NebulaByte AI Consultation - RAG Implementation Details

Date: October 6, 2025

Participants: Dr. Sarah Chen (Lead AI Architect), James Wilson (ML Engineer), Lisa Wong (Product Manager)

Retrieval-Augmented Generation (RAG) Implementation Plan:

The PDF RAG agent will be responsible for processing uploaded documents and enabling semantic search capabilities. The implementation will use the following components:

1. Text Extraction: - PyMuPDF (fitz) for PDF text extraction - Support for various PDF formats and layouts

2. Text Chunking: - Chunk size: 500 tokens with 50 token overlap - Semantic chunking to preserve context

3. Embedding Model: - SentenceTransformer 'all-MiniLM-L6-v2' for efficient encoding - 384-dimensional embeddings

4. Vector Store: - FAISS for fast similarity search - Index persistence for session continuity

5. Retrieval: - Top-K nearest neighbor search - Re-ranking based on relevance scores

Performance Targets: - Document processing: < 5 seconds for 10-page PDF - Search response: < 1 second - Accuracy: > 85% relevant results