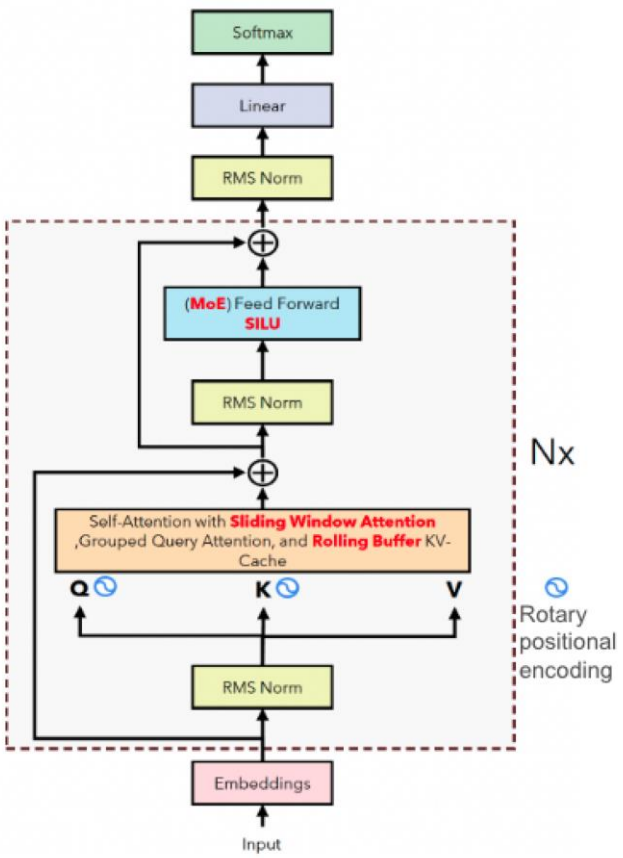
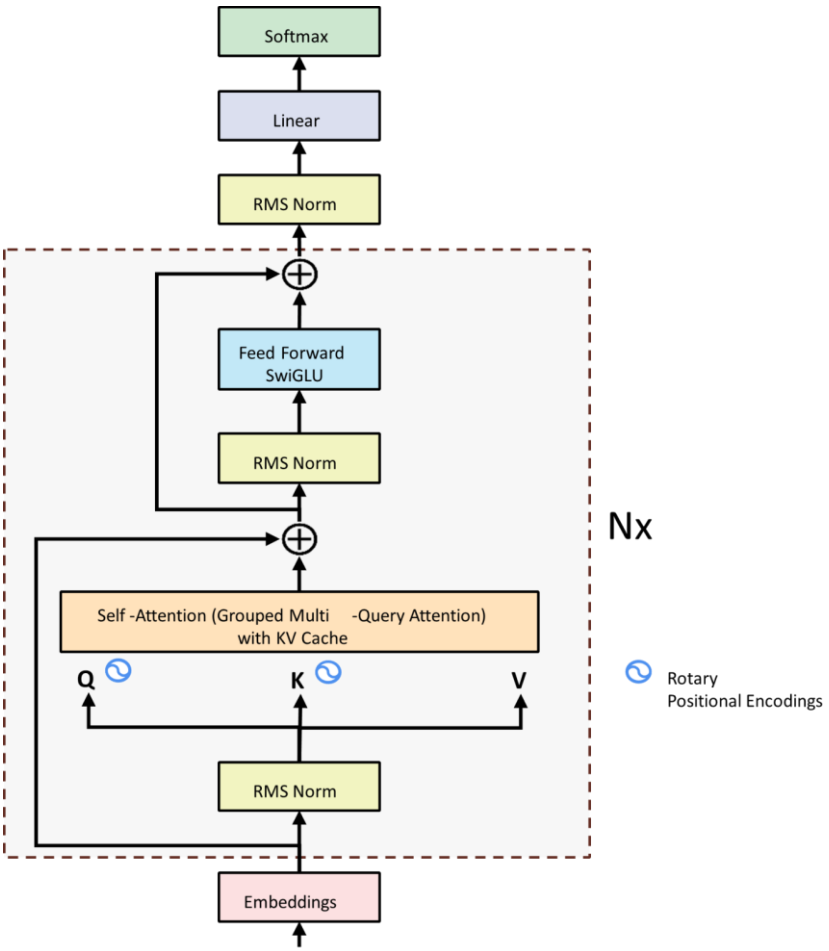
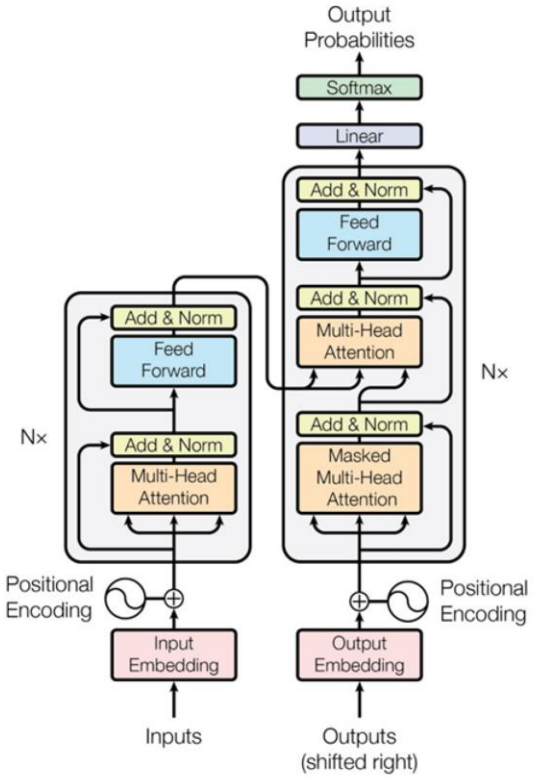
Mistral



LLaMA-2 and LLaMA-3



Transformers



Feature	Plain Transformers	Llama	Mistral
Architecture	Encoder-Decoder	Single Layer	Single Layer + Enhancements
Positional Encoding	Standard	Rotary	Rotary + Additional Innovations
Normalization	LayerNorm	RMS Norm	RMS Norm + Enhanced Mechanisms
Attention Mechanism	Self-Attention	Grouped Multi-Query Attention	Sliding Window Attention
Memory Efficiency	N/A	KV Cache	Rolling Buffer KV Cache
Activation in Feedforward	RELU	SwiGLU	SwiGLU + Mixture of Experts