Transformer Fundamentals

Large Language Models (LLMs) process text by first breaking it down into **tokens**, which are the atomic units of meaning (words, characters, or sub-words). Each token is converted into a numerical vector (embedding). These vectors then pass through a stack of **Transformer blocks**, the core engine of the model. Inside each block, the **Attention mechanism** allows every token to "look at" every other token in the sequence to determine which ones are most relevant to its own context. For example, in the phrase "the bank of the river," attention helps the word "bank" connect more strongly with "river" than with financial concepts.

After the attention layer has captured these relationships, a **Feed-Forward Network (FFN)** within the same block processes each token individually to refine its representation. This cycle of "relating" and "refining" repeats across multiple layers—often dozens or even hundreds in models like GPT-4. By the time the data reaches the final layer, the model has a high-dimensional understanding of the entire input, allowing it to predict the most probable next token as the **output**.

FlowChart