E-mail Semantic Search AI

GITHUB LINK: https://github.com/supriyaKU23/Email-Semantic-Search-AI

1. Introduction

Enron's historic e-mail corpus (> 1 million messages) contains priceless knowledge on **energy-trading tactics, regulatory strategy, and internal risk discourse**. Yet, keyword search buries insight beneath noise. Our solution converts this raw text into a **semantic knowledge service** that lets analysts, compliance officers, or litigators ask conversational questions and receive *curated, citation-rich answers in seconds*.

Value to the business:

- **Speed-to-insight ↑** — hours of manual digging collapse into seconds.
- **Reg-risk ↓** — pinpoint every mention of "round-trip trades" or "weather derivatives risk" for audit.
- **Knowledge retention** — institutional memory survives employee turnover.

DATASET: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset

2. Project Goals

#	Goal	Success Metric
1	Build a fault-tolerant vector index over the full 1 M-row corpus	100 % rows indexed w/ < 3 h wall-time @ Colab T4
2	Serve natural-language Q&A with < 5 s latency	p95 end-to-end latency < 5 s
3	Operate 100 % on commodity hardware (GPU opt-in)	Colab/T4 ✅ Local CPU fallback
4	Document architecture & automation for reproducibility	One-shot script + README + Flowchart

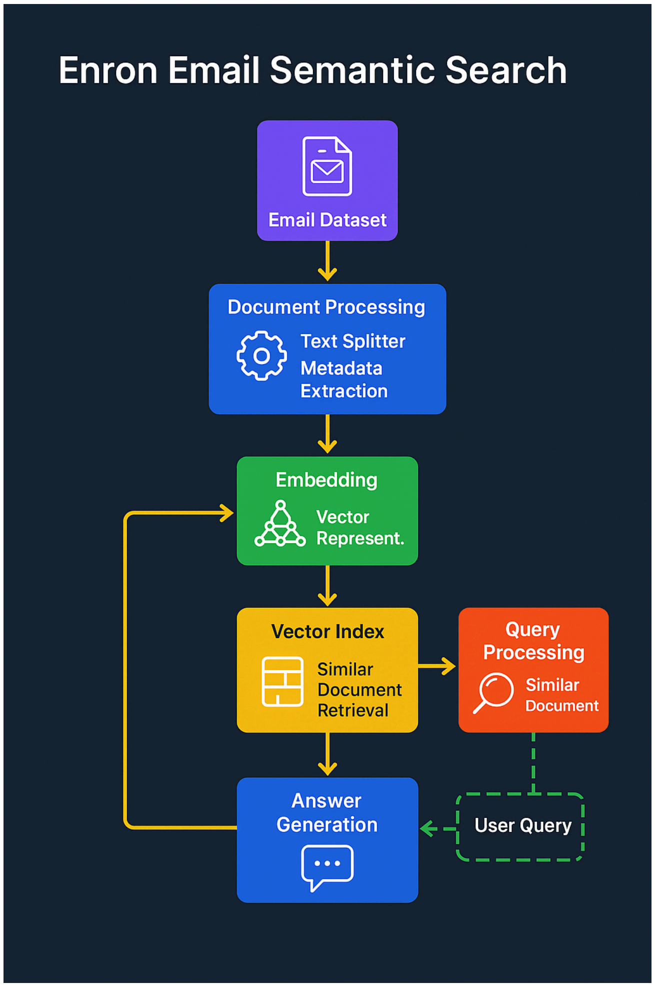
3. Data Sources

- **Raw** : `emails.csv` — cleaned Enron corpus (1 034 802 rows; columns: *file*, *message*, *date*, *subject*, …). After downloading from https://www.kaggle.com/datasets/wcukierski/enron-email-dataset

- **Vector store** : persisted **ChromaDB** collection `<DB_DIR>/emails` (~ 2 GB).

- **Index metadata** : auto-generated by *LlamaIndex*; persisted inside `<DB_DIR>`.

4. System-Design Flowchart



📧 1. Email Dataset

- **Purpose**: This is the source of raw data, specifically the Enron email dataset.
- **Dataset:** https://www.kaggle.com/datasets/wcukierski/enron-email-dataset
- **Contents**: Each email includes metadata (sender, receiver, subject, timestamp) and the main body of the email.
- **Role**: Acts as the input to the semantic search pipeline.

🛠️ 2. Document Processing

- **Text Splitter**: Breaks long email threads or documents into smaller, manageable text chunks. This helps fit within the token limits of embedding models.
- **Metadata Extraction**: Captures fields like sender, receiver, subject, and date. Useful for filtering and enhancing retrieval results.
- **Goal**: Clean and structure the raw emails for downstream processing.

🔗 3. Embedding (Vector Representation)

- **Role**: Converts text chunks into dense numerical vectors.
- **Technology**: Uses embedding models (e.g., OpenAI, BERT, Sentence Transformers).
- **Function**: Captures semantic meaning, enabling similarity comparison between chunks.

🫱 4. Vector Index (Similar Document Retrieval)

- **Role**: Stores and indexes the vector representations of emails.
- **Technology**: FAISS, Pinecone, Weaviate, or Milvus.
- **Function**: Supports fast similarity search using nearest neighbor algorithms.
- **Output**: Retrieves the most relevant email chunks for a given query vector.

🔍 5. Query Processing

- **Input**: User enters a natural language query.
- **Steps**:
  - Query is embedded using the same model as documents.
  - The query vector is compared to the index to find similar documents.
- **Goal**: Identify semantically similar email content.

💬 6. Answer Generation

- **Role**: Synthesizes a user-friendly response based on retrieved content.
- **Methods**:
  - Summarization of top documents.
  - Retrieval-Augmented Generation (RAG) using LLMs.
- **Output**: Final answer presented to the user.

🧑‍💻 7. User Query

- **Trigger**: The user inputs a question.
- **Connection**: Initiates query processing and downstream retrieval and generation.

♻️ Flow Summary

1. Load and process the raw Enron email dataset.
2. Split text and extract metadata.
3. Generate vector representations.
4. Store vectors in an index.
5. Embed user query and search the index.
6. Retrieve top documents.
7. Generate and return the answer.

---

5. Design Choices & Rationale

1. **ChromaDB (PersistentClient)** — fastest open-source vector DB w/ disk persistence; easy Colab install.

2. **HuggingFace `all-MiniLM-L6-v2`** — 384-dim embedding; balances recall vs. speed; no API cost.

3. **ThreadPool parallel chunking** — ~8 × faster than pandas loop; CPU-bound.

4. **2048-token chunks / 128 overlap** — matches GPT-3.5 context; avoids fragmenting long threads.

5. **Insert-nodes incremental upsert** — avoids rebuilding index each batch; memory-safe for 1 M rows.

6. **Progress-file resume** — survives Colab disconnect/sleep.

7. **OpenAI chat-LLM at query time** — zero cost during indexing; pay-per-query.

---

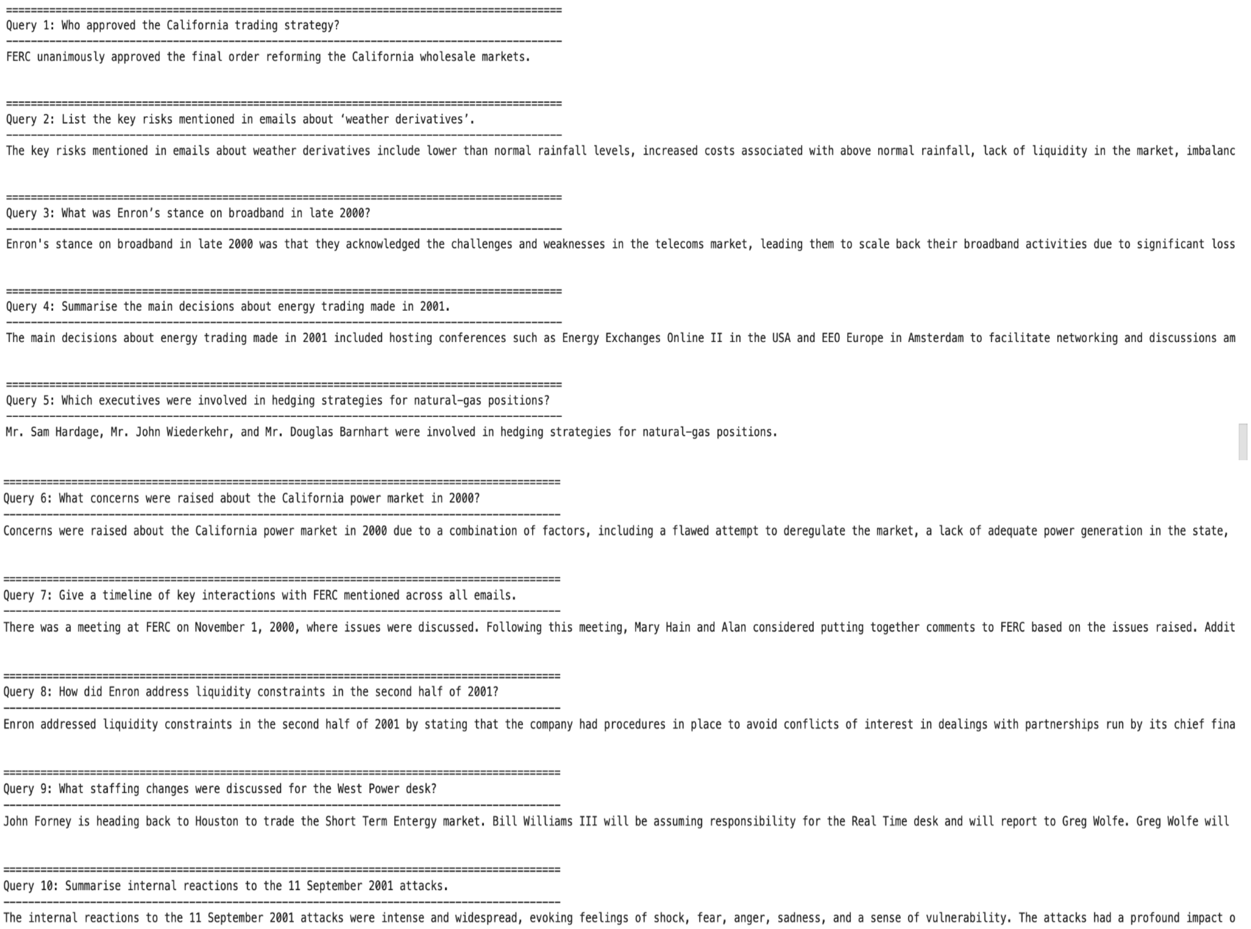
6. Key Challenges & Mitigations

Challenge	Mitigation
Colab 12 h idle timeout	Progress checkpoint + resume logic
OpenAI rate limit / key errors	API key via `os.environ`, exponential retry, local HF embeddings
Chroma telemetry spam	Silence via `posthog.api_key=""` env var
Memory pressure (> 8 GB)	Process 10 k-row batches, flush GPU cache, `gc.collect()`

---

7. Example Queries (expandable)

```
queries = [
    "Who approved the California trading strategy?",
    "List the key risks mentioned in emails about ‘weather derivatives’.",
    "What was Enron’s stance on broadband in late 2000?",
    "Summarise discussions on round-trip trades in 2001.",
    "Give 3 e-mail snippets illustrating concern over mark-to-market accounting.",
    "Which executives were looped in on weather-derivatives hedging proposals?",
    "Create a timeline of major energy-trading decisions between 1999-2001.",
    "Extract action-items from Jeff Skilling’s mails in Q4-2000.",
    "Did anyone flag regulatory risk for California power trades before 2001?",
    "Summarise positions on broadband build-out vs. lease in June 2000.",
]
for i,q in enumerate(queries,1):
    print(f"Q{i:02}: {q}\n{q_engine.query(q)}\n{'-'*60}")
```



```
==========================================================================================
Query 1: Who approved the California trading strategy?
------------------------------------------------------------------------------------------
FERC unanimously approved the final order reforming the California wholesale markets.


==========================================================================================
Query 2: List the key risks mentioned in emails about ‘weather derivatives’.
------------------------------------------------------------------------------------------
The key risks mentioned in emails about weather derivatives include lower than normal rainfall levels, increased costs associated with above normal rainfall, lack of liquidity in the market, imbalanc


==========================================================================================
Query 3: What was Enron’s stance on broadband in late 2000?
------------------------------------------------------------------------------------------
Enron's stance on broadband in late 2000 was that they acknowledged the challenges and weaknesses in the telecoms market, leading them to scale back their broadband activities due to significant loss


==========================================================================================
Query 4: Summarise the main decisions about energy trading made in 2001.
------------------------------------------------------------------------------------------
The main decisions about energy trading made in 2001 included hosting conferences such as Energy Exchanges Online II in the USA and EEO Europe in Amsterdam to facilitate networking and discussions am


==========================================================================================
Query 5: Which executives were involved in hedging strategies for natural-gas positions?
------------------------------------------------------------------------------------------
Mr. Sam Hardage, Mr. John Wiederkehr, and Mr. Douglas Barnhart were involved in hedging strategies for natural-gas positions.


==========================================================================================
Query 6: What concerns were raised about the California power market in 2000?
------------------------------------------------------------------------------------------
Concerns were raised about the California power market in 2000 due to a combination of factors, including a flawed attempt to deregulate the market, a lack of adequate power generation in the state,


==========================================================================================
Query 7: Give a timeline of key interactions with FERC mentioned across all emails.
------------------------------------------------------------------------------------------
There was a meeting at FERC on November 1, 2000, where issues were discussed. Following this meeting, Mary Hain and Alan considered putting together comments to FERC based on the issues raised. Addit


==========================================================================================
Query 8: How did Enron address liquidity constraints in the second half of 2001?
------------------------------------------------------------------------------------------
Enron addressed liquidity constraints in the second half of 2001 by stating that the company had procedures in place to avoid conflicts of interest in dealings with partnerships run by its chief fina


==========================================================================================
Query 9: What staffing changes were discussed for the West Power desk?
------------------------------------------------------------------------------------------
John Forney is heading back to Houston to trade the Short Term Entergy market. Bill Williams III will be assuming responsibility for the Real Time desk and will report to Greg Wolfe. Greg Wolfe will


==========================================================================================
Query 10: Summarise internal reactions to the 11 September 2001 attacks.
------------------------------------------------------------------------------------------
The internal reactions to the 11 September 2001 attacks were intense and widespread, evoking feelings of shock, fear, anger, sadness, and a sense of vulnerability. The attacks had a profound impact o
```

8. How to Re-Run or Extend

1. **Install**: `pip install "chromadb>=0.5.17" "llama-index==0.12.43"` (or latest matching pair).
2. **Swap Embeddings**: set `HF_MODEL="sentence-transformers/all-mpnet-base-v2"`.
3. **Deploy**: wrap `q_engine` behind FastAPI or Streamlit.