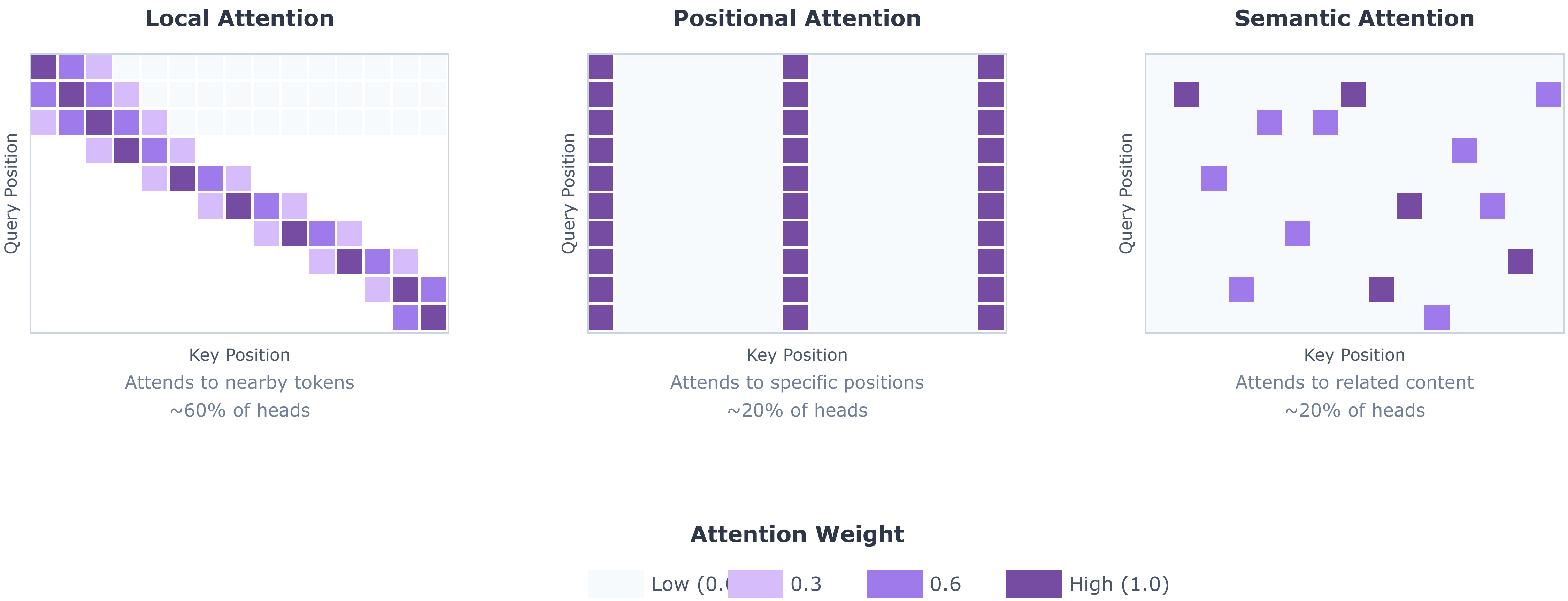
Learned Attention Patterns



Local Attention: Attends to nearby tokens
~60% of heads

Positional Attention: Attends to specific positions
~20% of heads

Semantic Attention: Attends to related content
~20% of heads

Optimization Opportunity

Local patterns enable sparse attention: skip low-weight computations
50-80% computation reduction with <1% accuracy impact