Date __/__/____

Assignment-3

1. Decoder only Transformers are more scalable for large language models for LLMs primarily due to architectural simplicity, training efficiency, inference efficiency and better utilization of large scale data.

Encoder- Decoder model requires
→ Full bidirectional encoding of input
→ Cross-attention
This results in
→ Higher memory usage
→ Increased compute cost

Decoder only models:
→ learn directly from raw text
→ Do not ~~text~~ require task specific formatting

Decoder only transformer scale better because they are simpler because they are simple, more compute-efficient, better aligned with large unlabeled data and optimized for fast autoregressive inference

2. Next-Token Prediction is sufficient because language itself encodes structure, logic & semantics and predicting next token forces the model to internalize properties.

Human written text already includes
→ Translations, summaries
By minimizing next token loss, model learns
→ How arguments unfold.

Page No.

Date __/__/____

To predict correct future tokens in:
→ Math proofs
→ Code
→ Multi step explanation

Since NTP conditions on context, model naturally learns these transformations into task specific objectives

as data & model size increases
→ Representations become more abstract
→ Patterns generalize across tasks

This explains why larger models suddenly exhibit
→ few shot learning
→ Instruction following

Next TOKEN Predictions works because it forces models to compress the statistical, semantic, and logical structure of language into a single predictive tasks