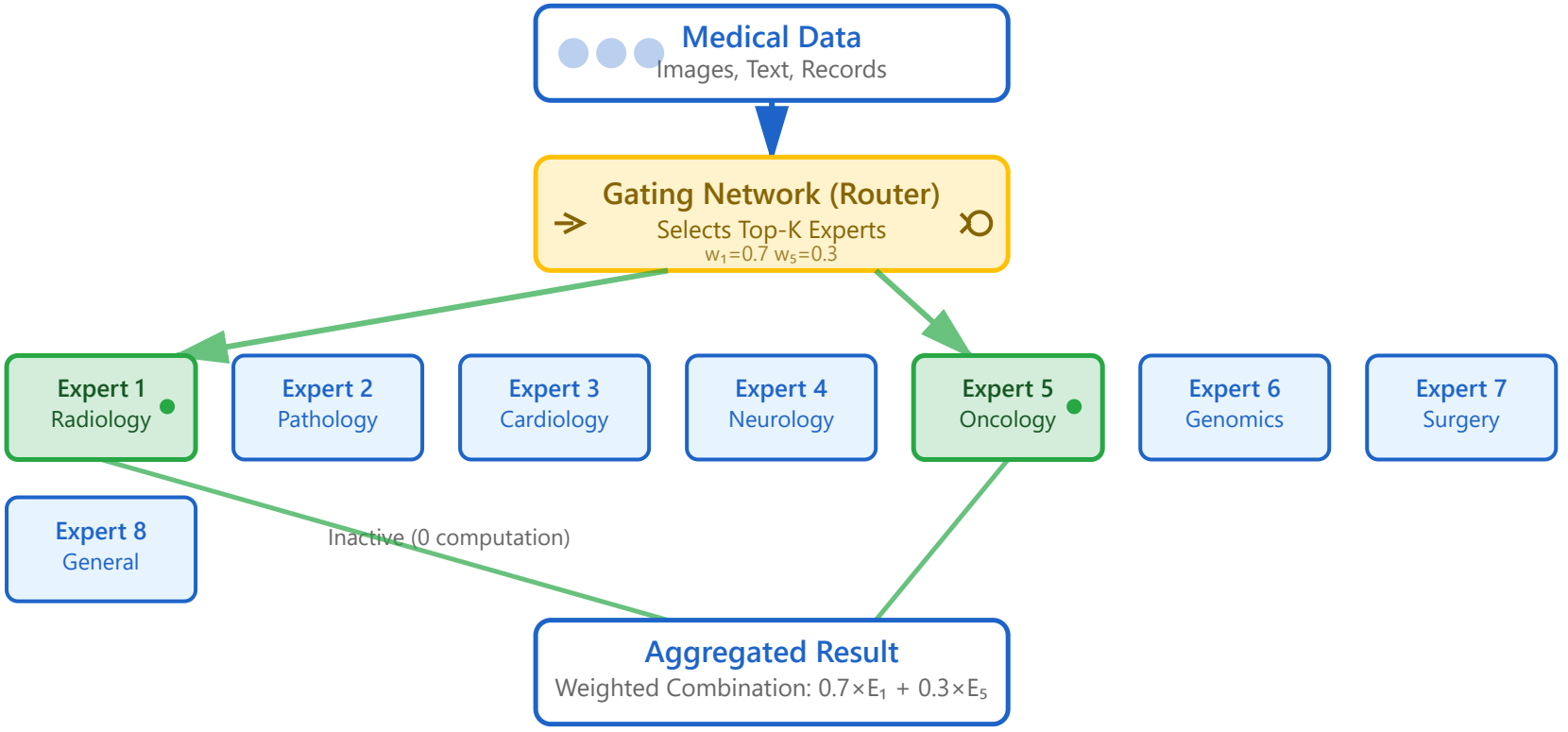
Mixture of Experts (MoE) Architecture & Gating Mechanism



🎯 Key Concept: Sparse Activation

Only 2-4 experts are activated per input, enabling massive scale with manageable computation. Gating network learns to route inputs to the most relevant domain specialists. **Green highlights show active experts** processing this input.