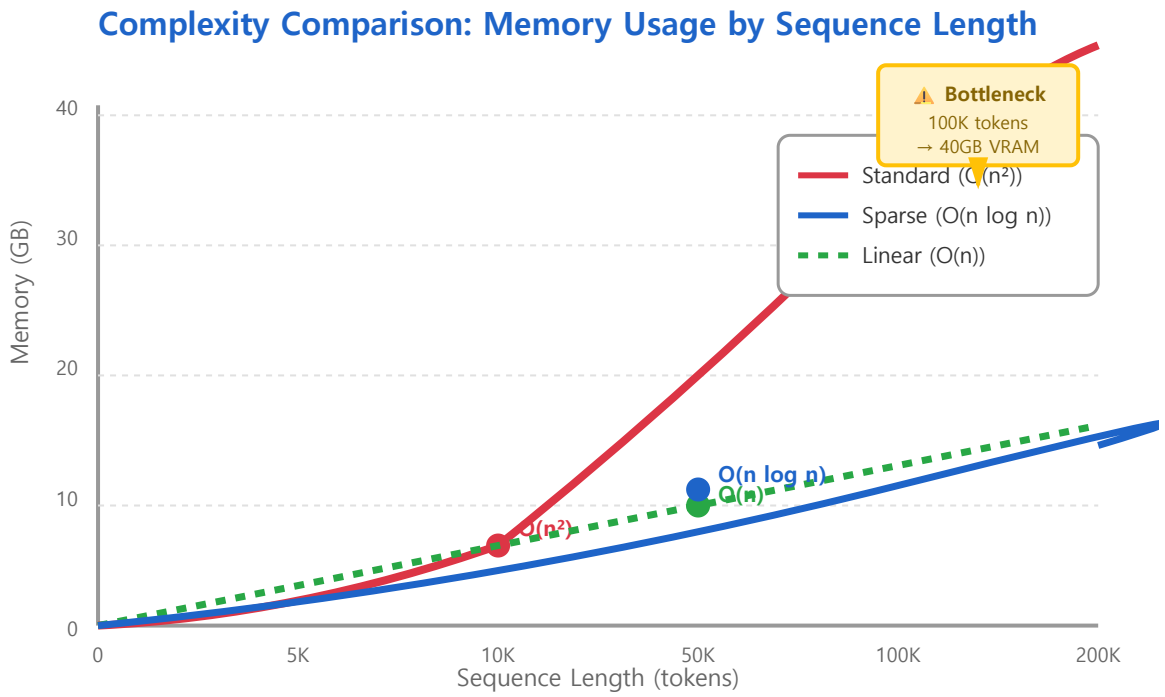
Efficient Attention Mechanisms



Standard Attention

- $O(n^2)$ complexity
- Memory bottleneck at 10K+
- Limits deployment

Sparse Attention

- Fixed patterns (stride, block)
- Learned sparsity
- Reduced memory footprint

Linear Attention

- Kernel approximation
- O(n) complexity
- Small accuracy trade-off

Hierarchical Attention

- Multi-scale processing
- Local + global attention
- Efficient for long sequences

Efficiency Gains

Efficient attention mechanisms reduce memory from $O(n^2)$ to $O(n \log n)$ or $O(n)$, enabling **10-100x longer contexts on same hardware**. Critical for processing entire patient histories.