Architectural Overview

Overview:

The **Internal Fintech Agent** is a **Retrieval-Augmented Generation (RAG)** system designed to extract precise, grounded answers from internal documentation (Fintech_intake.docx).
It combines **semantic retrieval using ChromaDB** with **Groq's Llama-3.3-70B-Versatile LLM**, ensuring high contextual accuracy, dynamic query understanding, and plain-text, citation-free responses.

1. Large Language Model (LLM)

Attribute	Details
Model	Groq Llama-3.3-70B-Versatile (default)
API Provider	Groq Cloud (https://console.groq.com)
Purpose	Performs natural-language understanding, re-ranking, and grounded answer generation
Temperature	0 (deterministic)
Response Mode	Plain text only (no JSON, no citations)

The Groq model is used through the official **Groq Python SDK** (client.chat.completions.create) to ensure deterministic, high-precision outputs suitable for enterprise environments.

2. Embedding Model

Attribute	Details
Embedding Framework	sentence-transformers
Model	all-MiniLM-L6-v2
Vector Dimension	384
Role	Converts document chunks and user queries into dense semantic embeddings for similarity comparison

Embeddings are generated locally (no API calls) for each paragraph-level chunk extracted from the Fintech document, enabling offline vector indexing and retrieval.

3. Vector Database

Attribute	Details
Database	**ChromaDB**
Storage Path	storage/chroma/
Persistence	Local on-disk vector store
Functionality	Stores embeddings and supports fast top-K semantic similarity search
Metadata Stored	Chunk ID, section path, raw text

ChromaDB provides a lightweight, embedded vector database that supports persistent, low-latency semantic retrieval without external dependencies or internet connectivity.

4. RAG Pipeline

The Fintech Agent implements a **6-step Retrieval-Augmented Generation pipeline**:

Step	Stage	Description	Implemented In
1	**Ingestion**	Parses Fintech_intake.docx into normalized text chunks, each tagged with a section path.	scripts/ingest.py
2	**Embedding**	Uses all-MiniLM-L6-v2 to encode each chunk into a dense vector representation.	scripts/index.py
3	**Indexing**	Stores all vectors and metadata in a persistent ChromaDB collection.	scripts/index.py
4	**Retrieval**	For each query, embeds the query and retrieves top-K similar chunks from ChromaDB.	src/retriever.py
5	**Re-Ranking & Generation**	Passes retrieved chunks and the query to Groq's Llama-3.3 LLM to synthesize a concise, contextually supported answer.	src/agent.py
6	**Response Delivery**	Returns a plain-text answer via Flask API to the web UI.	src/server.py, templates/index.html

5. Data Flow

```
Fintech_intake.docx
       |
       ▼
 [scripts/ingest.py] → Text Chunks → Embeddings (MiniLM-L6-v2)
       |
       ▼
   ChromaDB (storage/chroma/)
       |
       ▼
  Query → Embedded → Top-K Chunks
       |
       ▼
  Groq Llama-3 LLM → Plain-Text Answer
       |
       ▼
    Web UI (Flask)
```

6. Key Design Principles

- **Precision First:** The LLM is temperature-controlled and context-restricted to retrieved text only.
- **Dynamic Retrieval:** ChromaDB vector search allows semantic understanding across paraphrased queries.
- **Privacy & Portability:** Fully local vector store; no data leaves the environment except LLM API calls.
- **Simplicity & Maintainability:** Each stage (ingest, index, retrieve, generate) is modular and script-based.
- **Explainability:** Every answer is traceable to the retrieved text in ChromaDB (internally logged).

7. Summary

Component	Technology	Role
LLM (Generator)	Groq Llama-3.3-70B-Versatile	Generates factual, context-grounded answers
Embedding Model	SentenceTransformer all-MiniLM-L6-v2	Converts text/query into semantic vectors
Vector Database	ChromaDB	Performs similarity search over embeddings
Retriever	Custom ChromaRetriever	Connects query embeddings → Chroma results

Backend	Flask (Python 3)	Hosts /chat endpoint and UI
Frontend	HTML + JS	Provides simple internal query interface

In summary:
The Internal Fintech Agent implements a complete **vector-based RAG architecture** using **ChromaDB** for semantic retrieval and **Groq Llama-3** for generation.
It achieves **accurate, explainable, and reproducible** retrieval-augmented answers suitable for internal enterprise deployment.