Section 1:

Introduction

INTRODUCTION

Section 1:

About the exam & course setup

About the **AWS Certified AI Practitioner** Certification

Why getting certified?

- ✓ Impactful way to advance career
- ✓ Positioning as an expert
- ✓ Future proof + great job opportunities.

What is covered?

- ✓ AWS Certified AI Practitioner
- ✓ https://aws.amazon.com/certification/certified-ai-practitioner/

Demos

- ✓ Not needed for the exam.
- ✓ Help with memorizing.
- ✓ Give you practical foundation.

Goal

- ✓ Clear exam with ease.
- ✓ Knowledge for working with AWS

Passing Score

- ✓ 700 / 1000
- ✓ Goal: Achieve a score of 850+

Master the Exam

Free Trial Account

Not needed for the exam.
Help with memorizing
Give you a practical knowledge.

Exam Overview

https://aws.amazon.com/certification/certified-ai-practitioner/

Exam Duration

- ❑ Time: 120min

Exam Questions

- ❑ 85 questions
- ❑ Multiple Select, Multiple Choice

A company wants to use Amazon Rekognition to analyze images stored in Amazon S3. Which of the following is a benefit of using Amazon Rekognition?

- ❑ Automatic image enhancement and filtering AWS Glue
- ☑ Identification of objects, people, text, scenes, and activities in images
- ❑ Image compression and optimization for web delivery
- ❑ Real-time image rendering and editing

Recipe to clear the exam

Lectures
Step-by-step incl. Demos
~ 30-60 min / day

Quizzes
Practice and test your knowledge

Slides
Repeat and go through important points

Practice Test
Evaluate knowledge & weaknesses
Eliminate weaknesses

Book Exam

Confident & Prepared

aws certified
AI Practitioner
FOUNDATIONAL

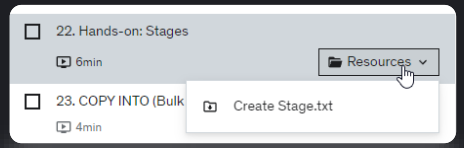
Final Tips

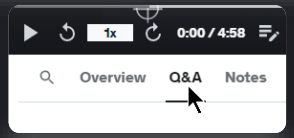
Resources

Q&A Section

Reviews

Connect & Congratulate



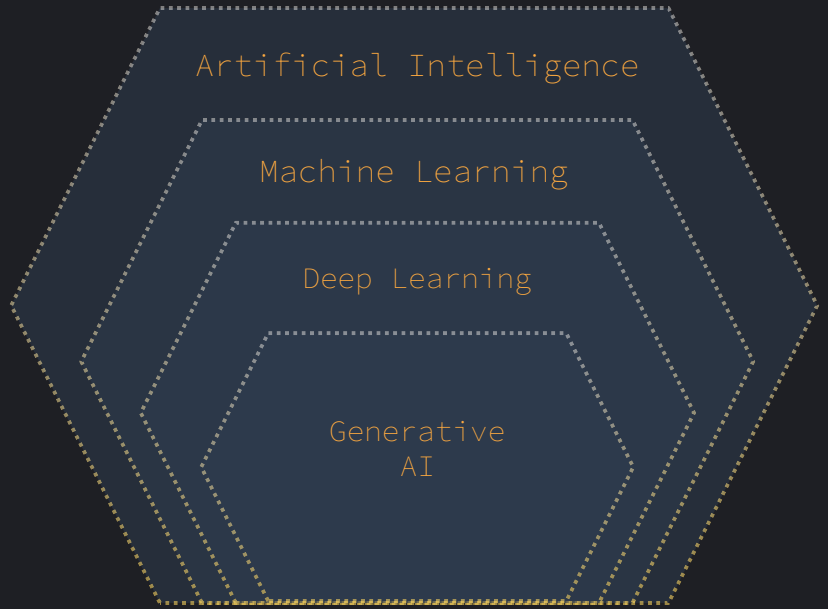


Generative AI & Machine Learning Basics

003-1040559 1250 003-77156.8 1760 0009-14563.7 73273

AI – Basics

- Artificial Intelligence:
  - Everything related to making machines smart.
- Machine Learning:
  - Teaching machines to learn from data.
- Deep Learning:
  - Designed to mimic the way brain work.
- Generative AI:
  - Not only learns from data but also creates new data.



Machine Learning - Basics

- Learn from data. Predict based on data.
- Train with large datasets, identify the patterns.

Key Concepts

Data	⇒ Quality and quantity of data impact model performance.
Algorithms	⇒ Formulas in scripts for solving problems.
Models	⇒ The result of training an algorithm with data
Training and Testing	⇒ The process of teaching a model and evaluating its performance on a different dataset

Introduction to Generative AI

- Generate new unique contents.

Text Generation

- Produce human-like writing on various topics.
- Creative Topics
- Technical reports

Amazon Bedrock

Image Generation

- Creative images from simple text prompts.

Amazon Bedrock

Audio and Speech Synthesis

- Generate realistic human-like voices.

Amazon Polly

Code Generation

- Auto-completing code
- Generate new code snippets

Amazon CodeWhisperer

Generative AI - Models

Foundation Models:

- Pre-trained models on large-scale internet data.
- Performs text-generation, chatbot interactions, information extractions.

Amazon Q Business

Amazon Q Business

Amazon Q Business

- AI-powered assistant.
- Intelligent answers, content generation, summaries and task automations.
- Accessible via web interface or APIs.
- Can integrate with other business platforms like Teams, Slack.

Amazon Q Business

Key Features:

Enterprise Integration	⇒ +40 pre-built connectors, Salesforce, Jira, ServiceNow, Zendesk…
User-Friendly Interface	⇒ Web based interface, integration with Microsoft Teams, Slack.
Rapid Deployment	⇒ Quick setup, without any code.
Managed Infrastructure	⇒ need for managing infrastructure.
Access Control	⇒ Respects user permissions within integrated enterprise applications.
Data Integration	⇒ Amazon S3, Salesforce, Oracle, Google Drive, Microsoft 365, and more.
Administrative Controls	⇒ specific guardrails and controls.

Amazon Q Business

Use Cases:

Content Creation

Marketing & Sales: Generate blog posts, social media headlines.

Research: Summarize academic papers, create new sections.

Enterprise Use-Cases

Knowledge Management: Find specific docs like company policies.

Support: Get customer support for common issues.

Executive Summaries: Summarize long meetings and project reports.

Key Insight Generation

Comparative Analysis: Compare documents, identify trends.

Market Research: Analyze market research, get insights

Amazon Q Business

Technical Details:

- IAM Identity Center:
  - Purpose: Manages user access
  - Function: Connects existing identity provider. Users can interact with Amazon Q Business.
- Retrieval Augmented Generation (RAG):
  - Purpose: Enhances gen AI models with up-to-date information.
  - Function: Retrieves relevant data from external sources. Ensures response accuracy.
- Enterprise Data Access Control:
  - Purpose: Ensures data security and compliance.
  - Function: Respects user permissions and integrates with SAML 2.0 supported identity providers like Microsoft Entra ID.
- Data Integration and Updates:
  - Purpose: Connects Amazon Q Business to various enterprise data sources..
  - Function: Uses pre-built connectors for easy integration.
- Plugins:
  - Purpose: Extends Amazon Q business' functionality
  - Function: Allows interaction with popular 3rd party applications like Jira, ServiceNow…

Language AI Services

Amazon Transcribe

Automatic Speech Recognition

Amazon Transcribe

What is Amazon Transcribe?

- **Definition**: Automatic Speech Recognition (ASR) service
- **Key Functions:**
  - Converts audio to text
  - Supports multiple languages and accents

Subtitles for videos

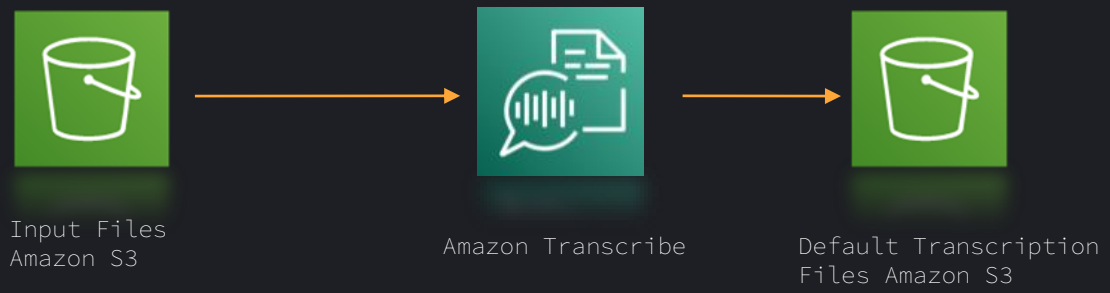
Call recording transcripts

Dictation transcripts

Amazon Transcribe

What is Amazon Transcribe?

- **Definition**: Automatic Speech Recognition (ASR) service
- **Key Functions:**
  - Converts audio to text
  - Supports multiple languages and accents



Amazon Transcribe

Key features:

Real-time Transcription
Transcribe audio streams in real-time with low latency

Batch Transcription
Transcribe pre-recorded audio files in batches

Custom Vocabulary
Add industry-specific terms for improved accuracy

Speaker identification
Identify and label different speakers in a conversation

Auto Punctuation
Automatically add punctuation and formatting to transcripts

Multi-language Support
Transcribe audio in multiple languages with high accuracy

Amazon Transcribe

Use Cases:

Subtitling and Closed Captioning

Add subtitles and closed captions to videos for accessibility and global audiences

Call Center Analytics

Analyze call transcripts to improve customer service and agent performance

Meeting Transcription

Automatically transcribe meetings to capture action items and decisions

Content Creation

Transcribe audio from podcasts and videos to generate articles, show notes, and more

Compliance and Regulation

Ensure compliance with regulations by transcribing sensitive audio recordings

Amazon Transcribe

Advantages:

Scalable

Handling varying volumes of audio content
⇒ Easy to scale with your business needs

Accurate

Advanced deep learning algorithms
⇒ high accuracy in transcriptions
⇒ continuously improving with usage

Cost-efficient

Pay-as-you-go model
⇒ significantly reducing costs compared to traditional transcription (manual labor)

Easy Integration

Integrates with existing services
⇒ Quick deployment + minimal disruption

Amazon Transcribe

Advantages

Feature	Amazon Transcribe	Traditional Transcription
Speed	Fast, near-instant processing	Slow, dependent on human effort
Accuracy	High accuracy with continuous learning	Variable accuracy, prone to human error
Cost	Cost-effective, pay-per-use model	Generally higher, fixed costs
Scalability	Highly scalable for large volumes	Difficult to scale
Customization	Supports custom vocabulary	Limited customization options
Integration	Seamless integration with AWS services	Requires manual setup

Amazon Transcribe

Advantages

Upload or Stream Audio

- Users can upload audio files or stream live audio directly to Amazon Transcribe.

Choose Transcription Settings

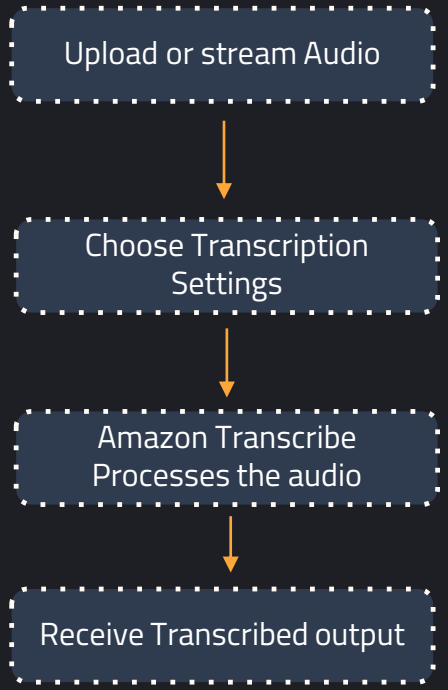
- Configure settings such as language, custom vocabulary, and speaker identification options.

Amazon Transcribe Processes the Audio

- The service uses advanced algorithms to analyze the audio and generate a text transcript.

Receive Transcribed Text

- Users receive the transcribed text in a structured format, ready for use.



Amazon Transcribe

Supported File Formats

Audio Formats:

- MP3
- MP4
- WAV
- FLAC
- OGG
- PCM encoding

AWS Service Integrations

- **Amazon S3**: For storing audio files and transcripts.
- **AWS Lambda**: For automating workflows and processing audio files.
- **Amazon Comprehend**: For analyzing transcribed text to extract insights.
- **Amazon Translate**: For translating transcripts into different languages.

Amazon Transcribe

Compliance and Security

HIPAA Eligible

- Eligible for use in systems covered by the Health Insurance Portability and Accountability Act (HIPAA)
  ⇒ Ensuring compliance for sensitive healthcare data.

Encryption in Transit and at Rest

- All data transmitted to and from Amazon Transcribe is encrypted using HTTPS.
- Audio files and transcripts are encrypted at rest using AES-256 encryption.

Amazon Polly

Text-to-Speech

Amazon Polly

What is Amazon Polly?

- Text-to-Speech(TTS) service.

- **Common Use Cases:**
  - **Media And Entertainment:** Voice-overs for videos and eBooks
  - **Business**: Interactive voice Response(IVR) and automated customer service
  - **Education**: Audio for learning materials

Amazon Polly

Key features:

Natural-Sounding Voices Lifelike speech synthesis that enhance user experience.
Wide Language Support Multiple languages and adjustable voices for global reach.
Custom Lexicons Define pronunciations for specific terms and names.
Speech Marks Detailed information for synchronization with visual elements.
Real-Time Generate speech on-the-fly for immediate applications.

Amazon Polly

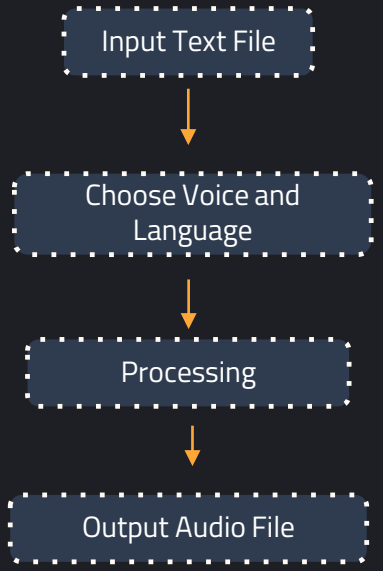
Process of Converting Text to Speech:

Input Text – Start with your written Content

Choose Voice and Language - Select from available options

Process Text - Amazon Polly synthesizes the speech.

Receive Audio Output - Get the final audio file



Amazon Polly

Supported Formats and Integrations

Audio Formats:

– **MP3**
Compressed, widely supported for mobile/web

– **OGG** (Vorbis)
High quality, smaller file sizes

– **PCM(WAV)**
Uncompressed, ideal for IoT devices

AWS Integrations:

– S3: Storage

– Lambda: Processing

– CloudFront: Content delivery

Amazon Polly

Compliance and Security

Data Security:

Supports encryption for data both in transit and at rest.

Compliance:

Amazon Polly is HIPPAA-eligible.

Generative AI: Selection and Metrics

Generative AI Models

Generative AI - Models

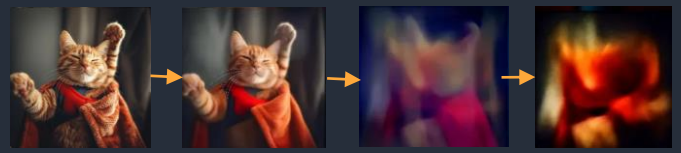
Diffusion Models:

- Produces high quality samples.
- Applicable to image, audio, text…
- Use Cases:
  - Generate high quality images,
  - Enhance resolution of low-quality image,
  - Filling missing parts of an image.
  - Generate audio

Generative AI - Models

Diffusion Models:

Forward Diffusion



- Gradually adds noise to the data.
- Creates a starting point for generation.
- Builds a robust training framework.
- Helps to understand data structure.

Backward Diffusion

- Generates new image by de-noising.
- Tries to predict noises in each step.
- Eventually reaches the brand-new image.

Generative AI - Models

Large Language Models (LLMs)

- Understand and generate human-like texts.
- Tokens:
  - Basic units of text. “The quick brown fox jumps over the lazy dog” > “The”, “quick”, “brown”, “fox”, …
- Embeddings and Vectors:
  - Numerical representations of tokens.
  - Vectors help model understand context.
  - Vector of word “king” can represent semantic similarity of words “queen” and “monarch”.

Generative AI - Models

Multimodal Models:

- Process and generate multiple types of data simultaneously.

Create captions for videos

Make Pictures from Text

Translate Languages with Visuals

Generative AI - Models

Generative Adversarial Networks (GANs)

- Class of machine learning frameworks.
- Uses two neural networks that compete against each other.
- Generator:
  - Starts with random noise and tries to produce data that mimics real data.
- Discriminator:
  - Gets both real data and fake data learns to tell the difference.
    Real Data: Its training data set.
    Fake Data: Generated by generator.

Generator creates fake data → Discriminator evaluates this data → Generator updates its parameters → Discriminator updates its parameters

Generative AI - Models

Variational Autoencders (VAEs)

- Generates new data with single neural network that contains, encoder and decoder.

Encoder

- Maps input data to latent space.
- It maps the input to a distribution, typically Gaussian

Latent Space

- Simplified representation of complex data.
- Encoder projects its output into latent space.

Decoder

- Reconstructs original data from latent variables.
- Obtains the new output that is different from the input

GANs

- Uses two competing networks, generator and discriminator
- Focuses on creating data that is indistinguishable from real data.

VS

VAEs

- Uses a single network with an encoder-decoder architecture.
- Generates data by sampling from a learned distribution.

Generative AI Capabilities & Challenges

Generative AI – Capabilities

Versatility

⇒ Can adapt for various tasks.

⇒ From *creative contents* to *data-driven decision-making*.

Real-Time Interaction

⇒ Produce content instantly.

⇒ Virtual assistants & chatbots

Task Simplification

⇒ Simplify complex workflows.

⇒ Reduces effort for repetitive tasks & report writings…

Innovation and Creativity

⇒ Can produce unique ideas & solutions & art.

Data Efficiency

⇒ Requires minimum data to produce valuable output

Customization

⇒ Create personalized content.

⇒ Recommendations & targeted marketing…

Scalability

⇒ Generate large amounts of content

⇒ Automated content generation & extensive data analysis

Generative AI – Real-World Use Cases

Media & Entertainment	⇒ Generate scripts for media contents. ⇒ Composes new songs or remixes of existing ones. ⇒ Create new characters.
Retail	⇒ Virtual fitting rooms. ⇒ Inventory optimization, demand predicting. ⇒ Personalized marketing campaigns.
Healthcare	⇒ Improved diagnostic accuracy. ⇒ Personalized treatment plans. ⇒ Simulation of new drug affects.
Finance	⇒ Risk assessment and investment strategies. ⇒ Model complex financial scenarios. ⇒ Generate personalized financial advices.
Manufacturing	⇒ Simulate and analyze manufacturing scenarios. ⇒ Generate multiple design options quickly. ⇒ Predict equipment maintenance.

Generative AI – Challenges

Compliance and Privacy

⇒ Outputs that reveals sensitive information.

⇒ Models should be trained cautiously.

Social Risks

⇒ Outputs that could harm organizations.

⇒ Models should be tested for those harmful content.

Toxicity

⇒ Offensive toxic content.

⇒ Training data should be filtered

Hallucinations

⇒ False or misleading information.

⇒ To prevent it, users can be warned, and some data can be labeled.

Security & Data

⇒ Sensitive data can lead privacy violations.

⇒ Data encryption, secure access controls

Complexity & Misintereption

⇒ Complex and easy to misunderstood outputs.

⇒ Provide clear explanations to prevent it.

Consistency & Reliability

⇒ Produce different results for same input.

⇒ Standardize model outputs through it training.

Factors of Selecting a Generative AI Model

Selecting Generative AI Model

- Evaluate multiple factors to choose the right model.

Task and Application	⇒ Define needed task or application requirements. ⇒ Select a model that is suitable for your task need.
Performance Requirements	⇒ Compare different model's performance based on your expectations.
Constraints	⇒ Assess the computational demands. ⇒ Ensure that you have sufficient proper data to fed your model. ⇒ Consider model deployment (on-premises or cloud).
Compliance	⇒ Evaluate model for potential biases. ⇒ Ensure the model adheres regulations and guidelines. ⇒ Consider ethical implications including privacy.
Cost	⇒ Evaluate cost of training deploying and usage of the model.

Selecting Generative AI Model

Model Types:

Amazon
Amazon Titan

- Text Summarization
- Embeddings
- Search
- Classification
- Image Generation

A121 Labs
Jurassic-2 Models

- Text Generation
- Summarization
- Paraphrasing
- Chat

Anthropic
Claude

- Text Generation
- Question Answering
- Summarization
- Code Generation

Stability AI
Stable Diffusion

- Realistic images from text.
- Quality improvement of existing images.

Meta
Llama

- Text Summarization
- Paraphrasing
- Classification
- Sentiment analysis

Business Metrics for Generative AI

003-1040559 1250 003-77156.8 1760 0009-14563.7 73273

Business Metrics – Generative AI

User Satisfaction

- Reflects overall user experience.
- Surveys and Feedback
- Net Promoter Score (NPS)
- Engagement Metrics

Average Revenue Per User (ARPU)

- Core financial metric.
- Revenue tracking.
- Pricing strategies.

Cross-Domain Performance

- Ability to perform different contexts.
- Versatility Testing.
- Domain Adaptation.
- Performance Metrics.

Conversion Rate

- Measures AI-driven action qualities like purchases, sign-ups…
- Conversion Tracking
- User Journey Analysis

Efficiency

- Delivering outputs cost-effectively
- Resource Utilization
- Time-to-Market

Generative AI Lifecycle

Generative AI Application Lifecycle

Idea and Planning

- ⇒ Identify use case
- ⇒ Feasibility Study

Select Foundational Model

- ⇒ Choose a model
- ⇒ Pre-trained models can be used.

Optimize Model

- ⇒ Fine-tuning

Evaluate Results

- ⇒ Validation
- ⇒ User testing
- ⇒ Ethical Review

Deploy

- ⇒ Integration
- ⇒ Monitoring
- ⇒ Maintenance

Amazon Bedrock

Amazon Bedrock

Amazon Bedrock

- Platform for building generative AI application.
- Offers access various foundational models.

Foundational Models	⇒ Diverse Model Selection: AI21 Labs, Stable Diffusion, Llama, Amazon Titan, Jurassic, Claude, Command…
Customization & Fine Tuning	⇒ Private Customization: Customize with own data. ⇒ Fine-Tuning: Tune models for specific domains.
Serverless	⇒ It eliminates the server management processes
Data Protection & Privacy	⇒ Prompt and model response secured. ⇒ Data is encrypted in transit and at rest.
Flexible Pricing	⇒ On-Demand Mode: Pay as you go ⇒ Provisioned Throughput Mode: for large and steady workloads.

Amazon Bedrock

- You can easily get benefits of other AWS Services.
  - Monitoring: AWS CloudWatch
  - Auditing: AWS CloudTrail
  - Storage: Amazon S3
  - Model Development: Amazon SageMaker

- Automation and Orchestration:
  - Agents for Amazon Bedrock: Handle complex tasks, use company data, enhance responses, and call APIs automatically.
  - Knowledge Bases for Amazon Bedrock: Provide company data, manage data intake and retrieval, support multi-turn conversations.

Amazon Bedrock

Benefits

Efficient Model Building

⇒ Rich variety of foundational models with a single API access.

⇒ Quick experimentation and model evaluation with playgrounds.

Secure Application Development

⇒ Data remains within AWS region encrypted.

⇒ AWS IAM provides fine-grained control over access.

Customizable Experiences

⇒ Automates complex tasks, integrates with existing data sources.

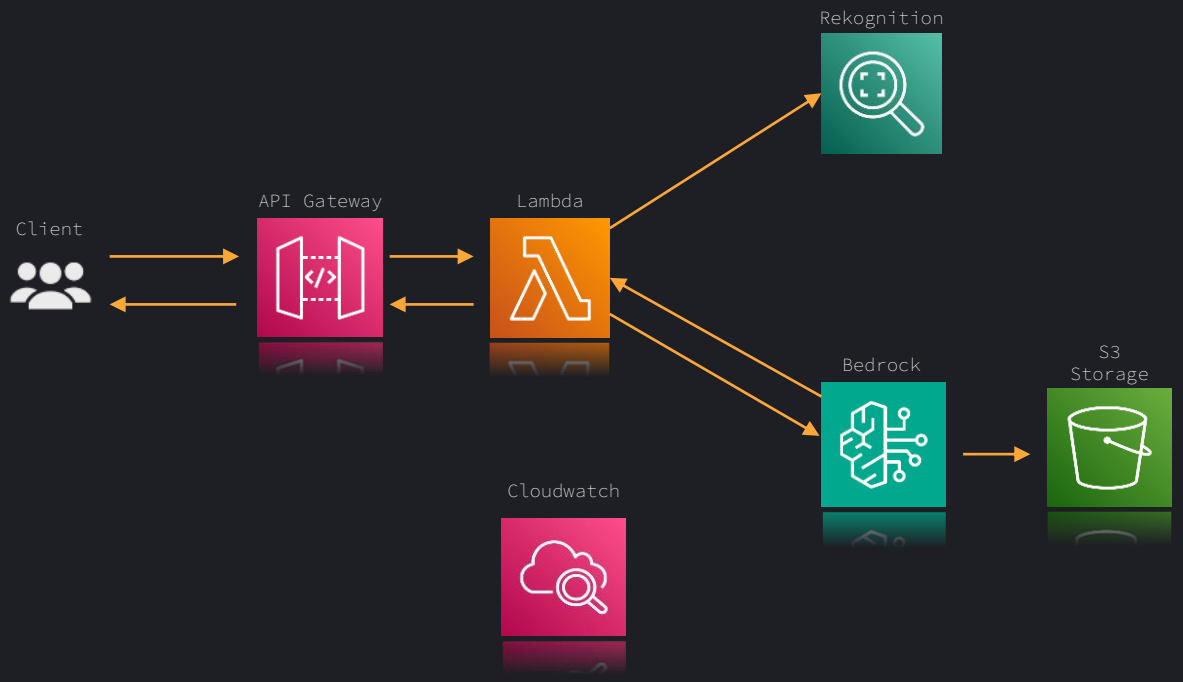
⇒ Users can fine-tune their models.

Amazon Bedrock

Use Cases

Content Creation	⇒ Generate dynamic content for various cases. ⇒ Personalize user experiences in real-time.
Customer Support	⇒ Chatbots and virtual assistants. ⇒ Automate repetitive support tasks.
Data Augmentation	⇒ Create data for training other ML models. ⇒ Enhance datasets.
Product Recommendations	⇒ Personalized product suggestions. ⇒ Enhance e-commerce platforms with dynamic content.

Amazon Bedrock – Sample Architecture



Why Improving Foundation Models

Why To Improve

- Foundation modals might not specialize in required domain.
- Some business needs might occur:
  - Model integration with companies' backend.
  - Collect data within model usage.
  - Restriction adjustments based on company policies.

Enhanced Accuracy

Increased Efficiency

Cost Reduction

Better User Experience

Scalability

Compliance and Security

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG)

- Improves LLM responses with data outside its training set.
- Integrates an external retrieval step.
- Why it is important:
  - Enhancing Trust: More accurate outputs as intended.
  - Cost-Effectiveness: Efficient than retraining FM with new data.

Latest Information

⇒ News sites

⇒ Company Databases

⇒ Social Media Feeds

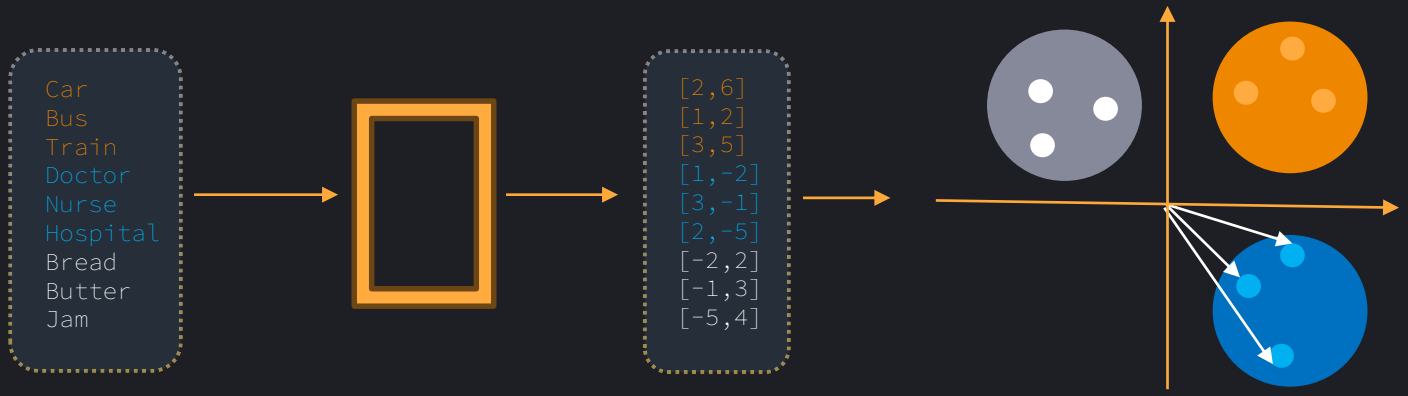
Developer Control

⇒ Control over information sources

RAG – How It Works

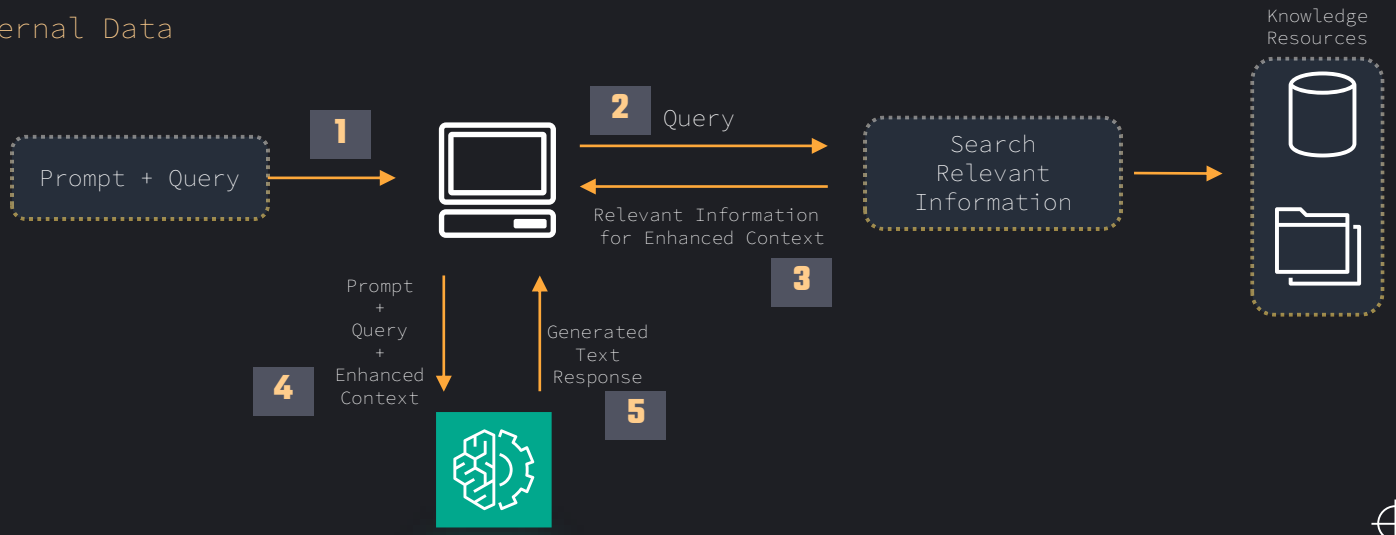
Vector Embeddings

- Each word represented as vectors (array of numbers).
- Model can understand meanings with numerical expressions.
- Vector Storing: AWS offers vector database solutions;
  - Amazon OpenSearch, pgvector extension in Amazon RDS for PostgreSQL and Amazon Kendra



RAG – How It Works

- Create External Data
- Retrieve Relevant Information
- Augment The LLM Prompt
- Update External Data



Agents

Agents

- Autonomous entities that interacts with an environment.
- Agents can make decisions and perform actions.

Intermediary Operation

⇒ Communication between AI Model & DBs,CRMs

⇒ Intelligent bridge between AI model and business backend.

Action Launch

⇒ Can perform tasks.

⇒ Adjust service settings, process transaction, retrieve documents.

Feedback Integration

⇒ Can help AI learning by gathering data based on actions.

⇒ Increase AI accuracy.

Agents

Improved Productivity	⇒ Autonomously perform specific tasks.
Reduced Costs	⇒ Reduce operational costs.
Informed Decision-Making	⇒ Process real-time data for better decisions.
Enhanced Customer Experience	⇒ Personalizing interactions

AGENT 1
Personalize Shopping Assistance

⇒ Provide product recommendations based on customer actions

AGENT 2
Customer Feedback Collection

⇒ Sends feedback surveys, analyzes responses in real-time

AGENT 3
Inventory Management

⇒ Updates inventory levels and manages stock information notify procurement team

Prompt Engineering

Essentials of Prompt Engineering

Essentials of Prompt Engineering

Prompt:
Is the input or query provided to a language model to generate a response.

Can be

- Question
- Statement
- Set of Instructions

Prompt Engineering: is the process of designing and preparing prompts.

Essentials of Prompt Engineering

Elements of a Prompt

Instruction:

Summarize the given article in no more than 50 words.

Context:

The article discusses the impact of artificial intelligence on various industries.

Input Data:

"AI is revolutionizing healthcare, finance, and transportation by enhancing diagnostics, improving fraud detection, and enabling autonomous vehicles. These advancements present both opportunities and challenges for society."

Output Indicator:

Provide a concise summary.

Example Output:

AI is transforming healthcare, finance, and transportation, improving diagnostics, fraud detection, and enabling autonomous vehicles, presenting new opportunities and challenges.

Essentials of Prompt Engineering

Negative Prompting

- Used to guide the model away from producing certain types of content or exhibiting specific behaviors.
- Used to prevent the model from producing

Hate speech | Explicit content | Biased language

- Helps steer the output towards more appropriate content.

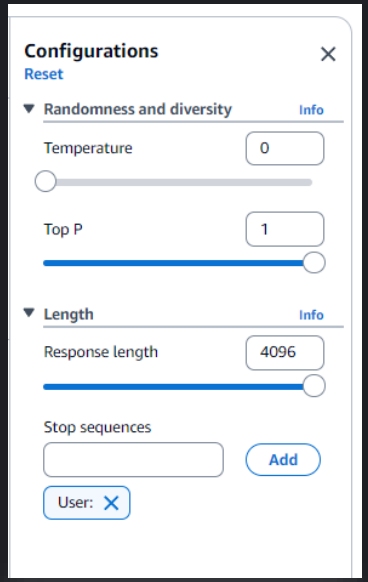
Optimizing Inference Parameters

Optimizing Inference Parameters

Inference Parameters

- Adjust, limit or influence the model response.
- Vary based on the model that you are using.

Example: randomness and diversity, and length.



Optimizing Inference Parameters

Randomness And Diversity

- Randomness: Is about how arbitrary the responses can be.
- Diversity: Is about variety of words utilized in the responses.
  - Control how varied and unpredictable the model's responses are.
  - The most common randomness and diversity parameters are

Temperature	Top P	Top K
Controls randomness	top words to reach a probability threshold	Fixed number of top words

Optimizing Inference Parameters

Temperature

- Controls the creativity of the model's output.
- It is set between 0 and 1.
- Higher temperature => more diverse and unpredictable output.
- Lower temperature => more focused and predictable output.

Optimizing Inference Parameters

Top P

- Limits the number of words the model can choose from.
- It is set between 0 and 1.

Higher Top P => broad range of possible words.

Lower Top P => less words.

Optimizing Inference Parameters

Top K

- Limits the number of words to the top k most probable words, regardless of their percent probabilities.
- Higher Top K setting => Broad range of possible words.
- Lower Top K setting => Less words from the total probability distribution.

Optimizing Inference Parameters

Length

- Used to limit the length of the response.
- The most common Length parameters are

Maximum Length

Stop sequences

Optimizing Inference Parameters

Maximum Length

- Determines the maximum number of tokens.

Stop Sequences

- Signal the model to stop generating further output.
- Useful in tasks where the desired output length is variable or difficult to predict in advance.

Prompting Best Practices

Prompting Best Practices

Your prompt should be clear and concise

- Prompts should be straightforward and avoid ambiguity

Your prompt should include context if needed

- Provide any additional context that would help the model respond accurately

Your prompt should use directives for the appropriate response type

- Summary, question, or poem

Prompting Best Practices

Your prompt should consider the output in the prompt

Example:

Calculate the area of a circle with a radius of 3 inches (7.5 cm).
Round your answer to the nearest integer.

Start prompts with an interrogation

- Who, What, Where, When, Why, and How.

Provide an example response

Prompting Best Practices

Break up complex tasks

- Divide the task into several subtasks.

Experiment and be creative

- Try different prompts to optimize the model's responses.

Use prompt templates

Prompt Engineering Techniques

Prompt Engineering Techniques

- Prompt Techniques are ways to give instructions to an AI model so it knows how to respond or do a task.
- Most comment prompt techniques include

Zero-shot

Few-shot

Chain-of-Thought

Prompt Engineering Techniques

Zero-Shot Prompting

- Prompt does not include any examples or demonstrations.
- Effective on a larger and more capable model.

Prompt Engineering Techniques

Few-Shot Prompting

- Uses contextual examples
- Providing more examples can help the model better understand the task
- Too many examples might introduce noise or confusion

Prompt Engineering Techniques

Chain-of-Thought Prompting

- Divides complex reasoning tasks into smaller, intermediary steps.
- Can be employed using either.

Zero-shot

Few-shot

- To initiate the chain-of-thought reasoning process in a machine learning model, you can use the phrase "Think step by step”.

Prompt Engineering Techniques

Using Zero-Shot

Prompt: *John has 3 apples. He buys 5 more apples and then eats 2. How many apples does he have now? Think step by step.*

Output:

```
John starts with 3 apples.

He buys 5 more apples: 3 + 5 = 8.

He eats 2 apples: 8 - 2 = 6.

John has 6 apples now.
```

Prompt Engineering Techniques

Using Few-Shot / One-Shot

Prompt:

Example #1:
Question: If there are 3 baskets and each basket contains 2 apples, how many apples are there in total?

Thought: There are 3 baskets and each basket contains 2 apples. To find the total number of apples, we multiply the number of baskets by the number of apples in each basket.
Calculation: 3 baskets * 2 apples/basket = 6 apples
Answer: 6 apples.

Now solve this in the same way:
Question: If there are 5 boxes and each box contains 4 books, how many books are there in total?

Answer:

Thought: There are 5 boxes and each box contains 4 books. To find the total number of books, we multiply the number of boxes by the number of books in each box.
Calculation: 5 boxes * 4 books/box = 20
Answer: 20 books

Prompt Misuses and Risks

Prompt Misuses and Risks

- Refers to the potential issues or challenges associated with the prompts used in AI or machine learning models.
- These risks can affect the accuracy, fairness, or reliability of the model's outputs.
- Common prompt risks

Poisoning, hijacking, and prompt injection

Exposure and prompt leaking

Jailbreaking

Prompt Misuses and Risks

Poisoning, Hijacking, And Prompt Injection

Poisoning	Intentional introduction of malicious or biased data
Hijacking and prompt injection	Influencing the outputs of generative models by embedding specific instructions within the prompts themselves.

Prompt Misuses and Risks

Exposure And Prompt Leaking

Exposure

Risk of exposing sensitive or confidential information to a generative model during training or inference

Prompt leaking

Refers to the unintentional disclosure or leakage of the prompts or inputs used within a model.

Prompt Misuses and Risks

Jailbreaking

- Is changing or breaking the rules of a computer model or AI assistant to make it do things it normally isn't allowed to do.
- Aims to bypass or exploit vulnerabilities in the AI system's filtering mechanisms or constraints.

Amazon Comprehend

Amazon Comprehend

- NLP service for extract insights and relationships from a text.
- Analyze text data without expertise in machine learning.
- Amazon Comprehend Medical.

Features

Entity Recognition

⇒ Identify specific entities. Names, locations, dates…
⇒ Amazon was founded by Jeff Bezos in Seattle.

Sentiment Analysis

⇒ Determines overall sentiment.
⇒ “I love this product, it’s amazing” → Positive sentiment.

Amazon Comprehend

Features

Key Phrase Extraction	⇒ Extract significant phrases. ⇒ The report highlights the significant growth in renewable energy.
Topic Modeling	⇒ Groups documents by common themes.
Topic Modeling	⇒ Categories text into custom categories defined by user.

Amazon Comprehend

Why Amazon Comprehend

Integrate Powerful NLP into Applications

⇒ API opportunities

Simplify Document Processing Workflows

⇒ Automation capabilities

Integration within AWS Services

⇒ Can combine with Amazon S3, Lambda

Enhanced Privacy, Security and Compliance

⇒ IAM

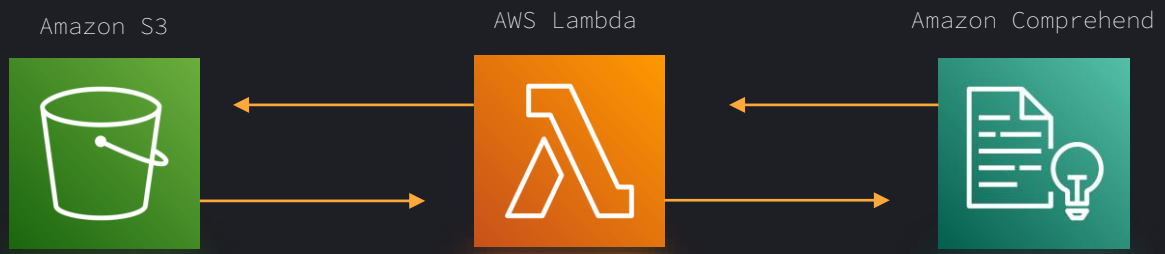
⇒ KMS

Cost Effective

⇒ Pay as you go

Amazon Comprehend

Sample Architecture



Amazon Rekognition

Amazon Rekognition

- Is a cloud-based image and video analysis service
- Analyzes any image or video file that’s stored in Amazon S3
- It can be used to :

Detect objects

Detect texts

Detect Unsafe content

Compare faces

Amazon Rekognition

Use cases

- Searchable Media Libraries
- Face-Based User Identity Verification
- Face Liveness Detection
- Facial Search
- Unsafe Content Detection
- Detection of Personal Protective Equipment
- Celebrity Recognition
- Text Detection

Amazon Personalize

Amazon Personalize

- Generate item recommendations for users
- Generates recommendations primarily based on item interaction data
- Interaction data can come from

Bulk interaction records

Real-time events

Amazon Personalize

Common Use Cases

- Personalizing a video streaming app
- Adding product recommendations to an ecommerce app
- Adding real-time next best action recommendations to your app
- Creating personalized emails
- Creating a targeted marketing campaign
- Personalizing search results

Amazon Fraud Detector

Amazon Fraud Detector

- Fully managed service that identifies fraudulent online activities.
  - Payment frauds, fake accounts, bots…
- It uses pre-build machine learning models and offers customization.
- Real-time detection.

Easy Integration

Automated Model Traning and Deployment

Flexible and Scalable

Amazon Fraud Detector

Workflow

Data Ingestion → Feature Engineering → Model Training → Model Deployment → Fraud Detection

Amazon Fraud Detector

- Reduced fraud losses.
- Improved customer experience.
- Operational efficiency.

E-commerce Transactions

Account Registration

Lending Applications

Online Gaming

Amazon Augmented AI

Amazon Augmented AI

- Enables a human review of machine learning (ML) systems.
- It makes building and managing human reviews for ML applications easy.
- Provides built-in human review workflows
- Supports custom human review workflows

Amazon Augmented AI

Use cases

- Amazon A2I with Amazon Textract
- Amazon A2I with Amazon Rekognition
- Amazon A2I to review real-time ML inferences
- Amazon A2I with Amazon Comprehend
- Amazon A2I with Amazon Transcribe
- Amazon A2I with Amazon Translate
- Amazon A2I to review tabular data

Overview: Security, Governance & Compliance

Overview: Security & Governance & Compliance

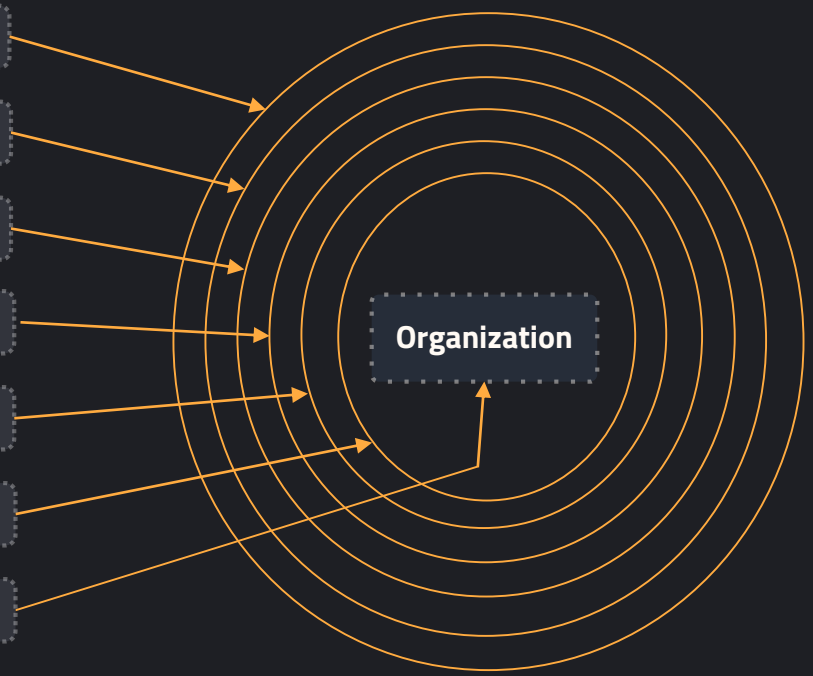
- Security & Governance & Compliance of AI solutions are essential for organization

Security	Ensure that Confidentiality, Integrity, and Availability are maintained
Governance	Ensure that an organization can add value and manage risk in the operation of business
Compliance	Is about following established standards and requirements

Overview: Security & Governance & Compliance

Defense in depth

7. Policies, procedures, and awareness
6. Threat detection and incident response
5. Infrastructure protection
4. Network and edge protection
3. Application protection
2. Identity and access management
1. Data protection



Overview: Security & Governance & Compliance

1. Data protection

Is all about making sure the data is protected

Data can be protected at

- At Rest
- In Transit

Overview: Security & Governance & Compliance

2. Identity and access management

Only authorized users, applications, or services are allowed.

AWS Identity and Access Management (IAM)

Overview: Security & Governance & Compliance

3. Application protection

Includes measures to protect against various application threats like

- Unauthorized access
- Data breaches
- Denial-of-service (DoS) attacks

Overview: Security & Governance & Compliance

4. Network and edge protection

Used to protect the network infrastructure and the boundaries of a cloud environment.

Overview: Security & Governance & Compliance

5. Infrastructure protection

Includes measures to protect against various Infrastructure threats like

- Unauthorized access
- Data breaches
- System failures
- Natural disasters

Overview: Security & Governance & Compliance

6.Threat detection and incident response

Identify and address potential security threats or incidents.

Threat detection:
Process of identifying and recognizing potential security threats

Incident response:
How an organization handles and reacts to a security breach or cyberattack

Overview: Security & Governance & Compliance

7. Policies, procedures, and awareness

Policies:
Set security expectations

Procedures:
Detailed steps for implementing these policies

Awareness:
Ensures employees understand and follow them

AI Compliance Standards and Regulated workloads

AI Compliance Standards and Regulated workloads

Standard : Is a set of rules that are agreed upon by experts or organizations

Compliance Standards : are standards that must be followed

Advantages of following standards

- Consistency and Quality
- Interoperability
- Safety
- Efficiency

AI Compliance Standards and Regulated workloads

- AI standards compliance differs from traditional software standards compliance

Complexity and opacity

Unique risks

Dynamism and adaptability

Algorithm accountability

Emergent capabilities

AI Compliance Standards and Regulated workloads

- Standards from National Institute of Standards and Technology (NIST)
- Standards from European Union Agency for Cybersecurity (ENISA)
- Standards from International Organization for Standardization (ISO)
- AWS System and Organization Controls (SOC)
- Health Insurance Portability and Accountability Act (HIPAA)
- General Data Protection Regulation (GDPR)
- Payment Card Industry Data Security Standard (PCI DSS)

AI Compliance Standards and Regulated workloads

Regulated workloads

- Refer to tasks, processes, or operations that are subject to specific rules, guidelines, or regulations set by authoritative bodies
- Industries with high degrees of regulatory compliance requirements

Financial services

Healthcare

Aerospace

AWS Services for Compliance and Governance

AWS Services for Compliance and Governance

- AWS Config
- AWS Artifact
- Amazon Inspector
- AWS CloudTrail
- AWS Audit Manager
- AWS Trusted Advisor

AWS Services for Compliance and Governance

AWS Config

Helps you in

- Resource administration
- Auditing and compliance
- Managing and troubleshooting configuration changes

AWS Services for Compliance and Governance

AWS Artifact

- Provides on-demand downloads of AWS security and compliance documents
- This includes AWS ISO certifications, PCI reports, and SOC Reports

AWS Services for Compliance and Governance

Amazon Inspector

- Continuously scans AWS workloads for Software Vulnerabilities and Unintended Network Exposure
- Including



- Provides a risk score

AWS Services for Compliance and Governance

AWS CloudTrail

- Used for auditing, governance, and compliance of your AWS account.
- Events => Actions taken by a user, role, or an AWS service.
- Events include actions taken in the AWS Management Console, AWS Command Line Interface (AWS CLI), and AWS SDKs and APIs.

AWS Services for Compliance and Governance

AWS Audit Manager

- Helps you continually audit your AWS usage
- Automates evidence collection

AWS Services for Compliance and Governance

AWS Trusted Advisor

- Continuously evaluates AWS environment using best practice checks across the categories of

Cost optimization	Performance
Resilience	Security
Operational excellence	Service limits

- Recommends actions to remediate any deviations from best practices

Data Governance Strategies for AI

Data Governance Strategies for AI

Data management concepts

Data lifecycles

Is the management of data throughout its entire lifespan

Collection > Processing > Storage > Consumption > Disposal or archiving

Data logging

Is the systematic recording of data related to the processing of an AI workload

Data Governance Strategies for AI

Data management concepts

Data residency

Is the physical location where data is stored and processed

Data monitoring

Is an ongoing observation and analysis of data used in AI workloads

It includes

- Monitoring data quality
- Identifying anomalies
- Tracking data drift

Data Governance Strategies for AI

Data management concepts

Data analysis

Is used to understand the characteristics, patterns, and relationships within the data used for AI workloads

Help to gain insights into the data

Data retention

Define how long data should be kept for AI workloads

Help organizations manage the lifecycle of data used in their AI systems

Data Governance Strategies for AI

Data governance strategies

Data quality and integrity	Establish data quality standards Implement data validation and cleansing
Data protection and privacy	Develop and enforce data privacy policies Establish data breach response and incident management procedures

Data Governance Strategies for AI

Data governance strategies

Data lifecycle management

Classify and catalog data assets

Implement data retention and disposition policies

Develop data backup and recovery strategies

Data Governance Strategies for AI

Data governance strategies

Responsible AI

Establish responsible frameworks and guidelines

Implement processes to monitor and audit AI and generative AI models

Provide training and support

Data Governance Strategies for AI

Data governance strategies

Governance structures and roles

Establish a data governance council or committee

Define clear roles and responsibilities

Provide training and support

Data Governance Strategies for AI

Data governance strategies

Data sharing and collaboration

Develop data sharing agreements and protocols

Implement data virtualization or federation techniques

Encourage data-driven decision-making

Security and Privacy Essentials for AI Systems

Security and Privacy Essentials for AI Systems

AI systems security : Measures and practices put in place to protect artificial intelligence systems from threats, vulnerabilities, and unauthorized access.

AI Systems should be secured because

- AI models process sensitive data
- AI Systems can be vulnerable to adversarial attacks
- Integration into critical applications and decision-making processes

Security and Privacy Essentials for AI Systems

Threat detection

Threat : Is any potential danger that could exploit a vulnerability to cause harm

Threat Detection : Is the process of identifying and analyzing potential threats.

AI-powered threat detection systems can be used to detect threats

Security and Privacy Essentials for AI Systems

Vulnerability management

Vulnerability : Weakness or flaw in a system that can be exploited by a threat actor

Vulnerability management : Is the process of identifying, assessing, and addressing vulnerabilities

Regularly conduct security assessments

Implement robust update processes

Security and Privacy Essentials for AI Systems

Infrastructure protection

Secure the underlying infrastructure that supports AI and generative AI systems

Prompt injection

Is an attempt to manipulate the input prompts to generate malicious or undesirable content

To reduce the risk

- Employ techniques, such as prompt filtering, sanitization, and validation
- Develop robust models and training procedures

Security and Privacy Essentials for AI Systems

Data encryption

Encryption : is the process of converting readable data into unreadable data using a cryptographic algorithm and an encryption key

Is used to protect the confidentiality and integrity of training data

Security and Privacy Essentials for AI Systems

OWASP Top 10 for LLMs

- Open Web Application Security Project
- Is the list of top 10 AI LLM vulnerabilities

1. Prompt injection
2. Insecure output handling
3. Training data poisoning
4. Model denial of service
5. Supply chain vulnerabilities
6. Sensitive information disclosure
7. Insecure plugin design
8. Excessive agency
9. Overreliance
10. Model theft

Security and Privacy Essentials for AI Systems

1. Prompt injection

Malicious user inputs that can manipulate the behavior of a language model

2. Insecure output handling

Failure to properly sanitize or validate model outputs

3. Training data poisoning

Introducing malicious data into a model's training set, causing it to learn harmful behaviors

4. Model denial of service

Techniques that exploit vulnerabilities in a model's architecture to disrupt its availability

Security and Privacy Essentials for AI Systems

5. Supply chain vulnerabilities	Weaknesses in the software, hardware, or services used to build or deploy a model
6. Sensitive information disclosure	Leakage of sensitive data through model outputs or other unintended channels
7. Insecure plugin design	Flaws in the design or implementation of optional model components that can be exploited
8. Excessive agency	Granting a model too much autonomy or capability, leading to unintended and potentially harmful actions

Security and Privacy Essentials for AI Systems

9. Overreliance

Over-dependence on a model's capabilities, leading to over-trust and failure to properly audit its outputs

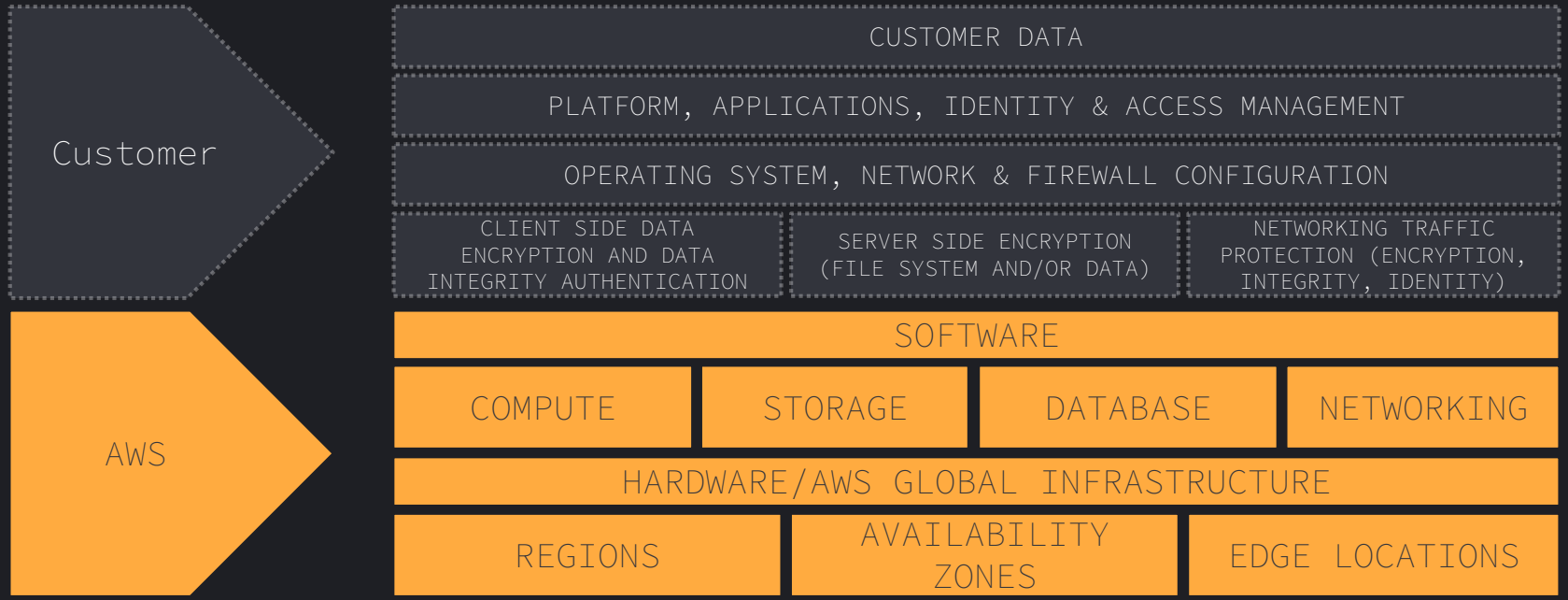
10. Model theft

Unauthorized access or copying of a model's parameters or architecture, allowing for its reuse or misuse

AWS Services for Securing AI Applications

AWS Services for Securing AI Applications

The AWS Shared Responsibility Model



AWS Services for Securing AI Applications

Foundational AWS security services

AWS Security Hub

AWS KMS

Amazon GuardDuty

AWS Shield

AWS Services for Securing AI Applications

Identify sensitive data before training models

Amazon Macie

- Used to scan S3 buckets for Personally Identifiable Information (PII), Personal Health Information (PHI), Financial Information, and other sensitive data
- Uses ML to automate sensitive data discovery at scale

AWS Services for Securing AI Applications

Manage identities and access to AWS services and resources

IAM

- Enables you to specify who or what can access services and resources in AWS
- Enables fine-grained permissions

IAM Access Analyzer : Ensure Least Privilege Access

AWS Services for Securing AI Applications

Protect data from exfiltration (data theft) and manipulation

Amazon VPC

AWS Network Firewall

AWS PrivateLink

AWS Services for Securing AI Applications

Protect AI workloads with intelligent threat detection

Amazon GuardDuty

Amazon Inspector

Amazon Detective

Defend your generative AI web applications and data

Defend your generative AI web applications and data

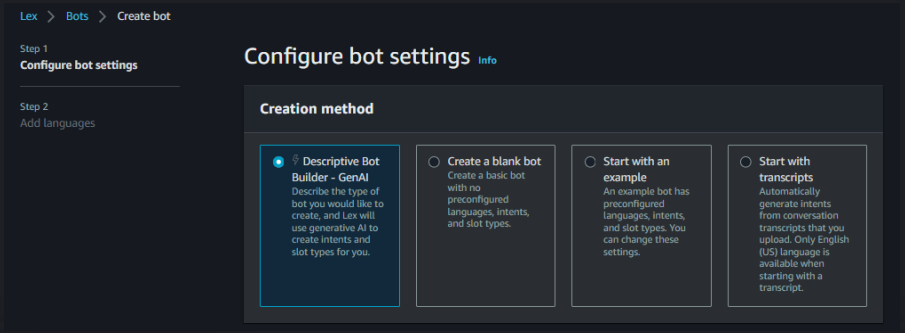
Amazon Lex

Amazon Lex

Service for building conversational Interfaces

Use Cases:

- Customer Service Chatbots
- Informational Bots
- Voice Assistants

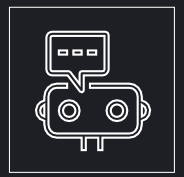


Key Features:

- Natural Language Understanding
- Speech Recognition
- Alias and Versioning support

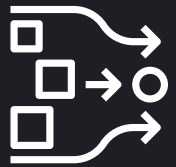
Amazon Lex

Building Conversational Interfaces



Create a bot

Test the bot



Publish a version
and
create an alias

Deploy the bot

Amazon Lex

Core Concepts

- Bot performs automated tasks like booking services
- Language language(s) configured in Amazon Lex
- Intent represents user-desired action
- Slot parameters required for intent
- Slot type defines the type of data slots hold
- Version a snapshot of a bot’s configuration
- Alias a pointer to a specific bot version