LLM Context Windows: Complete Comparison

Choosing the Right Model for Your Use Case

MODEL FAMILY	MODEL NAME	CONTEXT WINDOW	~WORD COUNT	COST TIER	BEST USE CASES
GPT (OpenAI)	GPT-3.5-Turbo	4K tokens	~3,000 words	💰 Low	Quick queries, chatbots
	GPT-3.5-16K	16K tokens	~12,000 words	💰💰 Medium	Document analysis
	GPT-4	8K tokens	~6,000 words	💰💰💰 High	Complex reasoning
	GPT-4-32K	32K tokens	~24,000 words	💰💰💰💰 Very High	Large documents
	GPT-4-Turbo-128K	128K tokens	~96,000 words	💰💰💰 High+	Massive analysis
Claude (Anthropic)	Claude 3 Haiku	200K tokens	~150,000 words	💰💰 Medium	Fast long-context
	Claude 3 Sonnet	200K tokens	~150,000 words	💰💰💰 High	Balanced performance
	Claude 3 Opus	200K tokens	~150,000 words	💰💰💰💰 Very High	Premium analysis
Mistral	Mistral 7B	32K tokens	~24,000 words	💰 Low (Open)	Research, fine-tuning
	Mistral Large	32K tokens	~24,000 words	💰💰💰 High	Enterprise apps
LLaMA (Meta)	LLaMA 2	4K tokens	~3,000 words	Free (Self-host)	Research, education
	Code Llama	16K tokens	~12,000 words	Free (Self-host)	Code development

Key Insights

- 🏆 Context Champions: Claude models lead with consistent 200K tokens across all variants.
- 💰 Cost Efficiency: Budget: GPT-3.5, Mistral 7B. Premium: Claude Opus, GPT-4-32K. Value: Claude Haiku, GPT-4-Turbo.
- 🎯 Use Cases: Short (≤4K): GPT-3.5, LLaMA 2. Medium (16-32K): GPT-4, Code Llama, Mistral. Long (100K+): Claude family, GPT-4-Turbo.

🧪 Quick Selection Guide

1. **Budget priority?** → GPT-3.5 or Mistral 7B
2. **Long documents?** → Claude Haiku/Sonnet
3. **Highest quality?** → Claude Opus or GPT-4-32K
4. **Code focus?** → Code Llama or Claude
5. **Research/Learning?** → LLaMA 2 (free)