RAG Performance Optimization Guide

Optimization Area	Technique	Expected Impact
Chunking Strategy	Semantic boundaries	15-20% accuracy gain
Embedding Model	text-embedding-3-large	10% better retrieval
Index Configuration	Hybrid search enabled	25% faster queries
Caching Layer	Redis for embeddings	60% latency reduction
Load Balancing	Auto-scaling groups	3x throughput capacity

Implementation Guidelines

The optimization strategies outlined above have been tested in production environments and demonstrate consistent performance improvements. Implementation should be done incrementally, with careful monitoring of system metrics at each stage.

Monitoring and Metrics

Key performance indicators to track during optimization:

- Query response time (target: <2 seconds)
- Retrieval accuracy (target: >85%)
- System throughput (target: >50 queries/second)
- Resource utilization (CPU, memory, storage)
- Cost per query and operational expenses