PRD One-Pager — Portfolio RAG (Local, Citations + Evals)

Grounded recruiter Q&A over Jeevan Deep Borugadda’s resume + portfolio PDFs

Owner: Jeevan Deep Borugadda | Date: 2026-02-26 | Status: v1 demo

Problem

- Recruiters and interviewers need fast, trustworthy answers about a candidate’s experience and project impact, but typical chatbots hallucinate or paraphrase without evidence.
- Goal is a QA assistant that answers only from approved documents with citations, and refuses when evidence is missing.

Target Users

- Recruiters / hiring managers (quick verification of role, impact, metrics).
- Interviewers (evidence-backed follow-ups and impact summaries).
- Jeevan (self-serve interview prep and rapid recall).

Goals (v1)

- Grounded answers over local PDFs with explicit citations per claim.
- Correct handling of perspective: user “I / my” refers to Jeevan (portfolio owner), not the assistant.
- No-guess behavior: refuse when sources do not contain evidence.
- Local-first demo: runs on a laptop with Ollama + Chroma (no paid APIs).

Non-Goals (v1)

- No web browsing or external data sources.
- Not a general-purpose chatbot — constrained to portfolio/resume documents.
- No public hosted backend in this version (demo is local).

Success Metrics (v1)

- Faithfulness (groundedness): ≥ 90% of golden-set answers supported by cited sources.
- Citation coverage: ≥ 95% of key claims contain citations.
- Refusal correctness: ≥ 90% correct refusals when evidence is absent.
- Median latency: < 4s locally on a typical Mac for short questions.

Solution Overview

- Ingestion: PDFs in demos/rag-local/data → text extraction (pypdf) → chunking (RecursiveCharacterTextSplitter).
- Chunking config (current): chunk_size=500, overlap=80, separators include newlines, bullets, hyphens.
- Embeddings: Ollama embeddings model nomic-embed-text → vectors stored in persistent local Chroma collection 'jeevan_portfolio'.
- Retrieval: vector search top-k (UI uses k=10 recommended). Resume-bias fallback for resume-type questions.
- Generation: Ollama chat model llama3.1:latest with strict grounding prompt; answer format is concise bullets with citations.

Key Product Behaviors

- Citations: each bullet must include [1], [2], ... referencing retrieved source chunks.
- Refusal: if no meaningful retrieved text, respond 'I don't have enough evidence in the documents to answer that.'
- Transparency: API returns retrieved chunks for debugging and iteration.

Risks & Mitigations

- PDF extraction issues (columns/spacing): add optional .txt versions of key docs; keep chunks smaller for resumes.
- Retrieval misses resume lines: increase top_k; resume keyword bias; consider a lightweight reranker later.
- Hallucinations: enforce 'answer only from sources' prompt + refusal + citation requirement.

Deployment / Demo

- Local backend: FastAPI on http://127.0.0.1:8000 (endpoints: /health, /ask).
- Local UI: Next.js route /demos/rag calling the local API.
- Portfolio integration: publish demo video + GitHub repo; optionally gate the live UI in production (show 'runs locally').