

Assignment - 3

Q1) Decoder-only Transformer Vs Encoder-Decoder Architecture

While both the models ^{can} generate and also predict text but Encoder-Decoder Architecture is mainly used for Tasks like Language Translation where it uses cross-attention. This requires additional computation as the number of parameters increases and also additional memory storage. Decoder-only Transformer are natively autoregressive whereas encoder-decoder models are designed for conditional generation i.e. it requires an input as well as an output. Decoders Scale efficiently with longer context whereas the Encoder-Decoder scales poorly and the cross-attention becomes a bottleneck. Decoders also enable efficient inference.

Q2) The Next Token Prediction learns the full probability Model of Language. NTP is self-supervised. All the Natural Language processing tasks like translation, summarization and reasoning appear different but they all exist naturally as text continuation. As the model Scale increases, efficient compression of language data leads to the emergence of abstraction and generalization, making advanced language abilities an emergent property of next-token prediction.