1. Architecture
   1. architecture and models
      1. various generative AI architectures and models, such as transformers, recurrent neural networks (RNNs), diffusion models, generative adversarial networks (GANs) and variational autoencoders (VAEs).
   2. LLM
   3. Libraries and tools (Pytorch, Hugging face)
2. Data Preparation (To convert the data into a format that LLM’s understand)
   1. Data Preparation
      1. Tokenization
         1. tokenizers for word-based, character-based, and subword-based tokenization.
         2. implement tokenization using various libraries such as nltk, spaCy, BertTokenizer, and XLNetTokenizer.
      2. Data Loaders
         1. data loaders for training generative AI models and list the PyTorch libraries for preparing and handling data within data loaders
         2. Create a data loader and use the collate function to process batches of text.
   2. Implement Tokenization
   3. Creating an NLP data Loader

Generative AI refers to deep learning models that can generate high quality text images and other content based on the data they were trained on. These models are developed and trained to understand patterns and structures within existing data and apply the understanding to produce new and relevant data. You will use the various types to generate different types of content, such as text, image, and audio. Generative AI models are also used for generating three-dimensional or 3D objects and music. Content Creation, Condensing Documents, Language Translation, Chatbots and Virtual assistants, Data Analysis

Architecture and models

These include recurrent neural networks, or RNNs, transformers, generative adversarial networks or GANs, variational autoencoders or VAEs, and diffusion models.

1. RNN
   1. Recurrent Neural Networks (RNNs) are a type of artificial neural network designed to work with sequential or time-series data. Unlike regular neural networks, RNNs have loops in their structure that allow them to remember previous inputs, which helps influence current inputs and outputs. This ability to retain information over time makes RNNs ideal for tasks that involve sequences and time-based dependencies, such as language modelling and speech recognition.
2. Transformers
   1. Transformers are advanced deep learning models that can translate text and speech in near real time. They process data, such as words or numbers, through multiple layers, starting from the input layer, passing through hidden layers, and ending at the output layer. Unlike traditional models, transformers use a self-attention mechanism that helps them focus on the most important parts of the input, making understanding and decision-making more efficient. This mechanism also allows them to handle different parts of the input sequence at the same time, enabling faster and more efficient training. During fine-tuning, most of the pretrained transformer model stays unchanged, and only the final output layers are usually trained for the specific task, while the self-attention and other core layers remain fixed.
3. GAN
   1. Generative Adversarial Networks (GANs) are generative AI models made up of two parts: a generator and a discriminator. The generator creates fake samples, while the discriminator checks if these samples are real or fake by comparing them with real examples from the dataset. It then gives each sample a probability score indicating how likely it is to be real. This process works like a friendly competition where the generator tries to make its outputs look real, and the discriminator improves at spotting fakes. Over time, both models get better, making GANs especially useful for tasks like image and video generation.
4. VAE (works on encoder and decoder frameworl)
   1. Variational Autoencoders (VAEs) are models that learn the underlying patterns in data by encoding and decoding it. The encoder compresses the input into a simplified abstract space that captures the essential features, while the decoder uses this condensed representation to recreate the original data. Unlike standard autoencoders, VAEs represent data as probability distributions in a latent space, which allows them to generate new samples with similar characteristics. This approach can produce multiple possible outputs for the same input, reflecting real-world uncertainty. VAEs are particularly useful in creative fields like art and design.
5. Diffusion Model
   1. A diffusion model is a type of generative model trained to create images by learning how to remove noise from data that has been heavily distorted. It gradually reconstructs the original data by reversing the process of adding noise. Based on a given prompt, diffusion models can generate highly creative images by leveraging the statistical patterns learned from their training data. They are particularly useful for producing high-quality images from noisy or low-quality inputs, such as restoring old or damaged photographs.

* In summary, different AI architectures use different training approaches. RNNs rely on a loop-based design, transformers use a self-attention mechanism, GANs work through a competitive process between two networks, VAEs learn patterns by representing data in a latent space, and diffusion models generate outputs based on statistical properties of data. Generative AI models are also linked to reinforcement learning, which traditionally helps agents interact with their environment to maximize rewards. In generative AI, reinforcement learning is often used during training to fine-tune and optimize model performance for specific tasks.
* AI Hallucinations
  + In AI hallucinations, the model generates output that it presents as accurate but is seen as unrealistic, inaccurate, irrelevant, or nonsensical by humans. It is similar to the way humans experience hallucinations. AI hallucinations are strongly associated with LLMs. Factors such as biases in the training data, limited training, complexity of the model, and lack of human oversight can cause AI hallucinations. Also, the outputs generated by the AI models might not be based on the patterns the models learned from the training data.
  + Some of the problems caused are: Generation of inaccurate information, Creation of biased views or misleading information, Wrong input provided to sensitive applications, such as those used in autonomous vehicles or medical domain
  + Methods to mitigate hallucinations : Eliminating any bias in the training data and performing extensive training of the models on high-quality data, Avoiding manipulation of the inputs that are fed into the models, Ongoing evaluation and improvement of the models, Fine-tuning a pre-trained LLM on domain-specific data
* Libraries and Tools
  + There are various libraries and tools that you can use to develop NLP applications using generative AI. Some tools are PyTorch, TensorFlow, Hugging Face, LangChain, and Pydantic.
  + PyTorch is an open source deep learning framework. It is a Python-based library well-known for its ease of use, flexibility, and dynamic computation graphs.
  + TensorFlow is an open-source framework for machine learning and deep learning. It provides tools and libraries to facilitate the development and deployment of machine learning models.
  + The tight integration of TensorFlow with Keras provides a user-friendly high-level neural networks API, facilitating rapid prototyping and building and training deep learning models.
  + Hugging Face is a platform that offers an open-source library with pretrained models and tools to streamline the process of training and fine-tuning generative AI models. It offers libraries such as Transformers, Datasets, and Tokenizers.
  + LangChain is an open-source framework that helps streamline AI application development using LLMs. It provides tools for designing effective prompts.
  + Pydantic is a Python library that helps you streamline data handling. It ensures the accuracy of data types and formats before an application processes them.