**Transformers & Their Role in BERT, GPT, and LLMs**

*— By Divyanshu Gupta*

Transformers are a breakthrough deep learning architecture introduced in 2017 through the paper *“Attention is All You Need”*. Unlike older models like RNNs or LSTMs that process words one-by-one, transformers process entire sequences **in parallel**, making them highly efficient. What really sets them apart is their **self-attention mechanism**, which allows the model to weigh the importance of each word in a sentence relative to every other word. This makes them excellent at understanding language context, even across long distances in the text.

At the core of transformer models are four major components. First, **multi-head self-attention** enables the model to focus on different parts of a sentence simultaneously. Second, **feed-forward layers** help transform the attention outputs into meaningful representations. Since transformers don't work sequentially like RNNs, they rely on **positional encoding** to understand the order of words. Lastly, **residual connections and normalization (Add & Norm)** help keep the model stable and trainable even when it's deep.

Thanks to these capabilities, transformers have become the backbone of most **Large Language Models (LLMs)** today — like ChatGPT, Google Gemini, and Meta’s Llama. These models use massive datasets and transformer layers to understand and generate human-like language, code, summaries, and more.

**🔹 BERT (Bidirectional Encoder Representations from Transformers)**

BERT is a transformer-based model built using **only the encoder** part of the architecture. It reads text **bidirectionally**, meaning it looks at both the left and right context of a word to understand its meaning — making it great for tasks that involve understanding language rather than generating it. BERT is trained using two main strategies: **Masked Language Modelling (MLM)**, where some words are hidden and the model has to guess them; and **Next Sentence Prediction (NSP)**, where the model predicts if one sentence follows another logically. It’s widely used for tasks like sentiment analysis, text classification, and question answering.

**🔹 GPT (Generative Pre-trained Transformer)**

GPT, on the other hand, is based on the **decoder** part of the transformer and works **left to right** (unidirectionally). It’s designed to **generate** text — meaning it predicts the next word in a sentence based on all the previous words. GPT models are trained on large corpora using **causal language modelling**, which helps them produce fluent, coherent, and often creative text. They are used in applications like AI chatbots, content writing, email drafting, and even code generation.

To help compare the two more clearly:

| **Feature** | **BERT** | **GPT** |
| --- | --- | --- |
| Role | Understand text | Generate text |
| Direction | Bidirectional | Unidirectional |
| Architecture | Encoder | Decoder |

Both models represent different strengths of transformers — BERT is better for understanding, while GPT excels at generation. Together, they demonstrate the flexibility and power of the transformer architecture, which is now the standard for modern NLP systems and LLMs.

In conclusion, transformers have redefined what machines can do with human language. Whether it's understanding context in a sentence or generating long, coherent paragraphs, transformer-based models like BERT and GPT make it possible. As students entering the AI/ML field, knowing how these models work gives us a strong foundation for building smarter, language-aware applications.