# Transformer Architecture: A Survey

The transformer architecture, introduced by Vaswani et al. in 2017, has revolutionized natural language processing and beyond.

1. Self-Attention Mechanism
The core innovation of transformers is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input sequence. Unlike RNNs, transformers process all positions simultaneously, enabling massive parallelization.

2. Multi-Head Attention
Rather than performing a single attention function, multi-head attention linearly projects queries, keys, and values multiple times with different learned projections. This allows the model to jointly attend to information from different subspaces.

3. Positional Encoding
Since transformers contain no recurrence or convolution, positional encodings are added to give the model information about token positions. The original paper used sinusoidal functions.

4. Applications Beyond NLP
Transformers have expanded beyond NLP. Vision Transformers (ViT) apply the architecture to image classification. DALL-E uses transformer components for image generation. AlphaFold 2 leverages attention mechanisms for protein structure prediction.

5. Scaling Laws
Research has shown that transformer performance follows predictable scaling laws. Larger models trained on more data consistently perform better, leading to GPT-4, Claude, and Gemini.

6. Efficiency Improvements
Various techniques improve transformer efficiency: sparse attention, linear attention approximations, mixture-of-experts, and quantization. Flash Attention reduces memory usage significantly.

7. Fine-tuning and Adaptation
Transfer learning with transformers involves pre-training on large corpora followed by fine-tuning. Parameter-efficient methods like LoRA allow adaptation with minimal additional parameters.