Paper: Attention Is All You Need (sample notes)
The paper introduces the Transformer architecture based on self-attention.
It replaces recurrence with multi-head attention and position-wise feed-forward blocks.
Key benefits include parallel training, strong sequence modeling, and improved translation quality.
This sample file is included for quick local RAG testing.