**Comparison Between GPT-4 and BERT: Architecture**

|  |  |  |
| --- | --- | --- |
| Feature | GPT-4 (Generative Pre-trained Transformer 4) | BERT (Bidirectional Encoder Representations from Transformers) |
| Model Type | Autoregressive Transformer (Decoder-only) | Autoencoding Transformer (Encoder-only) |
| Architecture | Uses only the decoder stack of the Transformer model. Generates text one token at a time (left-to-right). | Uses only the encoder stack of the Transformer model. Processes input bidirectionally. |
| Transformer Blocks | Stacked decoder layers with self-attention and cross-attention | Stacked encoder layers with self-attention |
| Self-Attention Mechanism | Causal self-attention (each token attends only to previous tokens) | Bidirectional self-attention (each token attends to all tokens in the input) |
| Training Objective | Causal Language Modeling (CLM): Predicts the next token given previous tokens. | Masked Language Modeling (MLM): Predicts missing words in a sentence. Next Sentence Prediction (NSP): Determines if one sentence follows another. |
| Bidirectionality | Unidirectional (Left-to-Right) | Bidirectional (Processes entire sentence at once) |
| Context Window | Large (e.g., GPT-4 can handle longer contexts) | Typically limited (e.g., BERT-base uses 512 tokens max) |
| Tokenization | Byte Pair Encoding (BPE) or SentencePiece | WordPiece Tokenization |
| Fine-tuning | Can be fine-tuned but mainly used as a general-purpose model | Pre-trained on large corpora, then fine-tuned for specific tasks |
| Use Cases | Text generation, chatbots, code generation, creative writing, summarization | Sentiment analysis, named entity recognition (NER), question answering, text classification |
| Scalability | Extremely large (trillions of parameters, optimized for massive-scale inference) | Typically smaller than GPT models (BERT-base: 110M parameters, BERT-large: 340M parameters) |
| Inference Efficiency | Slower due to autoregressive token-by-token generation | Faster since it processes the whole input at once |

**Detailed Explanation of Architecture:**

**GPT-4 Architecture:**

* Uses a **stack of transformer decoders**, each with multi-head self-attention, layer normalization, and feedforward layers.
* Implements **causal self-attention**, meaning each token can only attend to previous tokens, ensuring logical flow in generation.
* Large-scale model trained with reinforcement learning and fine-tuned for alignment with user intent.

**BERT Architecture:**

* Uses a **stack of transformer encoders**, which process input text bidirectionally.
* Employs **masked self-attention**, allowing each token to attend to all others in the input sequence.
* Designed primarily for understanding language rather than generating it.