Update Or Create Vector Store:

Vectors Update

Chat with PDF using AWS Bedrock's LLaMA 3 and Claude Models

Ask a Question from the PDF Files

what is attention

Claude Output

Based on the context provided, attention refers to an attention mechanism that allows a model to focus on relevant parts of the input when generating each output. The Transformer model described relies entirely on attention mechanisms instead of recurrence or convolutions. The attention mechanisms allow the model to draw global dependencies between input and output without regard to their distance. There are three types of attention used in the Transformer:

- Encoder-decoder attention where the queries come from the decoder and the keys and values come from the encoder. This allows the decoder to attend to all positions in the encoder input.
- Self-attention in the encoder where the keys, values and queries all come from the encoder output. This allows each position in the encoder to attend to all positions in the previous encoder layer.
- Self-attention in the decoder where positions can attend to all previous positions in the decoder.

The multi-head attention splits the attention into multiple "heads" to attend to different representation subspaces which helps the model jointly attend to information from different positions.

Done

Llama3 Output