageMaker Model Monitor

Amazon SageMaker Model Monitor monitors the quality of Amazon SageMaker machine learning models in production.

Model Monitor empowers you to implement continuous monitoring using various approaches;

- **Continuous monitoring with a real-time endpoint.**
- **Continuous monitoring with a batch transform job that runs regularly.**
- **On-schedule monitoring for asynchronous batch transform jobs.**
    - Model Monitor empowers you to establish alerts that notify model quality deviations. By promptly identifying these deviations, you can take corrective actions such as retraining models, auditing upstream systems, or fixing quality issues.
    - Model Monitor offers both pre-built and customizable monitoring options. You can use the pre-built features for quick setup or write your code for custom analysis.

Model Monitor offers various kinds of monitoring:

- **Monitor data quality -** Monitor drift in data quality.
- **Monitor model quality -** Monitor drift in model quality metrics, such as accuracy.
- **Monitor Bias Drift for Models in Production -** Monitor bias in your model's predictions.
- **Monitor Feature Attribution Drift for Models in Production -** Monitor drift in feature attribution.

Amazon SageMaker Model Monitor continuously monitors the performance of your machine-learning models in production. It automatically alerts you if any significant changes or quality issues are detected. Model Monitor uses rules to identify deviations from expected behavior and notify you promptly.

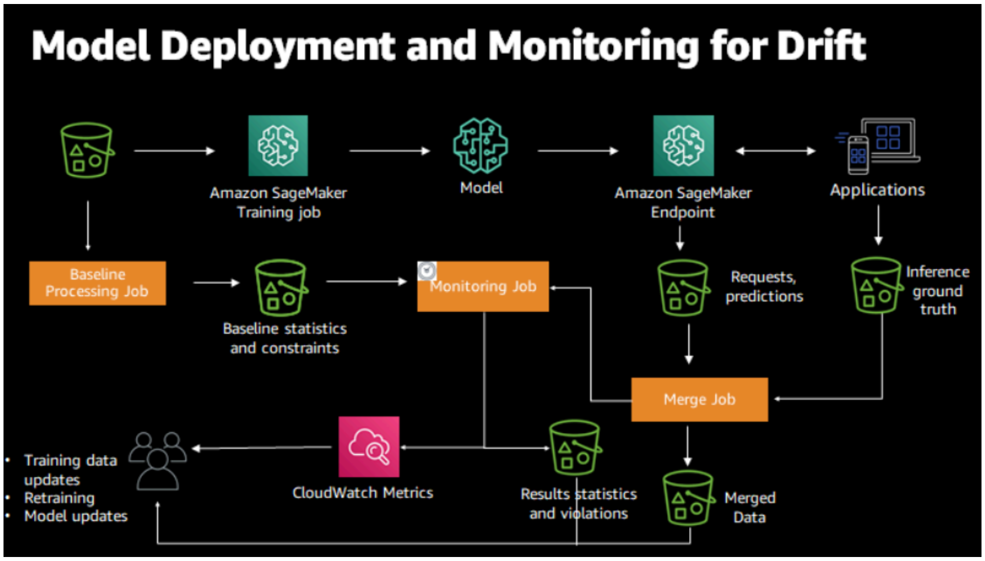


Figure: Amazon SageMaker Model Monitor

Amazon SageMaker Model Cards:

- Model governance is a framework that provides systematic visibility into machine learning (ML) model development, validation, and usage. Amazon SageMaker provides purpose-built ML governance tools for managing control access, activity tracking, and reporting across the ML lifecycle.
- Amazon SageMaker Model Cards are essentially a standardized template to document, retrieve, and share essential model information from development to deployment.

Features:

- Provide guidance on the appropriate use of the model.
- Support audit processes by offering detailed information on model training and performance metrics.
- Communicate specific business value that the model is expected to deliver.

AWS Managed AI Services

Amazon Polly

What is Amazon Polly?

Text-to-Speech Conversion: Transforms written text into spoken words.

Cloud-Based Service: Operates in the cloud, eliminating the need for local infrastructure.

Voice Options: Provides various voice choices, including Neural Text-to-Speech (NTTS) for natural-sounding speech.

Features

- **No Setup Fees**: Pay only for the text converted into speech, with no initial setup costs.
- **Multilingual Support**: Access various languages and Neural Text-to-Speech (NTTS) voices for creating speech-enabled applications.
- **Speech Caching and Replay**: Utilize caching and replay features for Amazon Polly's generated speech, available in formats like MP3.

Amazon Comprehend

What is Amazon Comprehend?

Natural Language Processing (NLP): Uses NLP to extract insights from document content.

Insight Extraction: Identifies entities, key phrases, language, sentiments, and other elements within documents.

Product Development: Leverage document structure understanding to develop new products.

Features:

- **Extract Insights from Diverse Text Sources**: Analyze text from documents, support tickets, product reviews, emails, and social media to uncover valuable insights.
- **Optimize Document Processing: I**mprove workflows by extracting key information, including text, phrases, topics, and sentiment from documents such as insurance claims.
- **Custom Document Classification**: Differentiate your business by training models to classify documents and recognize specific terms, without needing advanced machine learning skills.
- **Protect Sensitive Information**: Safeguard and manage access to sensitive data by identifying and redacting Personally Identifiable Information (PII) in documents.

Use cases:

- Analyze business and call center data
- Index and search through product reviews
- Manage legal briefs
- Handle financial document processing

Amazon Rekognition

What is Amazon Rekognition?

Cloud-Based Service: Utilizes advanced computer vision technology for image and video analysis.

No Machine Learning Expertise Needed: Accessible through an intuitive API.

Integration with Amazon S3: Quickly analyzes images and videos stored in Amazon S3.

Key Features:

Includes object and text detection, unsafe content identification, and facial analysis.

Facial analysis

- **User Verification:** Identifies and verifies individual identities.
- **Cataloging:** Organizes and manages face data for various applications.
- **Public Safety:** Enhances safety through surveillance and monitoring.
- Detect, analyze, and compare faces in both live streaming and recorded videos.

Image Analysis:

- **Object, Scene, and Concept Detection:** Detect and classify various objects, scenes, concepts, and celebrities present in images.
- **Text Detection:** Identify both printed and handwritten text in images, supporting multiple languages.

Video Analysis:

- **Object, Scene, and Concept Detection:** Categorize objects, scenes, concepts, and celebrities appearing in videos.
- **Text Detection:** Recognize printed and handwritten text in videos in different languages.
- **People Tracking:** Monitor individuals identified in videos as they move across frames.

Use cases:

- Simplify content retrieval with Amazon Rekognition's automatic analysis, enabling easy searchability for images and videos.
- Enhance security with Rekognition's face liveness detection, preventing identity spoofing beyond traditional passwords.
- Quickly locate individuals across your visual content using Rekognition's efficient face search feature.
- Ensure content safety with Rekognition's ability to detect explicit, inappropriate, and violent content, facilitating proactive filtering.
- Benefit from HIPAA Eligibility, making Amazon Rekognition suitable for handling protected health information in healthcare applications.

Amazon Lex

What Is Amazon Lex?

Build Chatbots: Create conversational interfaces using natural language processing.

Leverages Alexa Technology: Utilizes the same technology that powers Alexa for advanced language understanding.

Seamless Integration: Easily integrates with other AWS services to enhance chatbot functionality and user experience.

Features:

- Effortlessly integrate AI that comprehends intent, retains context, and automates basic tasks across multiple languages.
- Design and deploy omnichannel conversational AI with a single click, without the need to manage hardware or infrastructure.
- Seamlessly connect with other AWS services to access data, execute business logic, monitor performance, and more.

- Pay only for speech and text requests without any upfront costs or minimum fees.

Use Cases:

- **Enable virtual agents and voice assistants:** Provide users with self-service options through virtual contact center agents and interactive voice response (IVR), allowing them to perform tasks autonomously, like scheduling appointments or changing passwords.
- **Automate responses to FAQs:** Develop conversational solutions that answer common inquiries, enhancing Connect & Lex conversation flows with natural language search for frequently asked questions powered by Amazon Kendra.
- **Improve productivity with application bots:** Streamline user tasks within applications using efficient chatbots, seamlessly integrating with enterprise software through AWS Lambda and maintaining precise access control via IAM.
- **Extract insights from transcripts:** Design chatbots using contact center transcripts to maximize captured information, reducing design time and expediting bot deployment from weeks to hours.

Amazon Transcribe

What is Amazon Transcribe?

Speech-to-Text Conversion: Transforms audio speech into written text.

Deep Learning Technology: Utilizes automatic speech recognition (ASR) for accurate transcription.

Versatile Applications: Ideal for generating transcripts from various audio sources, such as meetings, interviews, and videos.

Features

- **Optimized for Specific Use Cases**: Ideal for customer service calls, live broadcasts, and media subtitling.
- **Medical Transcription:** Converts medical speech to text for clinical documentation with high accuracy.
- **Cost Structure:** Charges are based on the seconds of speech converted per month.

Use Cases

- **Customer Service:** Enhance customer interactions by transcribing service calls for analysis and improvement.
- **Live Broadcasts:** Generate real-time subtitles for live events and broadcasts.
- **Medical Documentation:** Streamline clinical documentation by transcribing medical speech accurately.

—Back to Index—

Amazon Translate

What is Amazon Translate?

Neural Machine Translation: Uses neural networks for accurate and natural text translations.
Language Pairs: Translates text between English and multiple other languages.
Source-Target Conversion: Converts text from a source language to a target language based on selected language pairs.

Benefits of Amazon Translate:

- **High-quality translations -** Provide precise and evolving translations across various applications.
- **Batch and real-time translations -** Integrate batch and real-time translation into your applications seamlessly using a single API call.
- **Customization -** Customize your ML–translated output to define brand names, model names, and other unique terms.

Use cases:

- **Translate user-generated content:** Automatically translate user-generated content, including social media posts, profiles, and comments, instantly in real-time.
- **Analyze online conversations in different languages:** Use a natural language processing application to analyze text in multiple languages and gain insights into public opinion about your brand, product, or service.
- **Create cross-lingual communications between users:** Implement real-time language translation capabilities in chat, email, helpdesk, and ticketing systems to enable English-speaking agents to communicate effectively with customers worldwide.

Amazon Mechanical Turk (MTurk)

What is Amazon Mechanical Turk?

- **Crowdsourcing Marketplace:** MTurk connects individuals and businesses with a global, virtual workforce for various tasks.
- **Task Types:** Includes simple data validation, research, survey participation, content moderation, and more.
- **How It Works:** Requesters post tasks (HITs) that Workers complete online. The system ensures workers are paid only for satisfactory work, and you can use qualification tests to select skilled Workers.

Advantages

- **Enhanced Efficiency:** Automate repetitive, manual tasks to streamline workflows. MTurk helps complete tasks quickly, freeing up internal resources for more strategic work.
- **Flexible Scaling:** Easily scale workforce up or down without the complexities of managing a temporary in-house team. MTurk provides access to a 24x7 global workforce.
- **Cost Reduction:** Lower labor and overhead costs with a pay-per-task model. MTurk helps manage expenses effectively while achieving results that might be challenging with a dedicated team.

Why Use MTurk?

- **Human Expertise:** Completes tasks better than computers, such as content moderation and data deduplication.
- **Efficient Crowdsourcing:** Breaks down complex projects into manageable microtasks for distributed workers, improving scalability and reducing manual effort.

MTurk Use Cases in Machine Learning

Data Collection and Annotation

- **Efficient Data Gathering:** MTurk simplifies the collection and labeling of large datasets needed for training ML models. It accelerates the process of annotating data, such as tagging images or categorizing text.

Model Improvement

- **Continuous Iteration:** Use MTurk for ongoing adjustments and enhancements to ML models. Human input helps in refining models by providing feedback and making necessary corrections.

Human-in-the-Loop (HITL)

- **Incorporating Human Feedback:** MTurk supports HITL workflows where human feedback is essential for model validation and retraining. For instance, annotating images with bounding boxes helps create precise datasets for computer vision tasks, especially when automated solutions fall short.

—Back to Index—

Amazon Augmented AI [Amazon A2I]

- Amazon Augmented AI (Amazon A2I) is a service that brings human review of ML predictions to all developers by removing the heavy lifting associated with building human review systems or managing large numbers of human reviewers.
- For example, extracting information from scanned mortgage application forms can require human review due to low-quality scans or poor handwriting.
- Building human review systems can be time-consuming and expensive because it involves implementing complex workflows, writing custom software to manage review tasks and results, and managing large groups of reviewers.

Amazon A2I offers built-in human review processes for typical ML tasks like content moderation and document text extraction.

- Content moderation
- Form extraction
- Image classification

Use cases:

- ➔ **Use Amazon A2I with Amazon Textract:-** Use Amazon Textract to randomly select and send documents from your dataset to humans for review.
- ➔ **Use Amazon A2I with Amazon Rekognition:-** Use Amazon Textract to randomly select and send images from your dataset to humans for review.
- ➔ **Use Amazon A2I to review real-time ML inferences:-** Use Amazon A2I to review real-time, low-confidence inferences made by a model deployed to a SageMaker. Continuously refine your model by incorporating feedback from A2I's output data.
- ➔ **Use Amazon A2I with Amazon Comprehend:-** Have humans to review inferences about text data using Amazon Comprehend.
- ➔ **Use Amazon A2I with Amazon Transcribe:-** Have humans to review transcriptions of video or audio files using Amazon Transcribe.

AWS DeepRacer

AWS DeepRacer is a reinforcement learning (RL)-enabled autonomous 1/18th-scale vehicle, designed for educational purposes with supporting services in the AWS Machine Learning ecosystem.

- AWS DeepRacer offers a platform for developing and testing deep reinforcement learning models. You can train these models in a simulated environment and then deploy them to a physical AWS DeepRacer car for autonomous racing.

AWS DeepRacer vehicles:

1. The original AWS DeepRacer device is a physical 1/18th-scale model car with a mounted camera and an onboard compute module. The compute module runs inference in order to drive itself along a track. The compute module and the vehicle chassis are powered by dedicated batteries known as the compute battery and the drive battery, respectively.
2. The virtual race car becomes the original AWS DeepRacer device, the Evo device, or various digital rewards that can be earned by participating in AWS DeepRacer League Virtual Circuit races.
3. The AWS DeepRacer Evo device is the original device with an optional sensor kit. The kit includes an additional camera and LIDAR (light detection and ranging), that allow the car to detect objects behind and lateral to itself.

Security, Compliance, and Governance for AI Solutions

Amazon Macie

What is Amazon Macie?

- Amazon Macie is a data security solution that employs machine learning algorithms and pattern recognition techniques to identify and safeguard sensitive data.
- By leveraging machine learning and pattern-matching capabilities, Amazon Macie not only detects sensitive data but also offers insights into potential data security threats.
- Additionally, it facilitates automated measures to mitigate these risks, enhancing overall data protection.

Features:

- Implement automated processes for detecting sensitive data on a large scale.
- Use Amazon Macie with Amazon Textract, Amazon Rekognition, and Amazon SageMaker to discover and secure sensitive data stored in Amazon S3.

AWS PrivateLink

What is AWS PrivateLink?

- AWS PrivateLink is a network service used to connect to AWS services hosted by other AWS accounts (referred to as endpoint services) or AWS Marketplace.
- Whenever an interface VPC endpoint (interface endpoint) is created for service in the VPC, an Elastic Network Interface (ENI) in the required subnet with a private IP address is also created that serves as an entry point for traffic destined to the service.

Interface endpoints

- It serves as an entry point for traffic destined to an AWS service or a VPC endpoint service. Gateway endpoints
- It is a gateway in the route-table that routes traffic only to Amazon S3 and DynamoDB.

Features:

- It provides security by not allowing the public internet and reducing the exposure to threats, such as brute force and DDoS attacks.
- With the help of AWS PrivateLink, VPC interface endpoint connects your VPC directly to the SageMaker API or SageMaker Runtime without using an internet gateway, NAT device, VPN connection.
- Define an interface VPC endpoint for Amazon Rekognition to connect your VPC to Amazon Rekognition.