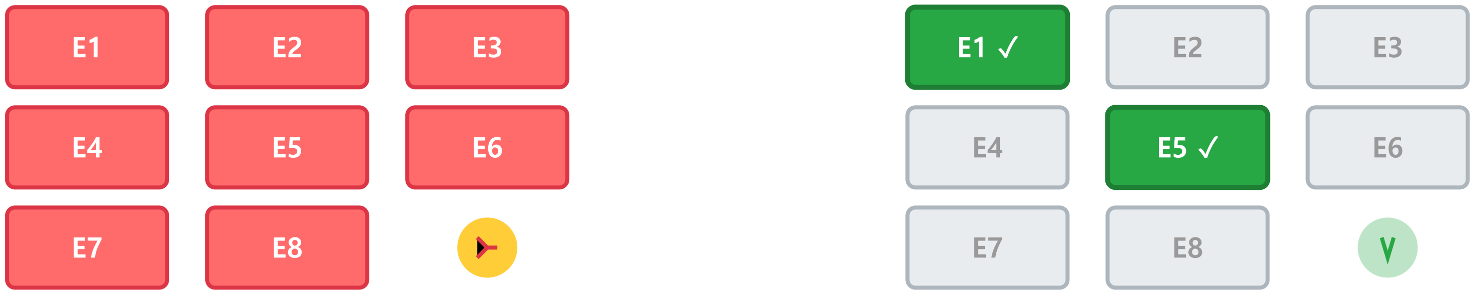
Sparse Activation: Computational Efficiency through Selective Expert Activation



All 8 Experts Active (100%)

High computation & memory usage

Only 2 Experts Active (25%)

75% computation reduction!

Active Experts	**8 / 8 (100%)**
Computation	**Full (100%)**
Memory Usage	**High**
Latency	**High**

Active Experts	**2 / 8 (25%)**
Computation	**75% Reduction**
Memory Usage	**Low**
Latency	**Low**

⚡ Efficiency Gains

Sparse activation enables models with billions of parameters to run with the computational cost of much smaller models. A 64-expert MoE with Top-2 routing achieves **32x parameter scaling with only 2x computation increase!**