Introducing Generative AI

FOUNDATION MODELS · TOKENS & EMBEDDINGS · RISKS · BEDROCK / Q / SAGEMAKER · USE CASES · PARTYROCK STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

01 **Foundation models (FMs):** deep learning models based on **complex neural networks**, extensively pre-trained on **massive amounts of data**. What sets them apart from other deep learning models: they produce **generalized responses** — you can ask things they were never specifically designed to answer.

02 **A prompt** is an instruction or question written in **natural language** that asks a generative AI model to perform a specific task. In ML terms the prompt is the *new data* given to a trained model to make an **inference** — so every output is a **prediction**, not a lookup.

03 **Text pipeline: tokenization** splits input into tokens (words / subwords / characters) → **embedding** converts each token to a numeric **vector** (dense; hundreds–thousands of dimensions; similar meanings sit close together) → the LLM predicts **what probably comes next**.

04 **Three AWS services,** least → most complex: **Amazon Q** (use generative AI without building an app) → **Amazon Bedrock** (build genAI apps on popular FMs through a single API) → **Amazon SageMaker AI** (build, train, and deploy your own models across the full ML lifecycle).

05 **Optimal outputs** rest on four components: **model choice**, **prompt design**, **data quality**, **safeguards**. To influence results, escalate: richer prompts → external knowledge base → additional training — complexity *and cost* rise with each step.

02 KEY TERMS

TERM	DEFINITION
Foundation model (FM)	Very large pre-trained deep learning model able to generalize across tasks
Large language model (LLM)	FM for text, pre-trained on massive text datasets (trillions of words)
Multimodal model	Combines data types — text, images, sound — for a more comprehensive understanding
Prompt	Natural-language instruction or question requesting a specific task from a model
Token	Small unit of text (word, subword, or character) produced by tokenization
Embedding	Numeric representation a token is converted into; “embedding” and “vector” are used interchangeably
Vector	Dense numeric representation with hundreds–thousands of dimensions; similar tokens cluster together
Inference parameters	Settings that make output more deterministic (accurate) or less deterministic (creative)
Hallucination	Output that sounds plausible but is verifiably untrue (e.g., invented citations)
Guardrails	Tools and policies that keep undesired content out of inputs and outputs
SageMaker JumpStart	ML hub in SageMaker AI: deploy publicly available FMs with a few clicks
PartyRock	Amazon Bedrock playground — build and share genAI apps, no code, no AWS account

03 FM CATEGORIES AND EXAMPLE PROMPTS

not hard-and-fast buckets; all accept natural-language input

CATEGORY	GENERATES	EXAMPLE USES	EXAMPLE PROMPT
Language	Text-based responses	Virtual assistance, chat-based assistants	“Summarize distinctions between dogs and cats.”

CATEGORY	GENERATES	EXAMPLE USES	EXAMPLE PROMPT
Image	High-resolution images	Product design, marketing	“Generate an image of a black puppy with large ears playing with a ball.”
Audio	Music, sounds, speech	Game sound effects, synthesized speech	“Generate the sound of a joyful puppy barking.”
Biomedical	Protein / molecular structures	Therapeutics design, new drug discovery	“Predict the 3D structure of canine olfactory receptor protein OR5B17.”

04 FOUR GENERATIVE MODEL ARCHITECTURES

exam asks for distinguishing features, not math

ARCHITECTURE	HOW IT GENERATES	KNOWN FOR
Diffusion	Adds noise to a clean input, then learns to reverse the process and remove it	High-quality images
GAN	Two networks compete: **generator** creates samples to fool the **discriminator**, which classifies real vs. fake	Realistic data samples
VAE	Learns a compressed **latent representation** of input data; generates new samples resembling the original	Efficient latent space
Transformer	Processes sequences **in parallel** and attends to different parts of the input simultaneously	Modern LLMs

05 HOW AN LLM ANSWERS A PROMPT

01 **Prompt** in natural language → 02 **Tokenize** into words / subwords / characters → 03 **Embed** tokens as dense vectors → 04 Predict **what probably comes next** → 05 Output = a **prediction**

06 THREE AWS GENERATIVE AI SERVICES

all managed · data encrypted by default

SERVICE	BUILT FOR	REMEMBER
Amazon Q	Using generative AI **without building an app** — Q Developer (coding assistant), Q Business (chat over enterprise data)	Built on Bedrock
Amazon Bedrock	Building genAI applications: choose / experiment / evaluate FMs, playgrounds, guardrails — no ML infrastructure	Many FMs, **one API**
SageMaker AI	Full ML lifecycle: build, train, deploy models incl. FMs; JumpStart hub; pipeline control	Most control, most work

07 IS GENERATIVE AI THE RIGHT TOOL?

GOOD FIT — REQUIREMENTS

work backward from the business problem

- Diverse, **unstructured** data sources
- Synthesizing complex information
- Creative content, human-like text, high-res images
- Repetitive content creation, done efficiently
- Personalization at scale

POOR FIT — WARNING SIGNS

choose another approach when…

- Ethical risks can’t be tolerated or mitigated
- Accuracy and consistency are critical
- Explainability required — many FMs are **black boxes**
- Ill-defined or constantly changing problem
- Not enough high-quality data, skills, or clear ROI for the cost

08 EXAM TRAPS

where the knowledge check gets you

- × "Tokens and embeddings are the same thing." — Tokens are the **text chunks**; embeddings are the **numeric vectors** tokens are converted into. (It's *embedding* and *vector* that are used interchangeably.)
- × "Diffusion models pit a generator against a discriminator." — That's a **GAN**. Diffusion models add noise, then learn to **reverse** the process.
- × "Making a model more deterministic increases creativity." — Backwards. More deterministic = more **restrictive** = more accurate and conservative; *less* deterministic = more creative.
- × "If an FM won't give the output you need, build your own model." — Not yet: enrich the **prompt**, add an **external knowledge base**, or run **additional training** on a task-specific dataset first.
- × "Amazon Q is the service for building custom genAI applications." — **Bedrock** is for building applications; Amazon Q delivers generative AI **without** building one.
- × "SageMaker JumpStart is a Bedrock feature." — JumpStart is the **SageMaker AI** ML hub of publicly available FMs.
- × "The task looks like a fit, so generative AI is the right choice." — Also weigh ethics, accuracy, explainability, data quality, skills, and **business value vs. cost** — a less complex approach may meet the need.
- × "PartyRock needs an AWS account and suits production apps." — Sign in with an **Amazon, Google, or Apple ID**; it's a free playground for experimenting, **not for scalable applications**.

09 SELF-QUIZ

answers are upside-down below

Q1 What is a prompt?

Q2 Which model architecture generates data by adding noise to a clean input and learning to reverse the process?

Q3 Name the two competing neural networks inside a GAN.

Q4 What is the numerical vector representation of a token called?

Q5 List the four challenges of FM-generated output.

Q6 You need consistent, accurate answers for an HR assistant. Which direction do you adjust inference parameters?

Q7 Name the three ways to influence an FM with additional data, least to most complex.

Q8 What are the four key components for getting optimal outputs from an FM?

Q9 Which AWS service lets you build, train, and deploy your own FMs across the full ML lifecycle?

Q10 Sample exam: a team wants the MOST efficient way to access and evaluate multiple FMs for comparison without managing infrastructure. Which service?

ANSWER KEY (rotate the page)

Q1 a natural-language instruction or question asking a generative AI model to perform a specific task **Q2** diffusion model **Q3** generator and discriminator **Q4** an embedding **Q5** hallucinations, toxicity, bias, illegality (IP imitation) **Q6** more deterministic (more restrictive) **Q7** richer prompts → external knowledge base → additional training on a task-specific dataset **Q8** model choice, prompt design, data quality, safeguards **Q9** Amazon SageMaker AI **Q10** Amazon Bedrock