LLaMA Family Context Windows

Meta's Research-Driven Open Models

LLaMA 2 (7B, 13B, 70B)

- **Context Window:** 4,096 tokens (~3,000 words)
- **License:** Custom license (commercial use allowed)
- **Strengths:** Strong foundational capabilities, efficient training
- **Best For:** Research, fine-tuning experiments, educational use
- **Community:** Large open-source ecosystem
- **Deployment:** Requires local hosting or cloud deployment

Code Llama (7B, 13B, 34B)

- **Context Window:** 16,384 tokens (~12,000 words)
- **Specialization:** Code generation, completion, and debugging
- **Languages:** Python, C++, Java, PHP, TypeScript, C#, Bash
- **Best For:** Software development, code analysis, programming education
- **Training:** Specialized on 500B tokens of code
- **Performance:** Competitive with GitHub Copilot

🔧 Research Innovations

- **RoPE (Rotary Position Embedding):** Enables context extension
- **YaRN:** Yet another RoPE extensioN technique
- **LongLLaMA:** Community extensions to 256K+ tokens
- **Code-specific optimizations:** Tailored for programming contexts

🔍 Fine-tuning Capabilities

- **LoRA (Low-Rank Adaptation):** Efficient parameter updates
- **QLoRA:** Quantized fine-tuning for consumer hardware
- **Instruction tuning:** Alpaca, Vicuna, and other derivatives
- **Domain specialization:** Medical, legal, scientific variants

4K Context (LLaMA 2) Applications

- Research experiments & model comparison
- Fine-tuning base models for specific tasks
- Educational projects and learning
- Efficient inference for simple tasks

16K Context (Code Llama) Applications

- Complete function and class analysis
- Multi-file code understanding & generation
- Documentation and docstring creation
- Code refactoring and optimization

🔬 Research & Community Focus

LLaMA models serve as a foundation for hundreds of research projects, fine-tuned variants, and community experiments, encouraging efficient prompt design and specialized fine-tuning approaches.