GPT Family Context Windows

OpenAI's Approach to Context Management

GPT-3.5-Turbo (Standard)

Context Window: 4,096 tokens (~3,000 words)

Best For: Quick queries, simple tasks, cost-effective solutions

Use Cases: Chatbots, simple content generation, basic Q&A

Cost: Most economical option

GPT-3.5-Turbo-16k

Context Window: 16,384 tokens (~12,000 words)

Best For: Medium documents, code analysis, longer conversations

Use Cases: Document summarization, code review, extended dialogue

Cost: 2x standard GPT-3.5 pricing

GPT-4 (Standard)

Context Window: 8,192 tokens (~6,000 words)

Best For: Complex reasoning, high-quality output, creative tasks

Use Cases: Analysis, creative writing, complex problem-solving

Cost: Premium pricing for superior capability

GPT-4-32k

Context Window: 32,768 tokens (~24,000 words)

Best For: Large documents, extensive code bases, comprehensive analysis

Use Cases: Book summarization, large codebase analysis

Cost: Highest tier pricing

GPT-4-Turbo-128k

Context Window: 128,000 tokens (~96,000 words)

Best For: Massive documents, entire codebases, comprehensive research

Use Cases: Academic papers, full application analysis

Cost: Premium with volume discounts

Practical Examples

- **4K Context:** Email responses, short articles, basic conversations.
- **16K Context:** Research papers, code files, detailed analysis.
- **32K Context:** Technical manuals, large datasets, extensive documentation.
- **128K Context:** Entire books, complete applications, comprehensive reviews.

🎯 Strategic Insights

Strategy: Start with smaller context models for prototyping, then scale up based on actual needs. GPT-3.5-16k often provides the best price/performance ratio for most applications.