Memory Efficiency Comparison Across Attention Techniques Standard Flash Attention Fused Softmax Multi-Query Sliding Window 10^1 Memory Usage (GB) 10^{0} 512 1024 2048 4096 **Sequence Length**