Generative AI: Foundation Models and Platforms

Lesson Summary

📌 Core Concepts Learned

1. Generative AI Overview

- **Definition**: AI systems capable of generating new content (text, images, code, etc.) that resembles human-created output.
- **Human-like Tasks**: Large Language Models (LLMs) and other generative models can write, summarize, translate, solve equations, generate code, and more.

---

2. Key Technologies Behind Generative AI

Model Type	Core Function	Key Feature
VAE (Variational Autoencoder)	Compresses and reconstructs data	Reduces data dimensionality effectively
GAN (Generative Adversarial Network)	Two networks compete (generator vs discriminator)	Produces highly realistic synthetic samples
Transformer	Uses an attention mechanism	Models long-range dependencies in text
Diffusion Model	Adds and removes noise through iterative processes	Great at high-quality image and data generation

---

3. Foundation Models

- **Definition**: Large, pre-trained models with billions of parameters.
- **Training**: Conducted on vast datasets using self-supervised learning.
- **Capabilities**:
    - Multimodal (text, image, audio, video)
    - Multidomain (applicable across industries)
    - Independent reasoning
    - Adaptable to new tasks and domains

Examples of Foundation Models:

- **GPT-4** (OpenAI): Excels in text generation, creative writing, code generation
- **Gemini** (Google): Strong in logic, math, and multimodal interaction
- **Stable Diffusion**: Generates realistic and stylized images from text prompts

---

4. Why Foundation Models Matter

- Act as **"building blocks"** of generative AI systems
- Allow developers and businesses to:
    - Reduce time-to-value (from months to weeks)
    - Customize AI solutions without training from scratch
    - Leverage cloud or edge platforms for scalability and privacy

---

Generative AI: Foundation Models and Platforms — Glossary

Term	Definition
Artificial neural networks (ANNs)	A collection of smaller computing units called neurons which are modeled in a manner similar to how a human brain processes information.
Bidirectional autoregressive transformer model (BART)	A text-to-text transfer transformer model developed by Facebook AI with a seq2seq translation architecture with bidirectional encoder representation like BERT and a left-to-right decoder like GPT.
Bidirectional encoder representations from transformers (BERT)	A family of language models by Google that uses pre-training and fine-tuning to create models that can accomplish several tasks.
Chatbot	A computer program that simulates human conversation with an end user. Though not all chatbots are equipped with artificial intelligence (AI), modern chatbots increasingly use conversational AI techniques like natural language processing (NLP) to make sense of the user's questions and automate their responses.
Clustering	An application of unsupervised learning wherein the algorithms group similar instances together based on their inherent properties.
Code2Seq	A text-to-code seq2seq model developed by OpenAI trained on a substantial text and code data set. It leverages the syntactic structure of programming languages to encode source code.
CodeT5	A text-to-code seq2seq model developed by Google AI trained on a large data set of text and code. CodeT5 is the first pre-trained programming language model that is code-aware and encoder-decoder based.

Convolutional neural networks (CNNs)	Deep learning architecture networks that contain a series of layers, each conducting a convolution or mathematical operation on a previous layer.
DALL-E	A text-to-image generation model developed by OpenAI that is trained on a large data set of text and images and can be used to generate realistic images from various text descriptions.
Deep learning	A type of machine learning focused on training computers to perform tasks through learning from data. It uses artificial neural networks.
Diffusion model	A type of generative model that is popularly used for generating high-quality samples and performing various tasks, including image synthesis. It is trained by gradually adding noise to an image and then learning to remove the noise. This process is called diffusion.
Dimensionality reduction	An application of unsupervised learning wherein the algorithms capture the most essential data features while discarding redundant or less informative ones.
Falcon	A large language model developed by the Technology Institute of Innovation (TII). Its variant, falcon-7b-instruct, is a 7-billion-parameter model based on the decoder-only model.
Foundational models	AI models with broad capabilities that can be adapted to create more specialized models or tools for specific use cases.
Generative adversarial network (GAN)	A type of generative model that includes two neural networks: Generator and discriminator. The generator is trained on vast data sets to create samples like text and images. The discriminator tries to distinguish whether the sample is real or fake.

Generative AI models	Models that can understand the context of input content to generate new content. In general, they are used for automated content creation and interactive communication.
Generative pre-trained transformer (GPT)	A series of large language models developed by OpenAI designed to understand language by leveraging a combination of two concepts: Training and transformers.
Google Flan	An encoder-decoder foundation model based on the T5 architecture.
Google JAX	A machine learning framework used for transforming numerical functions that combines autograd (automatic obtaining of the gradient function through differentiation of a function) as well as TensorFlow's XLA (accelerated linear algebra).
Hugging Face	An AI platform that allows open-source scientists, entrepreneurs, developers, and individuals to collaborate and build personalized machine learning tools and models.
IBM Granite	Multi-size foundation models that are specially designed for businesses. These models use a decoder architecture to apply generative AI to both language and code.
IBM watsonx	An integrated AI and data platform with a set of AI assistants designed to scale and accelerate the impact of AI with trusted data across businesses.
Imagen	A text-to-image generation model developed by Google AI trained on a large data set of text and images. Imagen is used to generate realistic images from various text descriptions.

Large language models (LLMs)	A deep learning model trained on substantial text data to learn the patterns and structures of language. They can perform language-related tasks, including text generation, translation, summarization, sentiment analysis, and more.
LLaMA	A large language model from Meta AI.
Natural language processing (NLP)	A subset of artificial intelligence that enables computers to understand, manipulate, and generate human language (natural language).
Neural code generation	A process that uses artificial neural networks like neural networks work in the human brain.
Neural network model	A type of text-to-text generation model that uses artificial neural networks to generate text.
Neural networks	Computational models inspired by the human brain's structure and functioning. They are a fundamental component of deep learning and artificial intelligence.
Open lakehouse architecture	A data lakehouse architecture that combines elements of data lakes and data warehouses.
PanGu-Coder	A text-to-code transformer model developed by Microsoft Research. It is a pre-trained decoder-only language model that generates code from natural language descriptions.
Pre-trained models	A machine learning model trained on an extensive data set before being fine-tuned or adapted for a specific task or application. These models are a type of transfer learning where the knowledge gained from one task (the pre-training task) is leveraged to perform another task (the fine-tuning task).

Pre-training	A technique in which unsupervised algorithms are repeatedly given the liberty to make connections between diverse pieces of information.
Prompt	An instruction or question given to a generative AI model to generate new content.
PyTorch	An open-source machine learning framework based on the Torch library. This framework is used for applications such as computer vision and natural language processing.
Recurrent neural networks (RNNs)	Deep learning architecture designed to handle sequences of data by maintaining hidden states that capture information from previous steps in the sequence.
Seq2seq model	A text-to-text generation model that first encodes the input text into a sequence of numbers and then decodes this sequence into a new one, representing the generated text.
Statistical model	A type of text-to-text generation model that uses statistical techniques to generate text.
Supervised learning	A subset of AI and machine learning that uses labeled data sets to train algorithms to classify data or predict outcomes accurately.
T5	A text-to-text transfer transformer model developed by Google AI trained on a substantial data set of code and text. It can be used for various tasks, including summarization, translation, and question-answering.
TensorFlow	A free and open-source software library used for machine learning and artificial intelligence.

Text-to-code generation model	A type of machine learning model used to generate code from natural language descriptions. It uses generative AI to write code through neural code generation.
Text-to-image generation model	A type of machine learning model used to generate images from text descriptions. It uses generative AI to make meaning out of words and turn them into unique images.
Text-to-text generation model	A type of machine learning model used to generate text from a given input. It is trained on a large text corpus and is taught to learn patterns, grammar, and causal information. Using the given input, the models generate the new text.
Training data	Data (generally, large data sets that also have examples) used to teach a machine learning model.
Transformers	A deep learning architecture that uses an encoder-decoder mechanism. Transformers can generate coherent and contextually relevant text.
Unsupervised learning	A subset of machine learning and artificial intelligence that uses algorithms based on machine learning to analyze and cluster unlabeled data sets. These algorithms can discover hidden patterns or data groupings without human intervention.
Variational autoencoder (VAE)	A generative model that is a neural network model designed to learn the efficient representation of input data by encoding it into a smaller space and decoding it back to the original space.
watsonx.ai	A studio of integrated tools for working with generative AI capabilities powered by foundational models and building machine learning models.
watsonx.data	A massive, curated data repository that can be used to train and fine-tune models with a state-of-the-art data management system.

watsonx.governance	A powerful toolkit to direct, manage, and monitor your organization's AI activities.