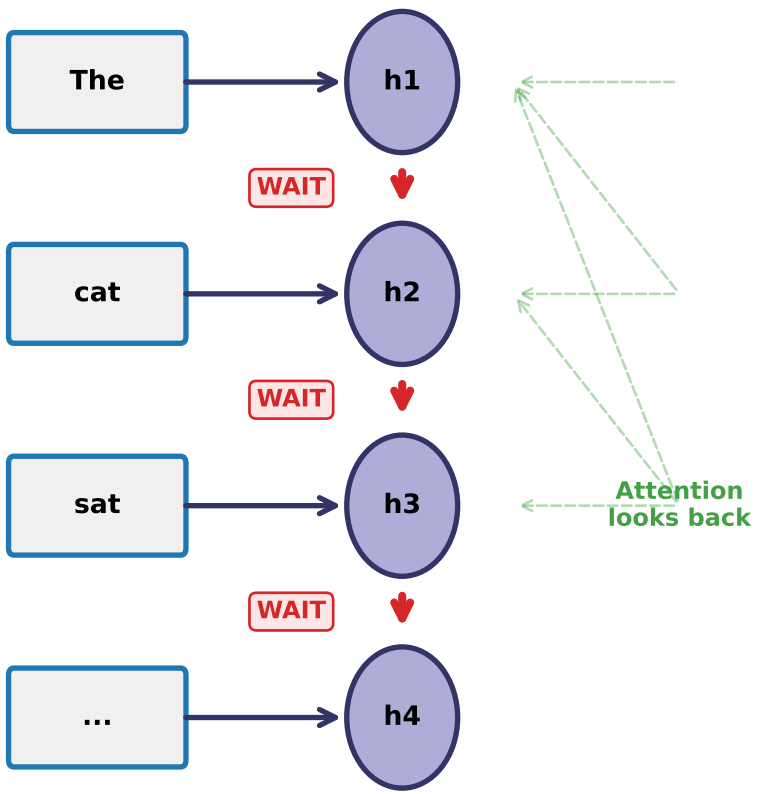
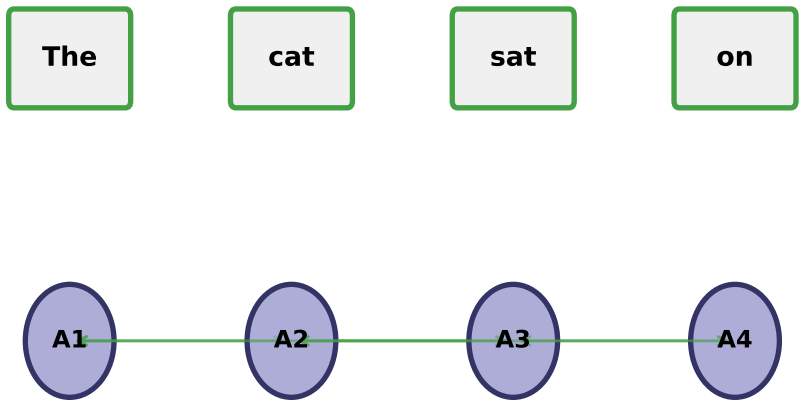
OLD WAY: RNN + Attention



⚠ Sequential Bottleneck

Process one word at a time
GPU utilization: 1-5%

NEW WAY: Pure Attention



✓ Full Parallelization

Process ALL words simultaneously

GPU utilization: 85-92%

No waiting! → 100x speedup