ENTERPRISE AI & BUSINESS INTELLIGENCE

Local Financial AI Assistant

Architected by Abdelmonem Hatem

01 Executive summary

Private, evidence-backed answers over complex financial reports — without sending source documents to a hosted LLM.

01 CHALLENGE

Privacy, trust and traceability

Financial filings may be sensitive. General-purpose LLMs can produce unsupported figures, and answers without document citations are difficult to verify.

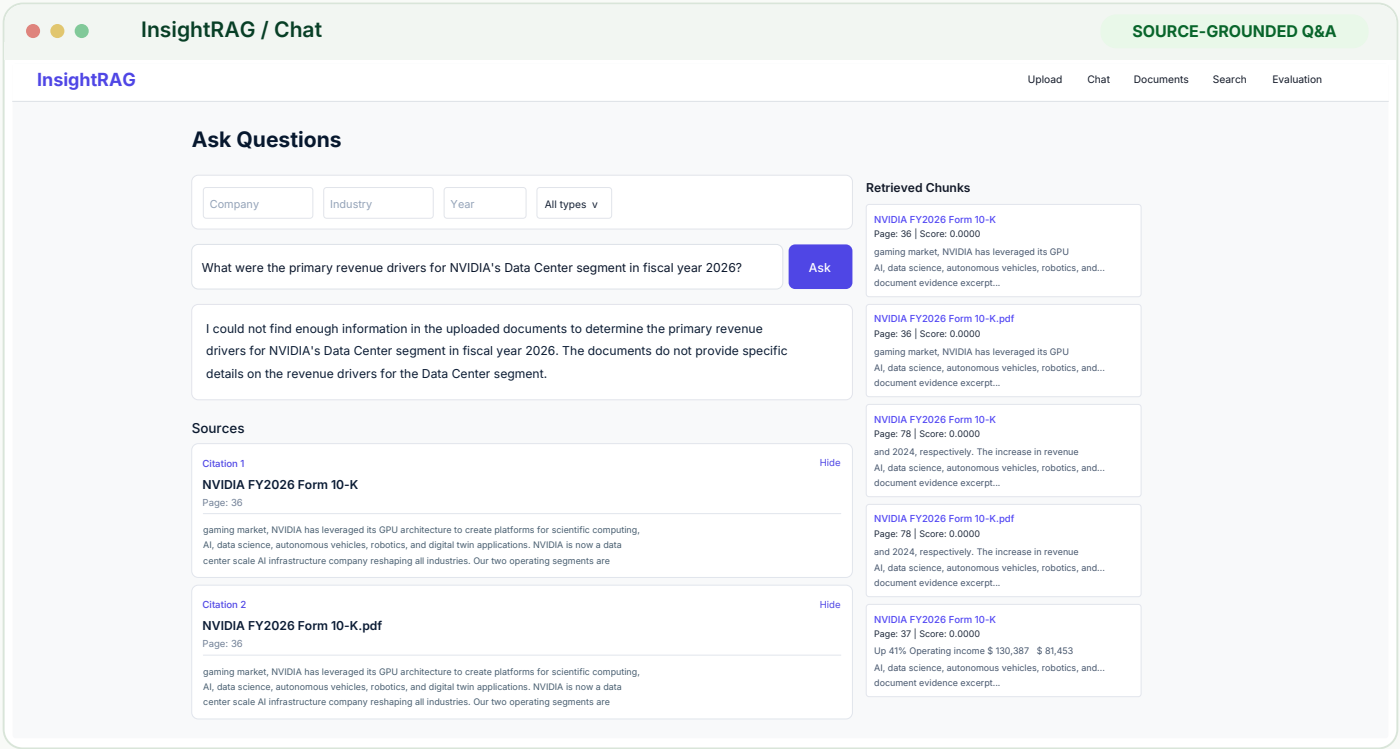
02 SOLUTION

Local, grounded financial RAG

A self-hosted retrieval workflow combines page-aware source chunks, Qdrant search and Ollama generation to return traceable, cited answers.

02 Product experience

Original InsightRAG interface · financial Q&A with source traceability



1 Upload reports → 2 Filter context → 3 Retrieve evidence → 4 Generate response → 5 Cite exact pages

03 Business impact

Project value proposition · quantitative figures are illustrative targets

PERFORMANCE

90%

Faster analysis

Target reduction in manual review time

DATA PRIVACY

100%

Local control

Offline-capable, no external model API

AUDITABILITY

PAGE

Citations

Grounding tied to the source document

OPERATING COST

$0

Token fees

Local inference; hardware costs remain

TECHNICAL ARCHITECTURE / LOCAL-FIRST RAG

System Architecture & Engineering

Lead Architect: Abdelmonem Hatem

01 Three-tier system architecture

A clearly separated interface, orchestration layer and local infrastructure — built for evidence-backed financial Q&A.

TIER 01 / UI

Next.js 14 Frontend

React + TypeScript web interface

USER OPERATIONS

- Upload 10-K documents
- Ask financial questions
- Filter by company and year
- Review citations and evaluations

HTTP / REST · JSON

TIER 02 / API

FastAPI + LangChain

Async endpoints, retrieval orchestration, prompting and page-aware response assembly.

REQUEST PIPELINE

- Parse query
- Apply filters
- Retrieve chunks
- Ground answer
- Score + cite

LOCAL SERVICES · DOCKER

TIER 03 / DATA + MODEL

Private infrastructure

Self-hosted services, isolated by role

Qdrant
VECTOR DATABASE
Cosine similarity + payload metadata filters

PostgreSQL 16
RELATIONAL PERSISTENCE
Document metadata, evaluations + audit logs

Ollama · Llama 3.1
LOCAL LLM RUNTIME
Response generation without hosted APIs

02 Engineering capabilities

01

Payload-filtered vector search

Qdrant cosine similarity combined with company, year and document payload filters.

02

Anti-hallucination guardrails

Page-aware evidence selection and source-linked answer generation.

03

LLM-as-a-judge evaluation

Faithfulness and relevance scoring with readable evaluation rationale.

03 Technology stack

- FastAPI
- Next.js 14
- LangChain
- Qdrant
- Docker
- Ollama
- PostgreSQL 16

AI **Custom AI & RAG Solutions**
Engineered by Abdelmonem Hatem