| Model | Provider | Cost/1K Tokens | Latency | Max Context | Strength | Use Cases |

|---------------|--------------|----------------|---------|-------------|-------------------------------------|-------------------------------------|  
| GPT-4o-mini | OpenAI | $0.01 (input) / $0.03 (output) | Low | 128K | Fast + cheap | Chat, summarization |  
| Claude 3 Sonnet | AWS Bedrock | ~$3.00 | Med | 200K | High reasoning, long inputs | Legal, reasoning, Q&A |  
| Mistral 7B | AWS Bedrock | ~$0.20 | Very Low| 32K | Cost-effective for short tasks | Summarization, classification |  
| Amazon Nova | AWS Bedrock | ~$0.10 | Very Low| 16K | Cheapest option | Lightweight apps, internal tools |