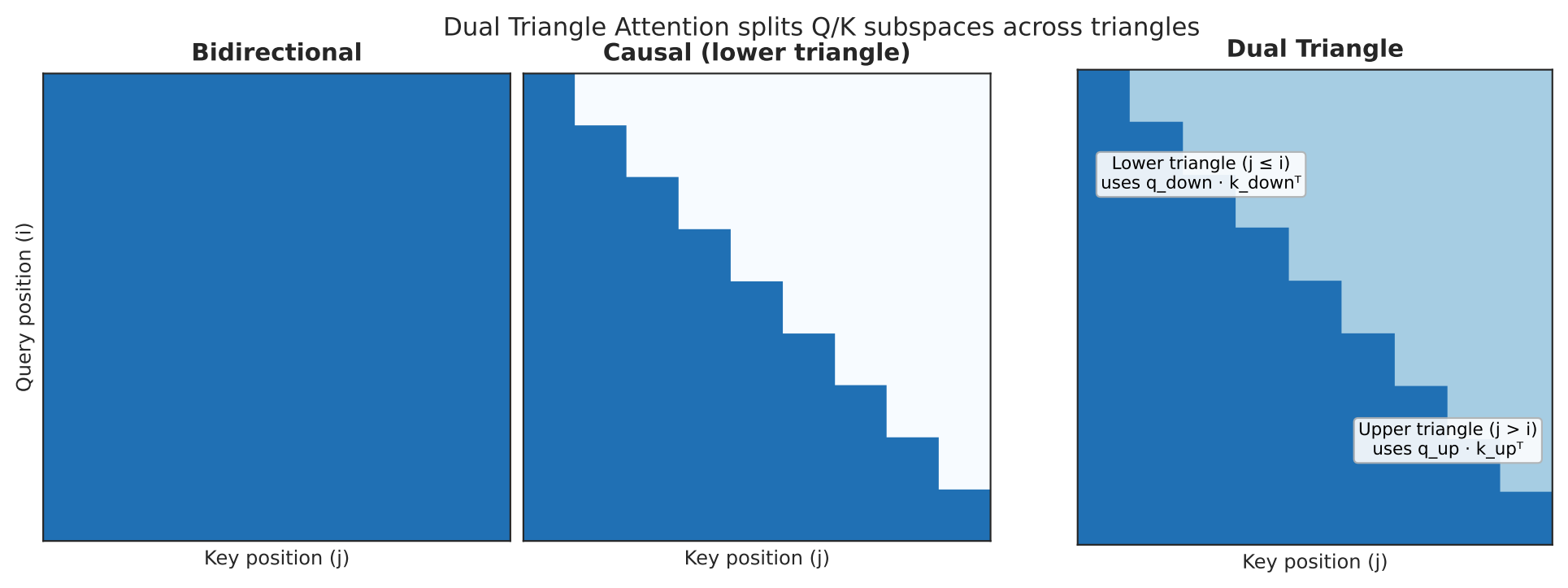
Dual Triangle Attention splits Q/K subspaces across triangles
Bidirectional
Causal (lower triangle)
Dual Triangle
Query position (i)
Key position (j)
Key position (j)
Key position (j)
Lower triangle (j ≤ i)
uses q_down · k_downᵀ
Upper triangle (j > i)
uses q_up · k_upᵀ