# The Architecture of Efficiency
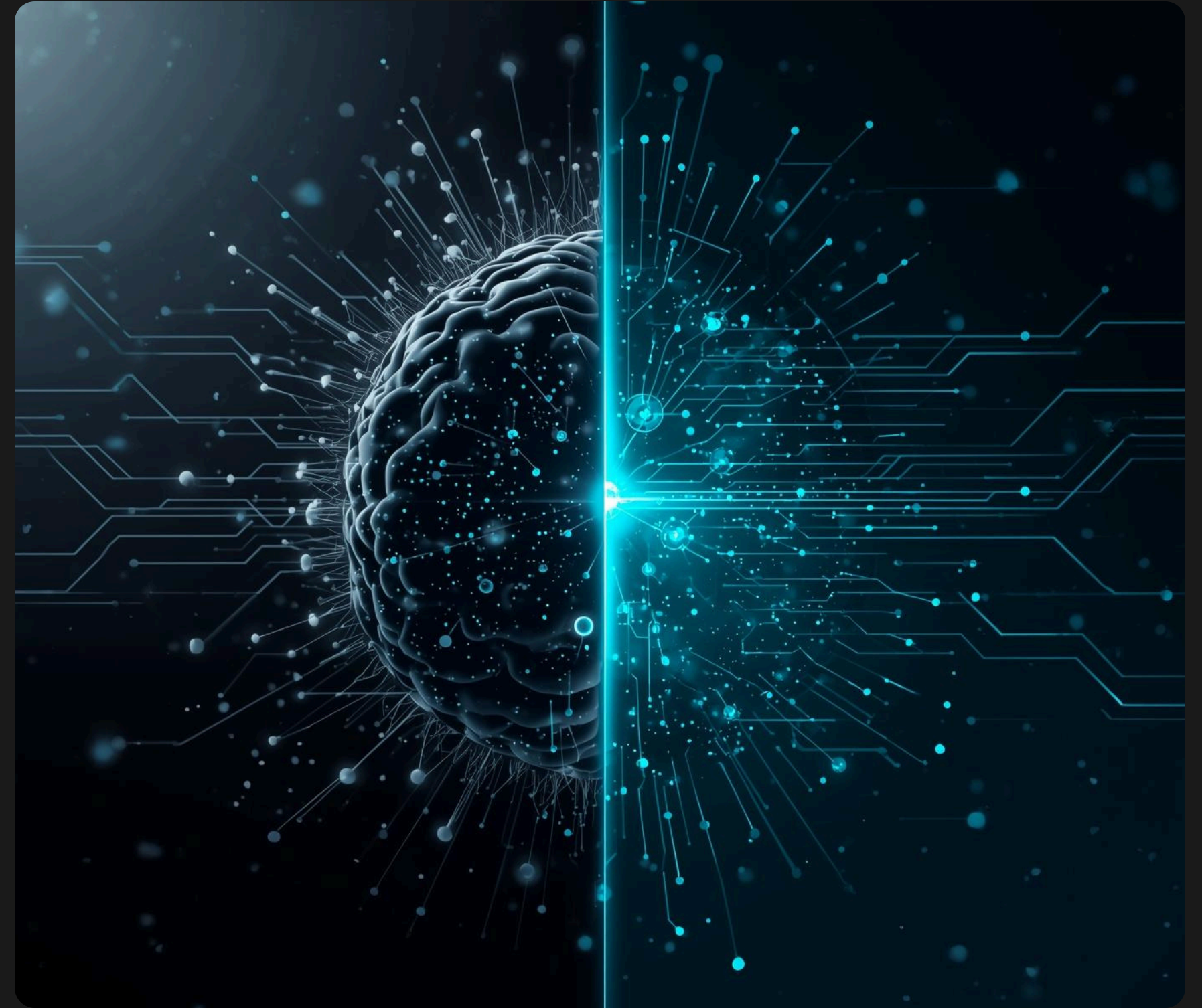
Scaling Intelligence while Reducing Costs

Mistral
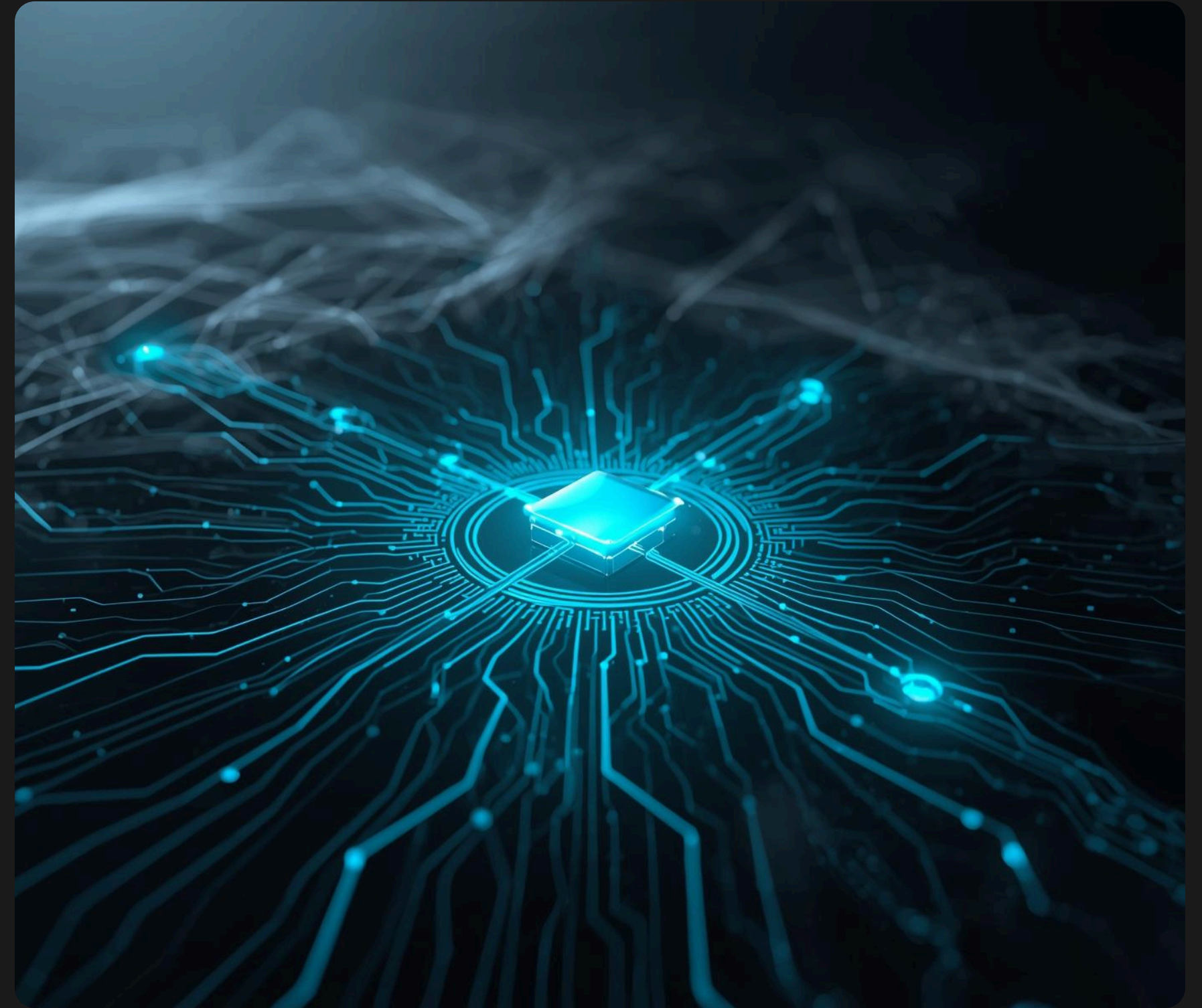
# Dense vs. Sparse Processing Efficiency



**HIGH EFFICIENCY THROUGH SELECTIVE ACTIVATION**

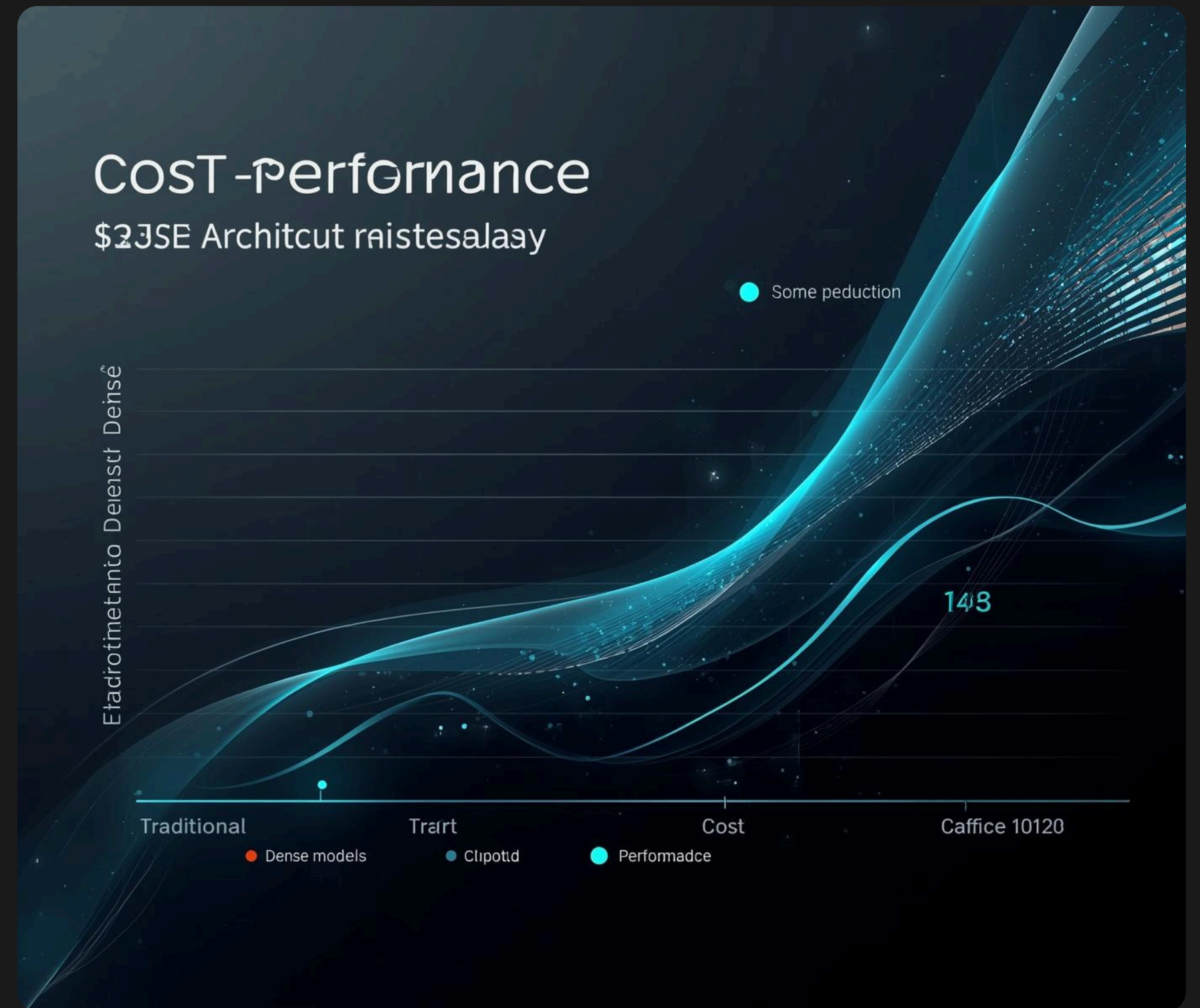Sparse models activate only relevant experts.

# Intelligent Task Delegation Mechanism



**DYNAMIC ROUTING FOR ENHANCED EFFICIENCY**

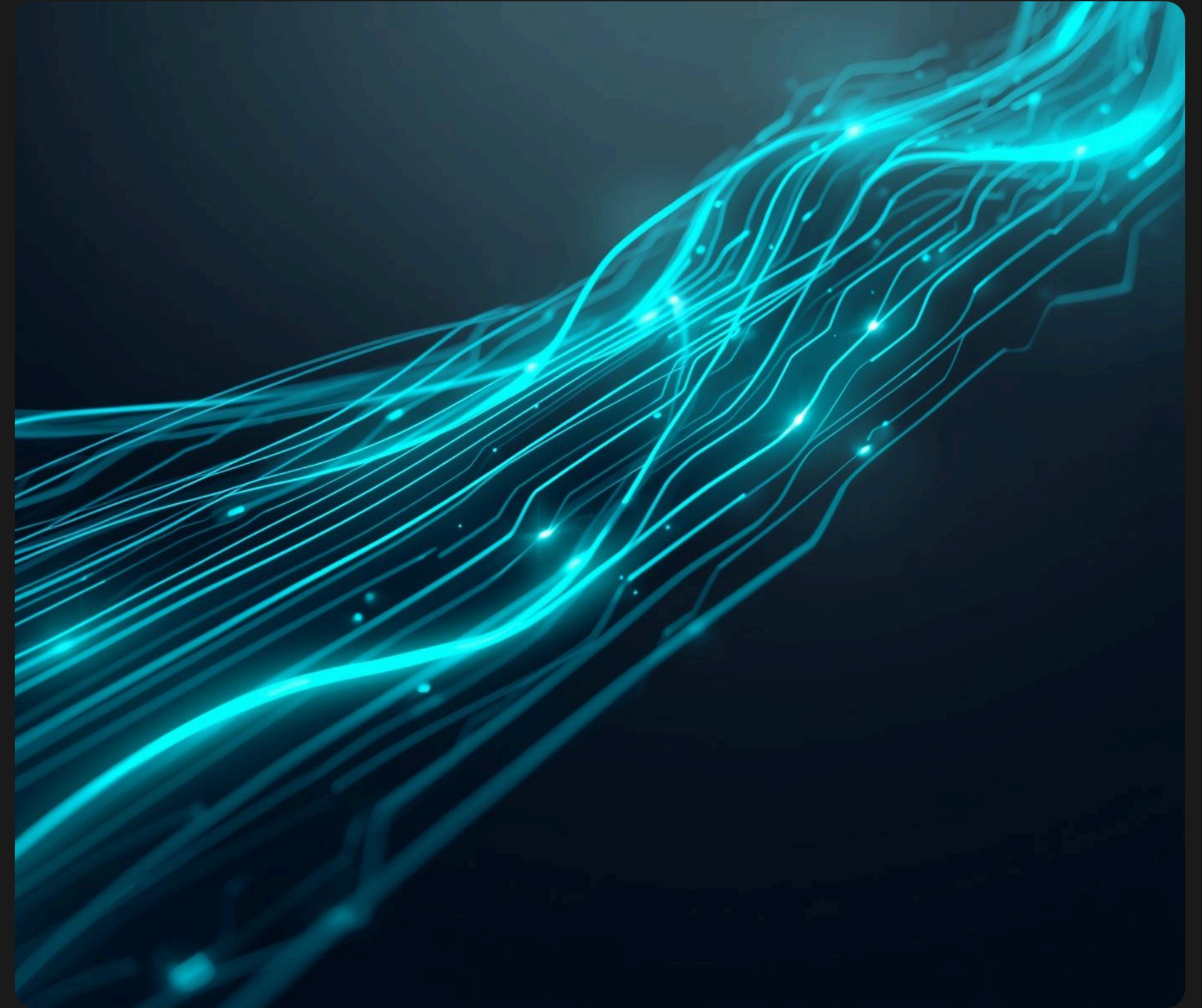Expert selection boosts processing effectiveness significantly.

# The Economic Advantage of MoE Architecture



## ACHIEVING SOTA PERFORMANCE WITH LOWER COSTS

MoE enables **8x cost savings** for enterprises.

# Doing More
with Less VRAM



**OPTIMIZED HARDWARE FOR ENHANCED PERFORMANCE**

Reduced latency improves user experience significantly.

# Tactical MoE Deployment: A New Standard



**FRONTIER INTELLIGENCE MEETS ENTERPRISE BUDGETS**

Transitioning AI from luxury to infrastructure