Assignment 3
Architecting Intelligence

Azaad Katiyar (240249)

January 12, 2026

1 Questions

Q1. Why Encoders got fired ?
Conceptually, why is a Decoder-only Transformer more scalable for LLMs than Encoder–Decoder architectures ?

Answer:

Encoder-Decoder transformers were initially proposed for sequence-to-sequence tasks such as translation where the source sequence and target sequence are not the same. However, the dominant task in Large Language Models is next-word prediction in a large body of text data; hence, the Decoder-only transformation is more appropriate.

Decoder-only models consist of one network based on causal self-attention, which is much simpler and more efficient to train. Encoder-Decoder models consist of two networks and are based on cross-attention.

Another major advantage here is that the decoder-only model also supports attention caching, in that it can use previous calculations in case it has to predict a long sequence of outputs.

Due to the simplicity, efficiency, and improved scalability for longer sequences of the Decoder-only Transformers architecture, it is preferred in the case of modern LLMs.

Q2. Weaponized Autocomplete
Why is Next-Token Prediction sufficient to learn complex language abilities such as reasoning, translation, and summarization ?

Answer:

Next-Token Prediction is successful because language contains many tasks within itself. When a machine learning model is trained on a huge text, the model gets to observe examples of translation, summarization, reasoning, and explanation in text form naturally occurring in language. Through next-token prediction, the model learns the underlying structures in these tasks.

For example, translated sentences, summaries, and logical explanations are all represented in the training data as text continued. However, for one to be able to predict the word that follows, one needs to comprehend grammar, meaning, and logic. This will result in the skill of reasoning and summarizing.

On the other hand, reasoning is also expressed in steps form in writing. Predictions about the next word require that the model trace logical paths.

So, in a way, having enough data and scaling appropriately allows a simple objective—”predict the next token in a sequence”—to model a wide range of complex phenomena in language without task-specific training.