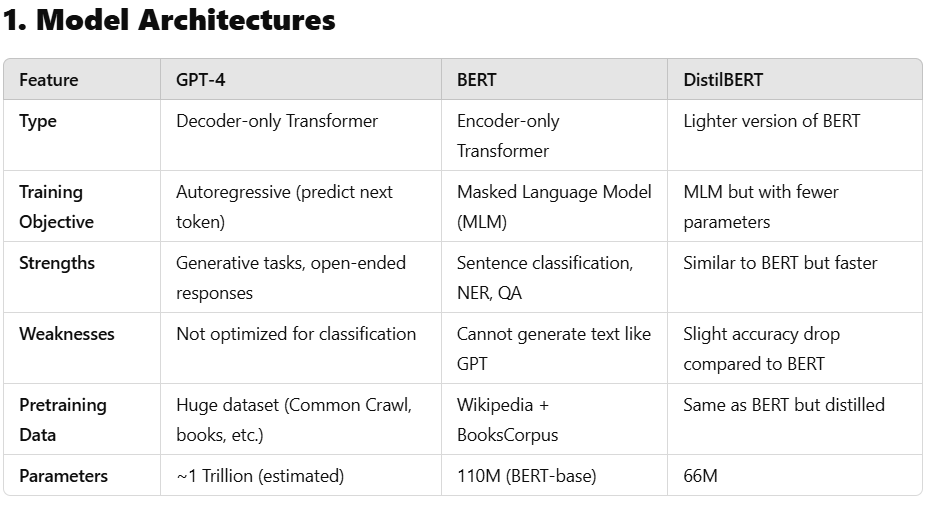
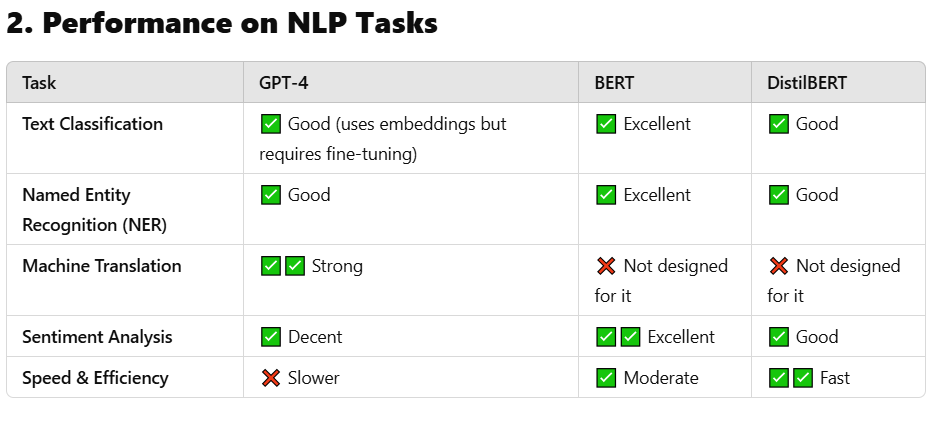
**Assiment2,رحمه نزيه عبد الغني   
20210328**

**Comparison of GPT-4, BERT, and DistilBERT**





**1. Understanding the Architecture**

BERT (Bidirectional Encoder Representations from Transformers) and GPT-4 (Generative Pre-trained Transformer 4) both belong to the transformer family, but they have fundamentally different architectures and training philosophies.

**BERT: The Bidirectional Encoder**

BERT is an **encoder-based** transformer model introduced by Google. Unlike traditional transformers that process text sequentially, BERT reads text **in both directions at the same time** (bidirectional attention). This allows it to **understand context better** because it looks at the words before and after the target word.

For example, in the sentence:  
*"The bank was located near the river."*  
BERT can determine that **"bank"** means a riverbank and not a financial institution because it looks at the surrounding words.

**GPT-4: The Decoder-based Model**

GPT-4, developed by OpenAI, is a **decoder-based** model, meaning it generates text **one word at a time** in a left-to-right manner. Unlike BERT, which focuses on understanding language, GPT-4 is designed to **generate human-like responses**.

For instance, when given the input:  
*"Tell me a story about a robot in space."*  
GPT-4 will **generate a full story**, making it more suitable for **creative and open-ended tasks** like chatbots, content generation, and summarization.

**T5: A Unified Encoder-Decoder Model**

T5 (Text-to-Text Transfer Transformer), developed by Google, combines aspects of both architectures. It **encodes and then decodes**, treating every NLP problem as a **text-to-text task**. Unlike BERT (which is just an encoder) and GPT-4 (which is just a decoder), T5 can both understand and generate text effectively.

### ****2. Training Objectives: What Are They Learning?****

Each model is trained with a different goal in mind:

* **BERT** uses a **Masked Language Model (MLM)**, where it randomly removes words from sentences and tries to predict them. It is also trained with **Next Sentence Prediction (NSP)** to understand sentence relationships.
* **GPT-4** is trained using **causal language modeling**, meaning it predicts the next word based on previous words only. This allows it to generate fluent and coherent text.
* **T5** is trained using a **text-to-text framework**, meaning every task—translation, summarization, classification—is converted into a text problem (e.g., "Summarize: [text]").

👉 **Key Difference**:  
BERT learns **deep context** for better understanding, GPT-4 learns to **generate text**, and T5 balances **both** understanding and generation.

**3. Performance on NLP Tasks**

Each model excels in different areas:

* **BERT** is strong in tasks that require deep understanding, such as **question answering, named entity recognition, and text classification**. It is commonly used in search engines and sentiment analysis.
* **GPT-4** is excellent for **text generation, chatbot responses, and long-form content creation** because it predicts the next word in a sequence.
* **T5** is **versatile**, capable of handling tasks like **translation, summarization, classification, and more** in a unified framework.

For example, in **sentiment analysis** on the IMDb dataset:

* **BERT** classifies sentences accurately due to deep contextual understanding.
* **GPT-4** can also classify sentiment but is better at explaining *why* a sentence is positive or negative.
* **T5** can **both** classify and generate a summary of the review.

**4. Computational Efficiency and Cost**

Each model has different computational requirements:

* **BERT** is efficient for classification tasks but requires fine-tuning on new datasets. It needs a **GPU** but is much faster than GPT-4.
* **GPT-4** is extremely large (with billions of parameters) and **expensive** to run. It requires **API calls** and is not as easy to fine-tune for specific tasks.
* **T5** is a middle ground—it's larger than BERT but smaller than GPT-4, making it **more flexible and efficient** for real-world applications.

👉 **Conclusion**: If you need **accuracy with efficiency, BERT is the best**. If you need **text generation and creativity, use GPT-4**. If you need a **multi-purpose model**, T5 is a great choice.