Transformer architecture, which is essential for modern large language models (LLMs) such as GPT, Gemini, and Llama. It begins by contrasting transformers with earlier models like recurrent neural networks (RNNs), which processed text sequentially and struggled to capture full context. Transformers revolutionized this by allowing simultaneous processing of sentences, enabling a more comprehensive understanding of text.

A notable feature of transformers is self-attention, which evaluates the importance of each word relative to others in the sentence. For example, in the sentence "The cat sat on the mat," the model can identify that 'cat' is more closely related to 'sat.' The lecture highlights the influential paper "Attention is All You Need" by Google in 2017, which introduced the transformer model for machine translation and later applications like text generation.

Transformers consist of an encoder and a decoder. The encoder transforms input text into contextual embeddings through multiple layers of self-attention and feed-forward networks, allowing a bidirectional understanding of context. The decoder generates output sequences, using masked self-attention to predict words one at a time. The relationship between the two is compared to that of a reader (encoder) and a writer (decoder).

Self-attention is explained through query, key, and value vectors, which assess word relationships, while positional encodings ensure word order is maintained. Transformers are trained initially on large text corpora for general language understanding and are then fine-tuned on specific datasets for targeted tasks, making them versatile for various applications.