Architectural Comparison: GPT-4 vs. BERT

1. Transformer Stack

GPT-4: Decoder-only Transformer

BERT: Encoder-only Transformer

- GPT-4 uses only the decoder blocks of the original Transformer architecture.
- BERT uses only the encoder blocks of the Transformer.

2. Attention Mechanism

GPT-4: Causal self-attention (left-to-right only)

BERT: Masked self-attention (bidirectional)

- GPT-4 masks future tokens to enable autoregressive generation.
- BERT masks random tokens to learn context from both directions.

3. Positional Encoding and Inputs

GPT-4: Likely uses learned or rotary positional encodings; processes single input sequence.

BERT: Uses absolute positional encodings; supports segment embeddings for sentence pairs.

4. Output Behavior (Pretraining Phase)

GPT-4: Predicts the next token (autoregressive); outputs one token at a time.

BERT: Predicts masked tokens in the sequence; multiple predictions per input.

5. Layer Structure and Flow

- Both use multi-head self-attention, feedforward layers, residual connections, and layer normalization.
- Main difference lies in attention masking and input directionality.