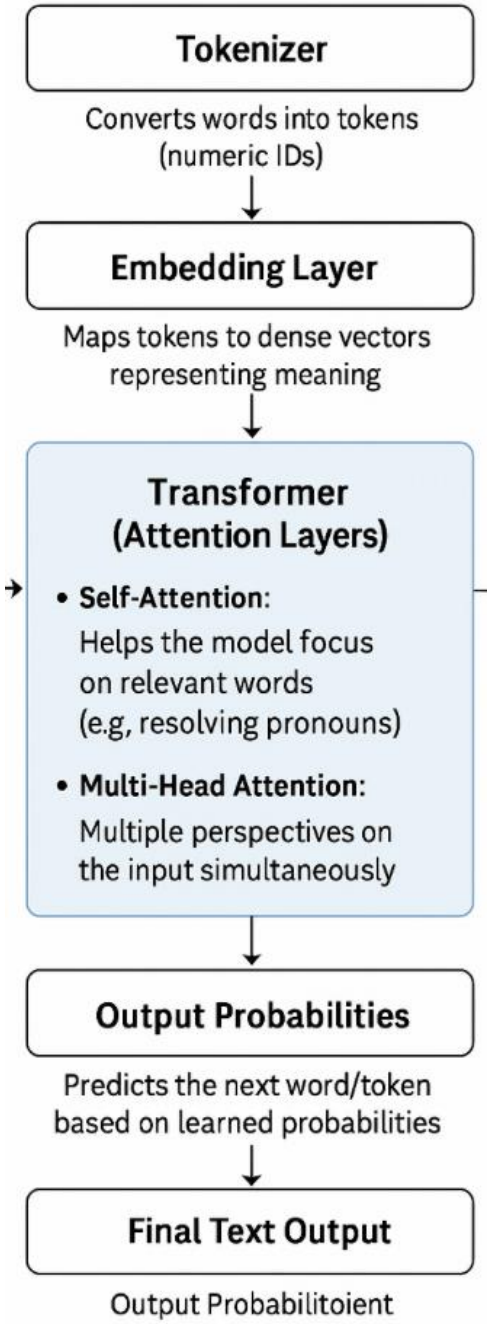
Lab 13 – Part A

How Does a Large Language Model (LLM) Work?

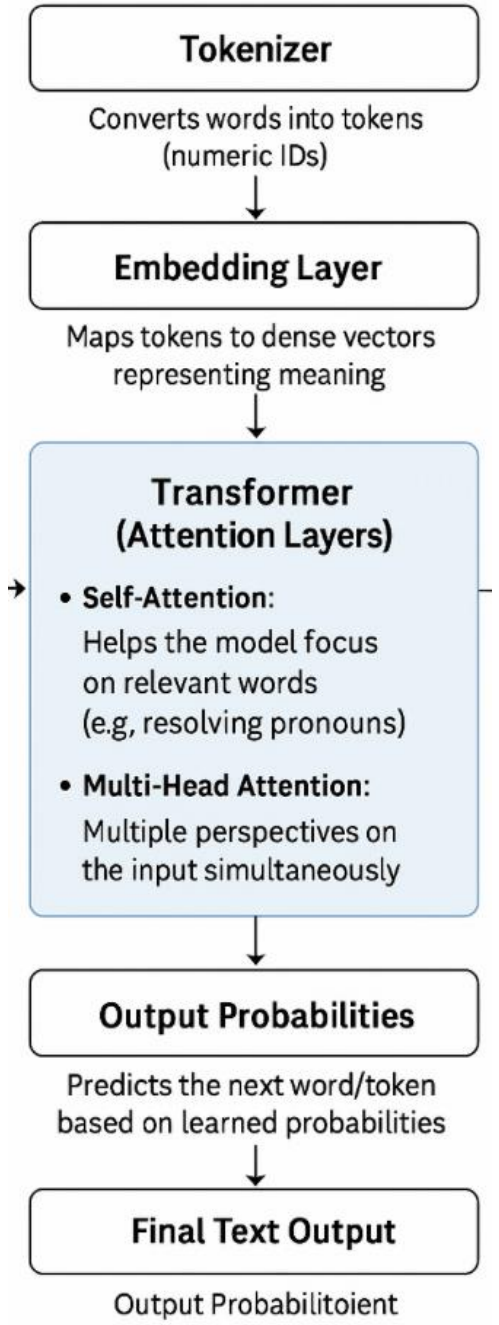
Overview

Tokenizer
Converts words into tokens (numeric IDs).
Embedding Layer
Maps tokens to dense vectors representing meaning.



Transformer Architecture (Core)

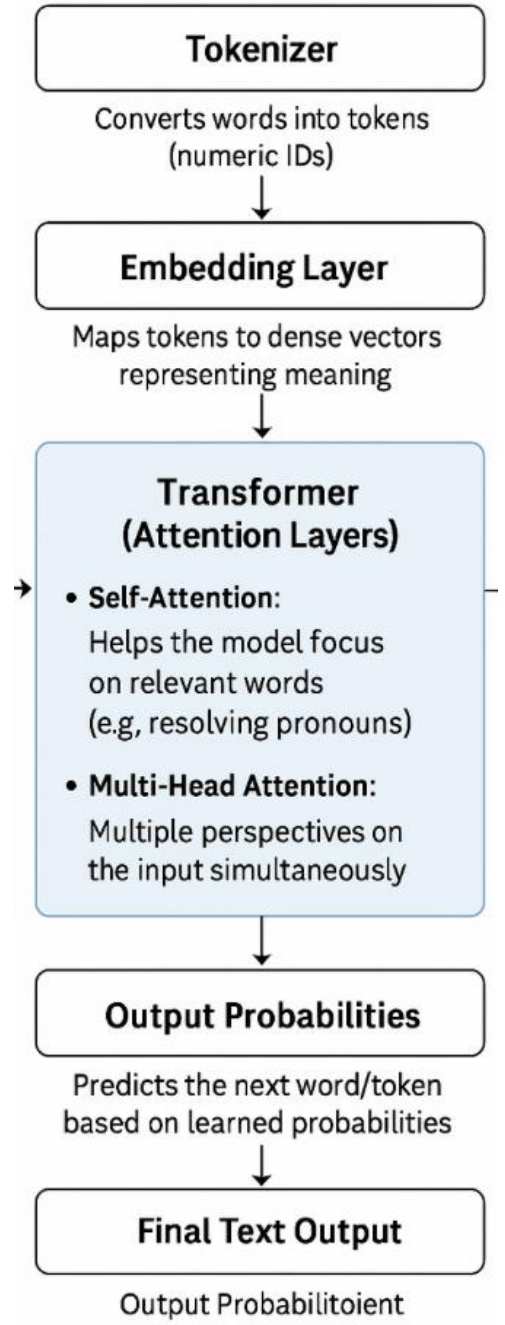
- **Self-Attention:** Helps the model focus on relevant words (e.g., resolving pronouns).
- **Multi-Head Attention:** Multiple perspectives on the input simultaneously.
- **Feedforward Layers:** Adds depth and learning power.
- **Stacked Layers:** More layers = deeper understanding.



Output Layer

Predicts the next word/token based on learned probabilities.

- **Training Principle**
  Learns by predicting the next word in a sentence—adjusts weights using massive datasets (unsupervised learning).



Important Concepts

- **Context Window:** The amount of input text LLM can “remember.”
- **Temperature & Top-p:** Control randomness and creativity of output.