Generative artificial intelligence (Generative AI, GenAI, or GAI) is a subfield of artificial intelligence that uses generative models to produce text, images, videos, audio, software code or other forms of data.These models learn the underlying patterns and structures of their training data and use them to produce new data based on the input, which often comes in the form of natural language prompts.



Generative AI tools have become more common since the AI boom in the 2020s. This boom was made possible by improvements in transformer-based deep neural networks, particularly large language models (LLMs). Major tools include chatbots such as ChatGPT, Copilot, Gemini, Claude, Grok, and DeepSeek; text-to-image models such as Stable Diffusion, Midjourney, and DALL-E; and text-to-video models such as Veo and Sora.Technology companies developing generative AI include OpenAI, xAI, Anthropic, Meta AI, Microsoft, Google, DeepSeek, Baidu and Yandex.

Generative AI is used across many industries, including software development,healthcare,finance,entertainment,customer service, sales and marketing, art, writing, fashion, and product design. The production of generative AI systems requires large scale data centers using specialized chips which require a lot of electricity for processing and water for cooling.

Generative AI has raised many ethical questions and governance challenges as it can be used for cybercrime, or to deceive or manipulate people through fake news or deepfakes.Even if used ethically, it may lead to mass replacement of human jobs. The tools themselves have been criticized as violating intellectual property laws, since they are trained on copyrighted works. The material and energy intensity of the AI systems has raised concerns about the environmental impact of AI, especially in light of the challenges created by the energy transition.



Types of Generative AI Models

1. Transformers or Autoregressive Models

Autoregressive Transformers Models generate sequences by predicting the next token based on all previous ones moving step by step through the text.

The architecture relies on the transformer’s self attention mechanism to capture context from the entire input so far making it highly effective for natural language and code generation.

Popular examples include GPT models which can produce coherent, context aware paragraphs, solve coding tasks or answer complex queries.

The autoregressive approach gives fine grained control over each output step but can be slower for long generations since tokens are generated one at a time.

2. Diffusion Models

Diffusion models generate data such as images or audio by starting with pure random noise and gradually refining it into a coherent output through a series of denoising steps.

Each step reverses a simulated diffusion process that added noise to real data during training.

This iterative approach can produce highly detailed and realistic results specially in image synthesis where models like Stable Diffusion and DALL·E 3 have set benchmarks.

Diffusion models are also versatile they can be adapted for inpainting, style transfer and conditional generation from text prompts.

3. Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs)

VAEs and GANs were among the first deep learning architectures for generative tasks.

A VAE encodes data into a compressed latent space and then decodes it back with a probabilistic twist that encourages smooth, continuous representations. This makes them good for controllable generation and interpolation between styles.

GANs in contrast use two networks against each other a generator that tries to produce realistic outputs and a discriminator that tries to detect fakes.

This adversarial setup leads to sharp, lifelike images though training can be unstable and prone to mode collapse.

4. Encoder Decoder Models

Encoder decoder architectures consist of two stages: the encoder processes the input into a dense representation and the decoder generates the desired output from that representation.

They are widely used for sequence to sequence tasks like language translation, summarization and image captioning.

The encoder captures the full context of the input before the decoder starts producing tokens, allowing for strong performance on tasks that require global understanding rather than token by token prediction.

Modern encoder decoder models often use transformers for both stages as in T5, BART and many multimodal system.