NYD RAG Pipeline
Progress Report

Technical Implementation Details

Data Processing
- Sentence Transformer: all-MiniLM-L6-v2
- Vector Dimension: 384
- Database: PostgreSQL with pgvector extension
- Storage Format: Optimized for quick retrieval

Answer Generation
- Model: Llama3-8b-8192 via Groq API
- Context Window: Up to 3 most relevant verses
- Response Length: Maximum 500 tokens

Next Steps
1. Enhance retrieval accuracy through:
   - Fine-tuning embedding model
   - Implementing cross-encoder reranking
2. Improve answer generation:
   - Context window optimization
   - Source attribution enhancement
3. System optimization:
   - Caching frequently accessed verses
   - Batch processing for multiple queries