Assument #2

A **large language model (LLM)** is a type of artificial intelligence (AI) model designed to process and generate human-like text. These models are based on deep learning techniques

**Key Features of LLMs:**

* **Trained on massive datasets**: They learn from vast amounts of text data from books, articles, and

websites.

**Capable of understanding context**: They generate coherent and contextually relevant responses.

* **Used in various applications**: Chatbots, content generation, code writing, and even medical or

legal text processing.

**How They Work**

LLMs use **deep learning**, specifically **transformer architectures**, to process and generate human-like text. They learn patterns, grammar, and contextual meaning from large datasets. The most well-known transformer model is **GPT (Generative Pre-trained Transformer)**, developed by OpenAI.

LLMs follow a two-step process:

* **Pre-training**: The model is trained on massive amounts of text to learn language patterns.
* **Fine-tuning**: It is then adapted for specific tasks like answering questions, summarization, or translation.

models

***t5 model***

|  |  |
| --- | --- |
| Feature | **T5 (2019, by Google)** |
| **Model Type** | **Seq2Seq (Encoder-Decoder) Transformer** |
| **Training Objective** | **Text-to-Text Framework** – Every NLP task is framed as a text generation task (e.g., translation, summarization, classification) |
| **Architecture** | **Encoder-Decoder** (like BERT for encoding + GPT for decoding) |
| **Context Understanding** | **Processes full input, then generates output** (good for both understanding and generation tasks) |
| **Fine-tuning Usage** | **Very flexible** – Can handle classification, Q&A, summarization, translation, and more |
| **Memory & Parameters** | **Varies (T5-Small: 60M, T5-Base: 220M, T5-Large: 770M, T5-XL: 3B, T5-XXL: 11B)** |
| **Real-time Adaptation** | Can be fine-tuned for different NLP applications but not designed for chatbot-style memory retention |

***Gpt2 model***

* **GPT-2 (Generative Pre-trained Transformer 2)** is a **language model** developed by **OpenAI** in **2019**. It is the second version in the **GPT series** and was designed to generate human-like text.
* **Key Features of GPT-2:**

1. **Transformer-Based Architecture**
   1. Uses **decoder-only architecture** (like GPT-3 and GPT-4).
   2. Based on **self-attention** mechanisms for text generation.
2. **Autoregressive Model**
   1. Predicts the **next word** based on previous words.
   2. Generates text **sequentially**, improving coherence.

***BERT model***

* BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model developed by Google AI in 2018. It is one of the most influential large language models designed for natural language processing (NLP) tasks.
* Key Features of BERT:

1. Bidirectional Understanding – Unlike traditional models that read text left-to-right or right-to-left, BERT reads the entire sentence at once, understanding context better.
2. Pre-trained on Large Datasets – BERT is trained on massive amounts of text data, such as Wikipedia and BooksCorpus.
3. Fine-Tuning for Specific Tasks – After pre-training, BERT can be fine-tuned for tasks like:
   1. Text classification (e.g., spam detection)
   2. Question answering (e.g., Google Search)
   3. Named entity recognition (e.g., identifying names in text)

***Key Differences: T5 vs. BERT vs. GPT***

|  |  |  |  |
| --- | --- | --- | --- |
| Model | BERT | GPT (GPT-2, GPT-4) | T5 |
| Architecture | Encoder-only | Decoder-only | Encoder-Decoder |
| Main Focus | Understanding text | Generating text | Understanding & Generating text |
| Training Objective | Masked Language Modeling (MLM) | Causal Language Modeling (CLM) | Text-to-Text Training |