1. Cài môi trường & thư viện

* Tạo venv/conda cho project → cài requirements.txt (hoặc environment.yml).
* Mục tiêu: máy bạn chạy được Python + Streamlit + Chroma + sentence-transformers.

1. Tạo file cấu hình .env

* Copy .env.example → .env.
* Điền OPENAI\_API\_KEY (nếu dùng cloud) hoặc OPENAI\_BASE\_URL (nếu dùng server local).
* Giữ DATA\_DIR, INDEX\_DIR, CHAT\_MODEL, EMBEDDING\_MODEL theo mặc định (có thể đổi sau).

1. Nạp dữ liệu vào data/

* Đặt tài liệu .txt/.md/.pdf vào data/.
* Đảm bảo .gitignore đang bỏ qua data/ và vector\_store/ (không push dữ liệu lên GitHub).

1. Build index (tạo embeddings)

* Chạy script build index để tạo/cập nhật vector\_store/.
* Đây là bước “chuẩn bị” cho truy hồi (retrieve).

1. Chạy ứng dụng (Streamlit)

* Khởi chạy UI → hỏi thử vài câu → xem câu trả lời & phần “Sources”.
* Kiểm tra nhanh độ đúng: top-K, chunk, hiển thị nguồn.

1. Vòng lặp cải thiện nhanh

* Điều chỉnh tham số (top-K, chunk size, model embedding).
* Bổ sung dữ liệu/định dạng lại tài liệu để dễ truy hồi.
* Khi ổn, commit/push thay đổi code/README (data & index vẫn giữ local).

1. (Tuỳ chọn) Đóng gói & triển khai

* Dockerfile sẵn: build image, run container (bind data/ & vector\_store/).
* Chọn nơi deploy sau nếu cần (giữ dữ liệu/index ở volume).

Bạn làm xong bước 1–5 là đã có bản chạy được để thử nghiệm; bước 6–7 là tối ưu & triển khai.