Physics Book RAG - Scaffold Documentation
Generated: 2025-10-04 14:38 UTC

This document explains every file and directory created for the zero-cost Physics Book RAG project scaffold. Each section covers the file's role, the key contents it holds, and why it was added to support a fully local retrieval-augmented generation workflow backed by Ollama. The latest iteration also describes the multimodal and Streamlit additions that extract textbook figures and expose a browser-based chat interface.

Directory: data/raw
Purpose: Entry point for source textbooks. Place physics PDFs here before running ingestion so the pipeline can discover and process them without manual path tweaks.
Rationale: Ensures a dedicated location for unprocessed materials and keeps the repository tidy.

Directory: data/processed
Purpose: Reserved for intermediate artifacts such as cleaned text caches. Currently unused by default but ready for future preprocessing enhancements.
Rationale: Creating this upfront avoids reshuffling the structure when you introduce text cleaning stages.

Directory: data/processed/images
Purpose: Holds PNG snapshots of figures extracted from each textbook page (one subfolder per PDF).
Rationale: Lets the retrieval layer link to diagrams and charts when the user asks for visual references.

Directory: data/vector_store
Purpose: Stores FAISS indexes and JSONL metadata for both text and image embeddings.
Rationale: Centralizes retrieval assets so they can be versioned, backed up, or regenerated easily.

Directory: models
Purpose: Optional parking spot for exported models or experiment assets. Not required when using Ollama, but available if you later store quantized files for other runtimes.
Rationale: Keeps large inference artifacts out of the source tree should you choose to add them.

Directory: scripts
Purpose: Placeholder for helper scripts such as evaluation or maintenance utilities. Now also houses the Streamlit web app entry point.
Rationale: Encourages automation and keeps ad-hoc tooling out of the main package namespace.

File: scripts/streamlit_app.py
Purpose: Streamlit interface that mirrors the CLI experience, displaying retrieved figures inline and persisting chat history per browser session.
Rationale: Provides a point-and-click option for teammates who prefer a graphical interface over terminals.

Directory: configs
Purpose: Configuration files (YAML) live here. The default template documents tunable knobs for paths, chunking, retrieval, image support, and Ollama runtime settings.
Rationale: Bundles environment-specific values in one place so you can maintain multiple profiles.