RAG System Technical Specification

Version: 1.0
Document ID: TECH-SPEC-001

1. OVERVIEW

This document describes the Retrieval-Augmented Generation (RAG)
system architecture and implementation details.

2. COMPONENTS

2.1 Vector Database
- PostgreSQL with pgvector extension
- Supports 768-dimensional embeddings
- Cosine similarity search

2.2 Storage Layer
- Google Cloud Storage for documents
- UUID-based file organization
- Support for PDF and TXT formats

2.3 Embedding Model
- Google's text-embedding-005
- 768 dimensions
- Multilingual support

3. API ENDPOINTS

POST /v1/documents/upload - Upload new document
GET /v1/documents - List all documents
DELETE /v1/documents/{id} - Remove document
POST /v1/query - Search similar content

4. TESTING

Comprehensive E2E tests validate:
- Document upload and deduplication
- Vector search accuracy
- File download with UTF-8 support
- Complete cleanup verification