| Model | Cost (Input / Output) | Avg. Tokens per Request | Est. Cost per 100 Requests |

|------------------|---------------------------|--------------------------|----------------------------|  
| GPT-4o-mini | $0.01 / $0.03 per 1K tokens | ~800 tokens | ~$2.40 |  
| Claude 3 Sonnet | ~$3 per 1K tokens | ~1000 tokens | ~$300 |  
| Mistral 7B | ~$0.20 per 1K tokens | ~600 tokens | ~$12 |  
| Amazon Nova | ~$0.10 per 1K tokens | ~500 tokens | ~$5 |

Total Cost = (Input Tokens + Output Tokens) / 1000 \* Price per 1K

Example: 800 input + 400 output = 1200 tokens → 1.2 \* $0.01 = $0.012