LLM Architecture at a High Level

Output Probabilities

Softmax

Linear

DECODER   Nx

Nx   ENCODER

Positional Encoding

Positional Encoding

Input Embedding

Output Embedding

High-Level Transformer Architecture

Output Probabilities

Softmax turns outputs into 0-1 probability distribution over vocabulary

Smoothens optimization: residual layers and layer normalization

Decoder "cross attention" to encoder

Repeated (stacked) attention blocks to "deep learn" hierarchical abstractions

Multi-head Self-Attention learns different patterns

Masked attention to predict future tokens in decoder

Inputs and positional encoding

Inputs

Outputs (shifted right)