Title: Transformers and Attention

Transformers use self-attention mechanisms to weigh token interactions.
Scaled dot-product attention computes weights via queries and keys.
Positional encodings inject order information into token embeddings.