Ass-3

DATE___ PAGE___

1) Decoder, Transformers are more scallable for LLM's because they use single transformer with self-attention (Causal) which directly matches the next token prediction objective used in large-scale language modeling.

This avoides extra encoder stack and cross attention required in encoder-decoder architectures.

2) Next-Token prediction is sufficient because many NLP tasks can be expressed as predicting the next word given a context.

When trained on massive text, the model learns the structural of language.

To correctly predict the next token, the model must learn meaning and logical independencies.

A single-next token objective acts as a universal training signal that enables the emergence of complex language abilities.