Performance Comparisons & Benchmarks

MoE vs Dense

- 10x parameters, 2x compute
- Better specialization
- Complex deployment

Long-Context Methods

- Flash Attention: 5-9x faster
- Mamba: O(n) vs $O(n^2)$
- RAG: cost-effective scaling

Novel Architectures

- Graph Transformers: +30% on relational tasks
- Neural ODEs: superior for time-series
- Quantum: TBD (future)

Medical Benchmarks

- MedQA accuracy: 85-92%
- Imaging F1: 0.90-0.96
- Clinical NER: 0.88-0.94

📊 Comparative Analysis

Different architectures excel at different tasks. MoE for scale, long-context for history, graphs for relationships. Hybrid approaches often yield best results.