NebulaByte: Retrieval-Augmented Generation

NebulaByte RAG Systems Documentation

Understanding RAG
Retrieval-Augmented Generation (RAG) is a technique that enhances LLMs by providing them with relevant external knowledge. This approach combines the benefits of:
- Large language models' natural language understanding
- External knowledge bases for factual accuracy
- Real-time information retrieval

RAG Architecture:
1. Document Ingestion: Convert documents to text
2. Chunking: Break text into manageable pieces
3. Embedding: Convert chunks to vector representations
4. Indexing: Store embeddings in a vector database
5. Retrieval: Find relevant chunks for queries
6. Generation: Use LLM to generate answers

Benefits:
- Improved factual accuracy
- Reduced hallucinations
- Domain-specific knowledge
- Up-to-date information
- Cost-effective compared to fine-tuning

Implementation Strategies:
Use FAISS or Chroma for vector storage, sentence-transformers for embeddings, and GPT or Claude for generation. Consider chunk size, overlap, and retrieval strategies.