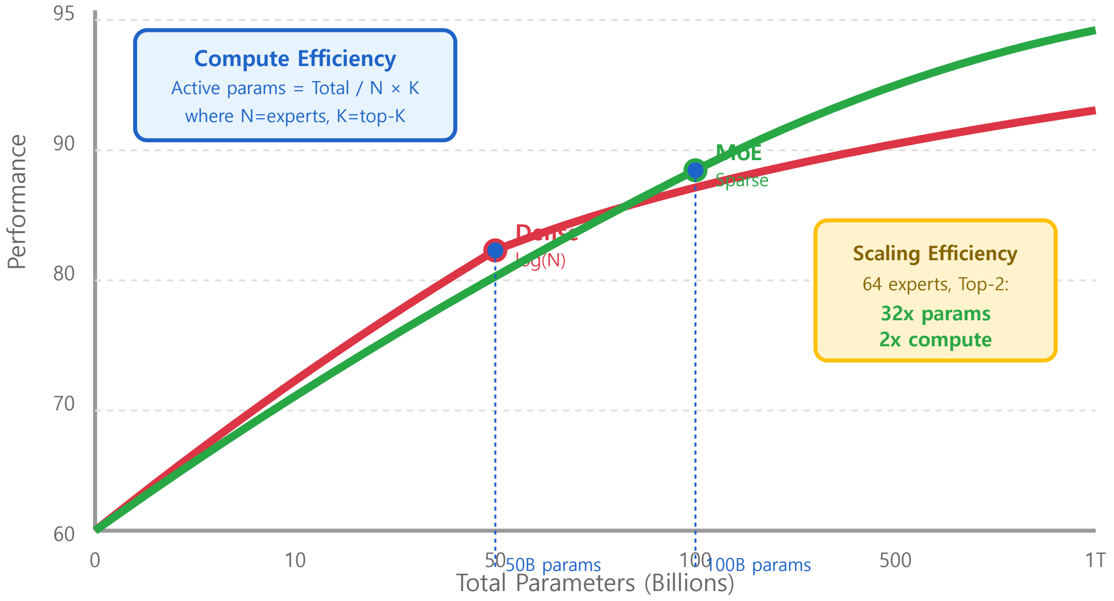
Scaling Laws for Mixture of Experts

Performance vs Model Size: MoE Scaling Advantage



Compute Efficiency

Active params = Total / N × K

where N=experts, K=top-K

Scaling Efficiency

64 experts, Top-2:

32x params

2x compute

Parameter Scaling

- Performance ∝ log(Parameters)
- Sub-linear gains beyond size
- Efficient with sparse activation

Compute Scaling

- Active = Total / N × K
- 64 experts, Top-2:

32x params, 2x compute!

Data Requirements

- More params need diverse data
- Medical: 10-100M samples
- Multi-modal beneficial

Inference Cost

- Latency: active experts only
- Memory: all weights loaded
- Bandwidth: critical factor

📊 Scaling Predictions

MoE models can scale to trillions of parameters while maintaining practical inference costs. **Medical MoE with 128 experts and 10B params performs like 200B dense model** with fraction of the compute.