# Large Language Models: Architecture and Processing

## Key Components of LLM Architecture

### Transformers

The transformer architecture serves as the backbone of modern large language models. Introduced in the seminal paper "Attention Is All You Need" (2017), transformers revolutionized natural language processing by replacing recurrent neural networks with a more parallelizable architecture. The transformer consists of encoder and decoder blocks, though many current LLMs use decoder-only architectures. Each block contains multiple layers of self-attention mechanisms and feed-forward networks, allowing the model to process entire sequences simultaneously rather than sequentially.

### Attention Mechanisms

Attention mechanisms enable LLMs to focus on relevant parts of the input sequence when processing each token. Self-attention allows every position in a sequence to attend to all positions, creating rich contextual representations. Multi-head attention runs several attention mechanisms in parallel, capturing different types of relationships between tokens. This mechanism is crucial for understanding long-range dependencies and contextual relationships that determine meaning in natural language.

### Embeddings

Embeddings convert discrete tokens into dense vector representations that capture semantic meaning. Token embeddings map vocabulary items to high-dimensional vectors, while positional embeddings encode the position of tokens within sequences. These embeddings are learned during training and form the foundation for all subsequent processing, enabling the model to understand both the meaning and context of words within sentences.

## Popular LLMs

**GPT (Generative Pre-trained Transformer)** models, including GPT-3 and GPT-4, use decoder-only transformer architectures optimized for text generation. They excel at producing coherent, contextually appropriate text across diverse tasks through autoregressive generation.

**BERT (Bidirectional Encoder Representations from Transformers)** employs an encoder-only architecture that processes text bidirectionally. This design makes BERT particularly effective for understanding tasks like question answering, sentiment analysis, and text classification.

**T5 (Text-to-Text Transfer Transformer)** frames all NLP tasks as text-to-text problems, using a full encoder-decoder architecture. This unified approach allows T5 to handle both understanding and generation tasks within the same framework.

## Text Processing and Generation

LLMs process text through tokenization, breaking input into discrete units called tokens that may represent words, subwords, or characters. During processing, the model converts tokens to embeddings, applies multiple transformer layers with attention mechanisms, and generates probability distributions over the vocabulary for next-token prediction.

For generation, LLMs use autoregressive decoding, predicting one token at a time based on previous context. Various sampling strategies like temperature scaling and top-k sampling control the randomness and creativity of generated outputs, enabling flexible text generation for different applications.

This architecture enables LLMs to capture complex linguistic patterns and generate human-like text across numerous domains and tasks.