Extended Context Windows: 100K+ Tokens

Traditional Limits

- BERT: 512 tokens
- GPT-3: 2,048 tokens
- Limited patient history view

Modern Long Context

- GPT-4 Turbo: 128K tokens
- Claude 3: 200K tokens
- Gemini 1.5: 1M tokens

Memory Requirements

- Attention memory: $O(n^2)$
- 100K context: ~40GB VRAM
- Optimization crucial for deployment

Medical Applications

- Full patient history analysis
- Multi-visit trend detection
- Comprehensive literature review

📈 Context Scaling

Extended context windows enable processing entire patient histories spanning years of medical records, lab results, and imaging reports in a single inference.