Model Pricing

Model family	Best for	Context window	Standard Pricing	
Claude Instant	Low latency, high throughput use cases	100,000 tokens	Prompt $0.80/million tokens	Completion $2.40/million tokens
Claude 2.0	Superior performance on tasks that require complex reasoning	100,000 tokens	Prompt $8.00/million tokens	Completion $24.00/million tokens
Claude 2.1	Same performance as Claude 2, plus significant reduction in model hallucination rates	200,000 tokens	Prompt $8.00/million tokens	Completion $24.00/million tokens