Implementing RAG Systems

Step-by-Step Implementation:

1. Document Processing:
   - Load PDFs, HTML, or text files
   - Split content into 512-token chunks with overlap
2. Embedding Generation:
   - Use models like BGE or MiniLM to create vectors
3. Vector Storage:
   - Use FAISS or Chroma for efficient similarity search
4. Query Processing:
   - Convert user query to embedding
   - Retrieve top 3 most relevant chunks
   - Pass context to LLM for response generation

Optimization Tips:

- Use hybrid search (keyword + semantic)
- Implement result re-ranking
- Add metadata filtering