**WHAT IS GENERATIVE AI?**

**Simple def**: Generative AI, sometimes called gen AI, is [artificial intelligence](https://www.ibm.com/think/topics/artificial-intelligence) (AI) that can create original content such as text, images, video, audio or software code in response to a user’s prompt or request

**Technical def:** Generative AI is sophisticated ML models “DEEP LEARNING” which are intended to mimic the human learning and decision making.  
The are trained on huge amount of data and they work by encoding and identifying patterns on that data, by this they understand the user input and try to answer them.

**What are Foundational Models?**  
They are large scale general purpose models (generalization refers to, they can do anything, after a bit of fine tuning or the patterns they’ve learned from the data sets) trained on huge amount of data and are capable of performing wide variety of tasks.   
As well as they can be adapted for specific applications, after fine-tuning.  
  
**-TYPES OF FOUNDATIONAL MODELS**

1. TEXT BASED: they handle natural language processing, i.e. giving text responses for text responses.
2. VISUALS BASED: they handle visual data like images and videos
3. MULTI MODEL: intended for processing multiple kinda data like images and text at once.
4. AUDIO BASED: handles data like music, audios, songs.
5. CODE BASED: handles programming data.
6. AGENT BASED: combines reasoning, decision making, and interaction capabilities.

**HOW IT WORKS?**It works in three phases namely, **TRAINING, TUNING, GENERATION**

1. TRAINING: foundational models (deep learning algorithms) train on huge amount of data and practitioners, make them take millions of “fill in the blanks” exercises. So, they become enable of predicting the next words line or thing.
2. TUNING: foundational models knows a lot but can’t perform specific tasks, so that’s why they are tuned for specific applications.  
   there are ways of TUNINGS, namely FINE TUNING, REINFORCEMENT.
3. GENERATION, EVALUATION AND TUNING: developers and users assess the outputs of the model and may tune the model again for greater accuracy and relevance.

**MODEL ARCHITECTURES**

1. VAEs: variation autoencoders, a deep learning algo that encodes large amount of data and decode it into multiple variations.
2. GANs: generative adversarial networks, works with 2 neural networks, one generator (to generate content) and second discriminator that evaluates the accuracy of the output. Mostly used in image and video generation.
3. DIFFUSION models: It works on this principal, that add noise to the data that it becomes unrecognizable, then iteratively diffuse that noise to achieve the outcome.
4. TRANSFORMERS: they also work with encoders and decoders, but the differentiating factor is “ATTENTION”, which helps them on focusing on specific parts or the most important one, which helps them focus on entire sequences, capture context and can work on multiple, inputs parallelly.

**BENEFITS OF GENAI**

1. ENHANCED CREATIVITY: generates original content and help professionals like artists, designers etc. explore fresh ideas.
2. ACCELERATES RESEARCH & DEV: gen ai helps in reducing research times, helping speeding up innovation.
3. EMPOWERING NON-EXPERTS: It enables even non-experts to generate quality content in various domains.

**LIMITATIONS OF GENAI**

1. DEPENDANT ON DATA: The amount of relevance and accuracy of output depends on the data it is trained on.
2. LIMITED CONTROL OVER OUTPUTS: we don’t have much control over outputs, which sometimes can result in irrelevant outputs, leading to hallucination.
3. HIGH COMPUTATIONAL COSTS: Training and running a genai model is resource intensive task, models demand significant computing power. So, it becomes too costly.
4. ETHICAL CONCERNS: Generative ai can be abused in things like creating deepfakes and harmful content, resulting in spread of misinformation or privacy violations.