

FEATURED PROJECT · ONE-PAGE SUMMARY

Demo RAG V5: Trusted Knowledge Service

A production retrieval-augmented generation service built for one job: put trusted, cited, source-grounded answers in front of the person who depends on them. Built for AI Empower (medical education) as the V5 Storage Decoupling release; the same pattern serves the operator in the field: SOPs, maintenance manuals, and policy.

Role: Architect & Lead Engineer

Stack: Python 3.11 · FastAPI · Cloud Run · Pub/Sub · GCS · Firestore Native + Vector · Vertex AI (Gemini 2.5 Pro, text-embedding-004) · Terraform · Streamlit

Links: github.com/burnsgregm/ai-empower-rag-v5 · demo-rag-v5.streamlit.app

WHAT IT IS

An asynchronous GCP fan-out pipeline that ingests textbook-scale PDF/PPTX content, indexes it with a Parent-Child strategy in Firestore Vector Search, and serves conversational, cited answers backed by Gemini 2.5 Pro, fronted by a live Streamlit demo.

KEY CAPABILITIES

- **Async fan-out ingestion:** GCS to Eventarc to dispatcher to Pub/Sub to workers at page-level granularity; solved monolithic ingestion timeouts on 500+ page documents.
- **Parent-Child retrieval:** child vectors for precision, parent chunks for rich LLM context, plus conversational memory.
- **Dual-write storage decoupling:** child vectors persist to Firestore for real-time search while all RAG data archives to Parquet in Cloud Storage.
- **Migration-ready by design:** the Parquet layer prepares a library-scale upgrade to a dedicated vector engine without re-ingesting the corpus.
- **Multi-tenant retrieval API:** tenant isolation, chat history, and grounded answer generation.
- **Infrastructure as code:** three containerized Cloud Run services plus front end, provisioned via Terraform.

KEY OUTCOMES

Textbook → Library

Scale path without re-ingestion

Page-level

Fan-out ingestion granularity

Cited

Every answer grounded in source

BUSINESS IMPACT

Answers an operator can trust are answers with sources. The dual-write decoupling decision keeps retrieval fast today and the source of truth durable for tomorrow's scale, which is the difference between a demo and a system an organization can run.