RAG Chatbot (1M Token Dataset)

Overview:
This project implements a scalable Retrieval-Augmented Generation (RAG) chatbot capable of handling approximately 1,000,000 tokens of indexed content using LangChain and Chroma.

Architecture:
Dataset (~1M tokens) → Recursive Chunking (400 / 80 overlap) → OpenAI Embeddings → Chroma Vector Store → Adaptive Retriever (Similarity + MMR) → LLM (gpt-4o-mini) → Answer with Source Citations

Chunking Strategy:
chunk_size = 400
chunk_overlap = 80

Adaptive Retrieval:
TOP_K = 15
FETCH_K = 60
lambda_mult = 0.5

Technology Stack:
LangChain, Chroma, OpenAI Embeddings, gpt-4o-mini, Gradio, Google Colab.

Key Features:
- Handles ~1M token datasets
- Persistent vector storage
- Duplicate-safe indexing
- Source citation
- Adaptive retrieval

True RAG Design:
The model never receives 1M tokens at once. The vector database stores the knowledge externally and retrieves only relevant chunks per query.