Lab 13 Part A

How a Large Language Model (LLM) Works

A Large Language Model (LLM) like GPT is a type of artificial intelligence trained to understand and generate human-like text. It uses a transformer architecture with three main components:

- **Embedding Layer**: Converts words into numerical vectors. - **Transformer Blocks**: Each block contains attention mechanisms and feedforward networks that analyze relationships between words. - **Output Layer**: Translates processed vectors back into text.

LLMs are trained on large text datasets to learn grammar, facts, reasoning, and writing styles. During inference, the model predicts the next word/token based on the previous ones using probabilities.