DocuChat: Scope and Success Criteria

1. Project Overview
DocuChat is an end-to-end document Q&A microservice that allows users to ingest documents (PDF, Markdown, plain text), index them with embeddings, and ask natural-language questions, retrieving precise answers via a large language model.

2. Scope

- **In-Scope**
  - Document ingestion pipeline for PDF, Markdown, and TXT formats.
  - Text chunking (≈500 tokens) and normalization.
  - Vector embedding using a pretrained model (e.g., OpenAI or Sentence-Transformers).
  - Vector storage and retrieval service (Pinecone, Chroma, or Weaviate).
  - LLM integration for question answering (OpenAI GPT-4 or open-source alternative).
  - REST API endpoints (`/ingest`, `/query`) with authentication (API key/JWT).
  - Basic web-based chat UI (React) for querying.
  - Containerization (Docker) and CI pipeline (GitHub Actions).
- **Out-of-Scope (for MVP)**
  - Fine-tuning LLMs on custom data.
  - Multi-language support beyond English (optional extension).
  - Advanced analytics (usage dashboards, unanswered question tracking).
  - Slack/Teams bot integrations.

3. Success Criteria

- **Functional Metrics**
  - **Ingestion**: Able to process and index a batch of ≥100 pages (PDF or MD) in ≤2 minutes.
  - **Retrieval**: Return top-3 most relevant chunks in ≤300 ms (average).
  - **Answer Quality**: For a curated set of 20 sample questions, accuracy ≥80% (via manual evaluation).
- **Performance & Reliability**
  - **Uptime**: ≥99% service availability over 1 week of testing.
  - **Latency**: 95th percentile end-to-end (query to LLM response) ≤2 seconds.
- **Security & Maintainability**
  - Authentication enforced on all API endpoints.
  - Automated tests covering ≥70% code coverage for backend services.
- **Deliverables**
  - Hosted demo environment accessible via public URL.
  - Comprehensive README with architecture diagram and setup instructions.
  - Demo video (3–5 min) walking through ingestion, query, and deployment.

4. Constraints & Assumptions

- **Budget**: Use free tiers or open-source tools wherever possible.
- **Team**: Single developer (you) dedicating ~8–10 hours/week.
- **Timeline**: Phase 1 (Week 1) completes by May 1, 2025.
- **Assumptions**: Stable network connection for cloud APIs; sample documents available for testing.

Next Step: Review scope items, adjust as needed, and finalize success metrics before moving to ingestion-phase setup.