Generative AI refers to deep-learning models that can take raw data — say, all of Wikipedia or the collected works of Rembrandt — and “learn” to generate statistically probable outputs when prompted. At a high level, generative models encode a simplified representation of their training data and draw from it to create a new work that’s similar, but not identical, to the original data.

Generative models have been used for years in statistics to analyze numerical data. The rise of deep learning, however, made it possible to extend them to images, speech, and other complex data types. Among the first class of models to achieve this cross-over feat were variational autoencoders, or VAEs, introduced in 2013. VAEs were the first deep-learning models to be widely used for generating realistic images and speech.

“VAEs opened the floodgates to deep generative modeling by making models easier to scale,” said Akash Srivastava, an expert on generative AI at the MIT-IBM Watson AI Lab. “Much of what we think of today as generative AI started here.”

Autoencoders work by encoding unlabeled data into a compressed representation, and then decoding the data back into its original form. “Plain” autoencoders were used for a variety of purposes, including reconstructing corrupted or blurry images. Variational autoencoders added the critical ability to not just reconstruct data, but to output variations on the original data.

This ability to generate novel data ignited a rapid-fire succession of new technologies, from generative adversarial networks (GANs) to diffusion models, capable of producing ever more realistic — but fake — images. In this way, VAEs set the stage for today’s generative AI.

They are built out of blocks of encoders and decoders, an architecture that also underpins today’s large language models. Encoders compress a dataset into a dense representation, arranging similar data points closer together in an abstract space. Decoders sample from this space to create something new while preserving the dataset’s most important features.

Transformers, introduced by Google in 2017 in a landmark paper “Attention Is All You Need,” combined the encoder-decoder architecture with a text-processing mechanism called attention to change how language models were trained. An encoder converts raw unannotated text into representations known as embeddings; the decoder takes these embeddings together with previous outputs of the model, and successively predicts each word in a sentence.

Through fill-in-the-blank guessing games, the encoder learns how words and sentences relate to each other, building up a powerful representation of language without anyone having to label parts of speech and other grammatical features. Transformers, in fact, can be pre-trained at the outset without a particular task in mind. Once these powerful representations are learned, the models can later be specialized — with much less data — to perform a given task.

Several innovations made this possible. Transformers processed words in a sentence all at once, allowing text to be processed in parallel, speeding up training. Earlier techniques like recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks processed words one by one. Transformers also learned the positions of words and their relationships, context that allowed them to infer meaning and disambiguate words like “it” in long sentences.

By eliminating the need to define a task upfront, transformers made it practical to pre-train language models on vast amounts of raw text, allowing them to grow dramatically in size. Previously, people gathered and labeled data to train one model on a specific task. With transformers, you could train one model on a massive amount of data and then adapt it to multiple tasks by fine-tuning it on a small amount of labeled task-specific data.

Transformers have come to be known as foundation models for their versatility. “If you wanted to improve a classifier, you used to have to feed it more labeled data,” said Srivastava. “Now, with foundation models, you can feed the model large amounts of unlabeled data to learn a representation that generalizes well to many tasks.”

Language transformers today are used for non-generative tasks like classification and entity extraction as well as generative tasks like translation, summarization, and question answering. More recently, transformers have stunned the world with their capacity to generate convincing dialogue, essays, and other content.

Language transformers fall into three main categories: encoder-only models, decoder-only models, and encoder-decoder models.

Encoder-only models like BERT power search engines and customer-service chatbots, including IBM’s Watson Assistant. Encoder-only models are widely used for non-generative tasks like classifying customer feedback and extracting information from long documents. In a project with NASA, IBM is building an encoder-only model to mine millions of earth-science journals for new knowledge.

Decoder-only models like the GPT family of models are trained to predict the next word without an encoded representation. GPT-3, at 175 billion parameters, was the largest language model of its kind when OpenAI released it in 2020. Other massive models — Google’s PaLM (540 billion parameters) and open-access BLOOM (176 billion parameters), among others, have since joined the scene.

Encoder-decoder models, like Google’s Text-to-Text Transfer Transformer, or T5, combine features of both BERT and GPT-style models. They can do many of the generative tasks that decoder-only models can, but their compact size makes them faster and cheaper to tune and serve.

Generative AI and large language models have been progressing at a dizzying pace, with new models, architectures, and innovations appearing almost daily.