RAG Performance Optimization Strategies

Strategy	Implementation	Performance Gain
Chunking Optimization	Semantic boundaries	18% accuracy improvement
Embedding Caching	Redis-based cache	65% latency reduction
Hybrid Search	Semantic + keyword	23% better relevance
Result Reranking	Cross-encoder model	12% accuracy boost
Context Compression	Selective inclusion	30% token reduction