Assignment-3

Q.1) The shift toward Decoder-only architectures for large language Models (LLMs) is driven by three main factors.

i) **Training Efficiency:-** In a Decoder-only setup, every token in a sequence is used as a training signal (predicting the next token). In Encoder-Decoder models, the encoder's tokens only serve as context, making the training process less computationally "dense" per flop.

ii) **KV caching & Inference:-** Decoders use Key-Value (KV) Caching, which stores previous computations to avoid redundant work.

- Decoder-Only → The cache grows linearly within a single unified stack.
- Encoder-Decoder → Requires managing separate self-attention and cross-attention catches, increasing memory management complexity and latency during auto-regressive generation.

iii) **Simpler Scaling:-**

- **Unified Information Flow:-** Without the "bottleneck" of a separate encoder and cross-attention bridge, the model can more easily scale to massive content lengths (e.g. → 100K+ tokens).
- **Zero-Shot Versatility:-** Decoder-only models are naturally task-agnostic. By treating everything as a continuation of a prompt, they exhibit better emergent behaviors for diverse tasks compared to the "input-output" structure of Encoder-Decoders.

Q.2) Next-Token Prediction (NTP) is sufficient, because it forces the model to develop an internal world model to minimize prediction error.

i) Compression as Intelligence:- To predict the next token accurately, a model cannot simply memorize; it must understand the underlying latent structure of the data.

- Logical Reasoning → Predicting the conclusion of a proof requires the model to learn the rules of logic.
- Semantic Understanding → To predict a word like "photosynthesis," the model must capture the relationship between sunlight, plants, and energy.

ii) Task-Agnostic Generalization:- NTP treats all NLP tasks as a single, unified problem of sequence continuation

- Translation → learned by predicting the next token in multilingual contexts (e.g. English: "Apple", French: "Pomme")
- Summarization → Learned by predicting tokens that condense preceding information.
- Reasoning → Emerges as the model learns to follow step-by-step "chains of thought" present in its training data.

iii) High-Density Training Signal:- NTP provides a supervision signal for every single token in a dataset. This massive, constant feedback loop allows the model to capture extremely subtle nuance in grammar, style, and factuality that task-specific objectives often miss.