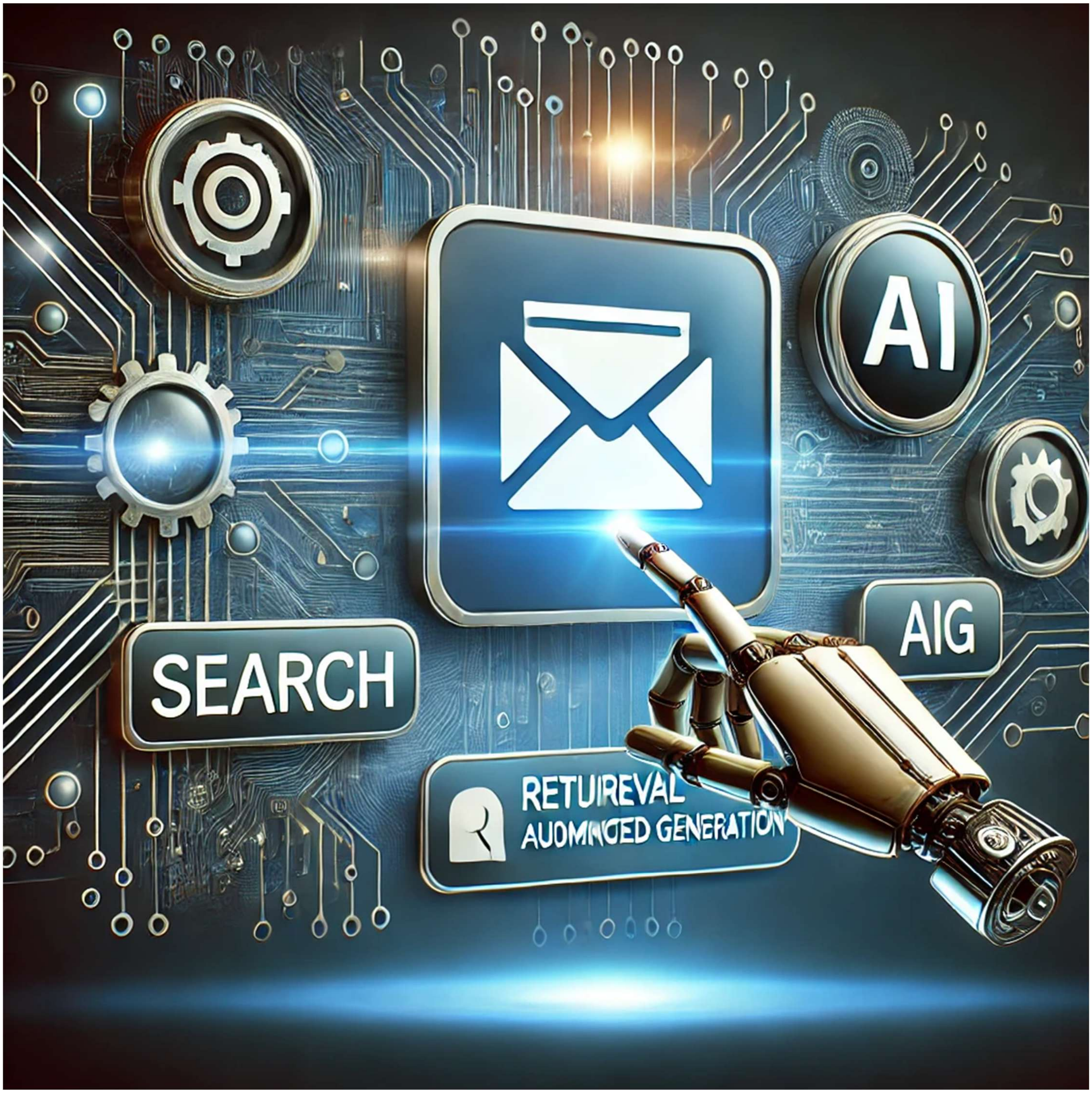
Email Search AI – Project Report



Prepared by

Kasi Viswanadh Maddala

Problem Statement

In enterprise environments, email threads often contain critical discussions, decisions, and context across multiple stakeholders. However, locating specific information within large, unstructured, and nested email threads is time-consuming and inefficient using conventional keyword-based search tools. Professionals face challenges in extracting relevant insights without wading through entire conversations manually. There is a pressing need for an intelligent system that enables **semantic search** and **automated summarization** of email threads.

Project Overview

Email Search_AI is an end-to-end **RAG-based AI assistant** designed for semantic search and summarization of email threads. It uses **Sentence Transformers** for creating dense **vector embeddings** of email content, stores them in a **ChromaDB vector store**, and enables **semantic index search**. Upon receiving a user query, it retrieves the most relevant chunks using **vector similarity**, improves result quality through **cross-encoder-based reranking**, and finally, generates context-aware responses using **OpenAI's LLM (e.g., GPT-3.5-turbo)**.

The system employs a **Retrieval-Augmented Generation (RAG)** architecture with **caching** for efficiency and performance.

Project Objectives

- **Semantic Understanding of Emails**: Use **Sentence Transformers (all-MiniLM-L6-v2)** to convert email chunks into vector representations capturing semantic meaning.
- **Vector Database Indexing**: Store email embeddings in **ChromaDB**, enabling fast approximate nearest neighbor (ANN) vector search.
- **Semantic Search & Retrieval**: Support user queries via **embedding-based similarity search** across indexed email chunks.
- **Result Reranking**: Improve retrieval accuracy with **cross-encoder reranking (ms-marco-MiniLM-L-6-v2)**, scoring relevance between query and result pairs.
- **Contextual Answer Generation**: Use a **Retrieval-Augmented Generation (RAG)** pipeline to feed retrieved results into OpenAI GPT models for answer synthesis.
- **Query Caching**: Implement a file-based **caching** layer to store query results and avoid repeated computations.

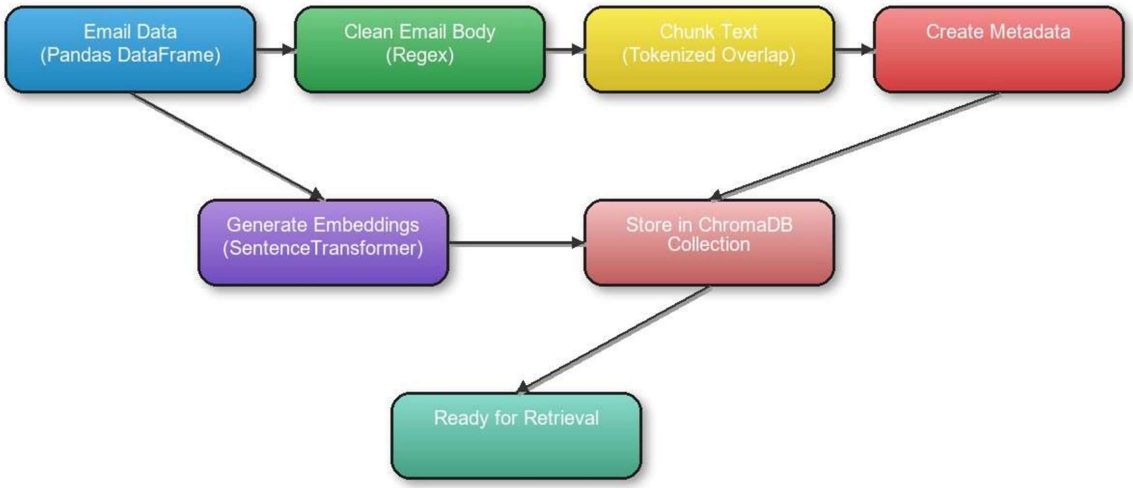
Functional Features:

Component	Description	Technology Used
Email Preprocessing	Cleans raw email bodies by removing quoted replies and normalizing text	Regex, custom cleaning
Chunking	Splits cleaned emails into overlapping token-limited chunks	Custom logic
Embeddings	Transforms email chunks into dense semantic vectors	SentenceTransformer (all-MiniLM-L6-v2)
Vector Indexing	Stores and indexes embeddings for fast similarity search	ChromaDB
Query Embedding	Embeds natural language queries for semantic comparison	SentenceTransformer
Initial Vector Search	Retrieves top-N similar email chunks using ANN search	ChromaDB
Reranking	Reorders retrieved results by true semantic relevance	CrossEncoder (ms-marco-MiniLM-L-6-v2)
Retrieval-Augmented Generation (RAG)	Combines retrieved chunks with the query to form a prompt for GPT	OpenAI GPT-3.5-turbo
Answer Generation	Synthesizes a coherent answer based on context	OpenAI Chat Completion API
Caching	Stores query results using hashed query keys for faster repeat access	JSON file-based custom Cache class

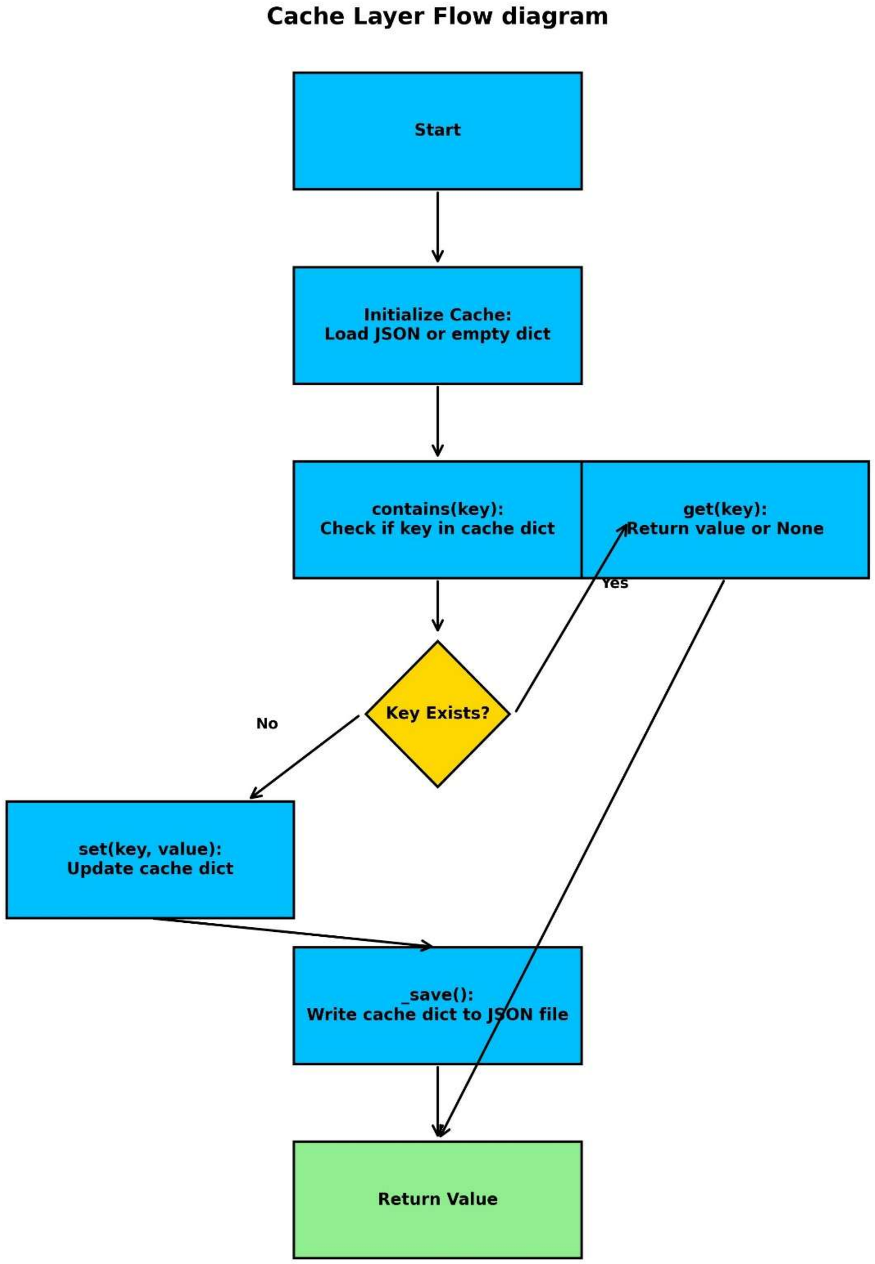
Overall Structure of the code:

```
email-search-ai/
│
├── email_dataset/
│   └── email_thread_details.csv
│   └── email_thread_summaries.csv
│
├── src/
│   ├── embedding_layer.py
│   ├── search_layer.py
│   ├── cache.py
│   ├── generation_layer.py
│   └── utils.py
│
├── screenshots_of_outputs/
│   ├── search_screenshots/
│   ├── generation_screenshots/
│
├── docs/
│   └── project_documentation.md
│
├── main.py
├── app.py
└── requirements.txt
```

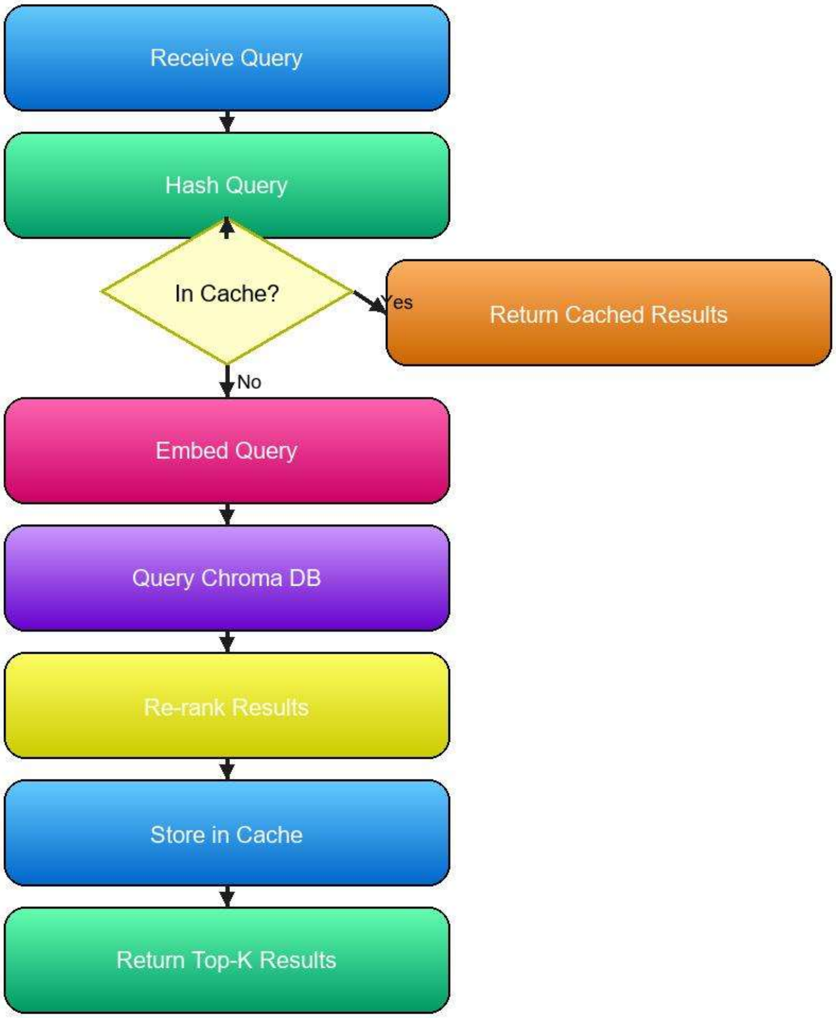
Embedding Layer



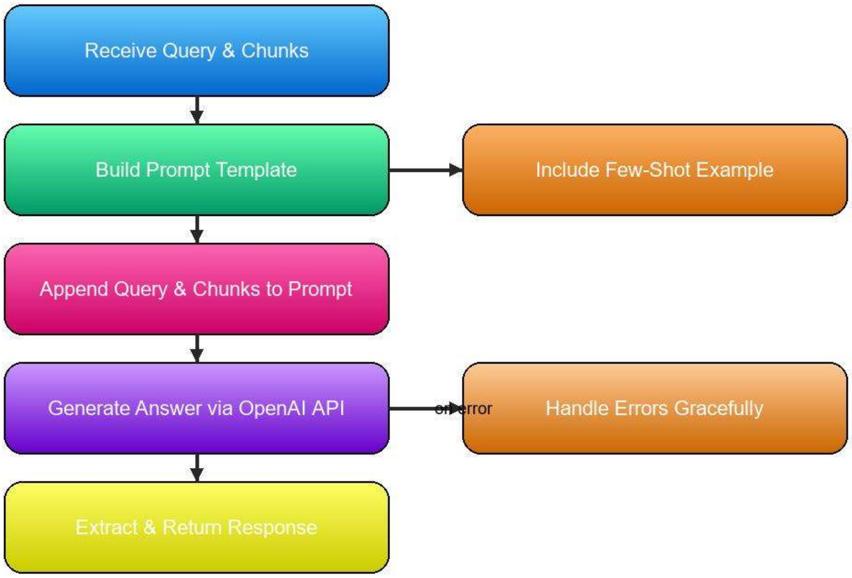
Cache Layer



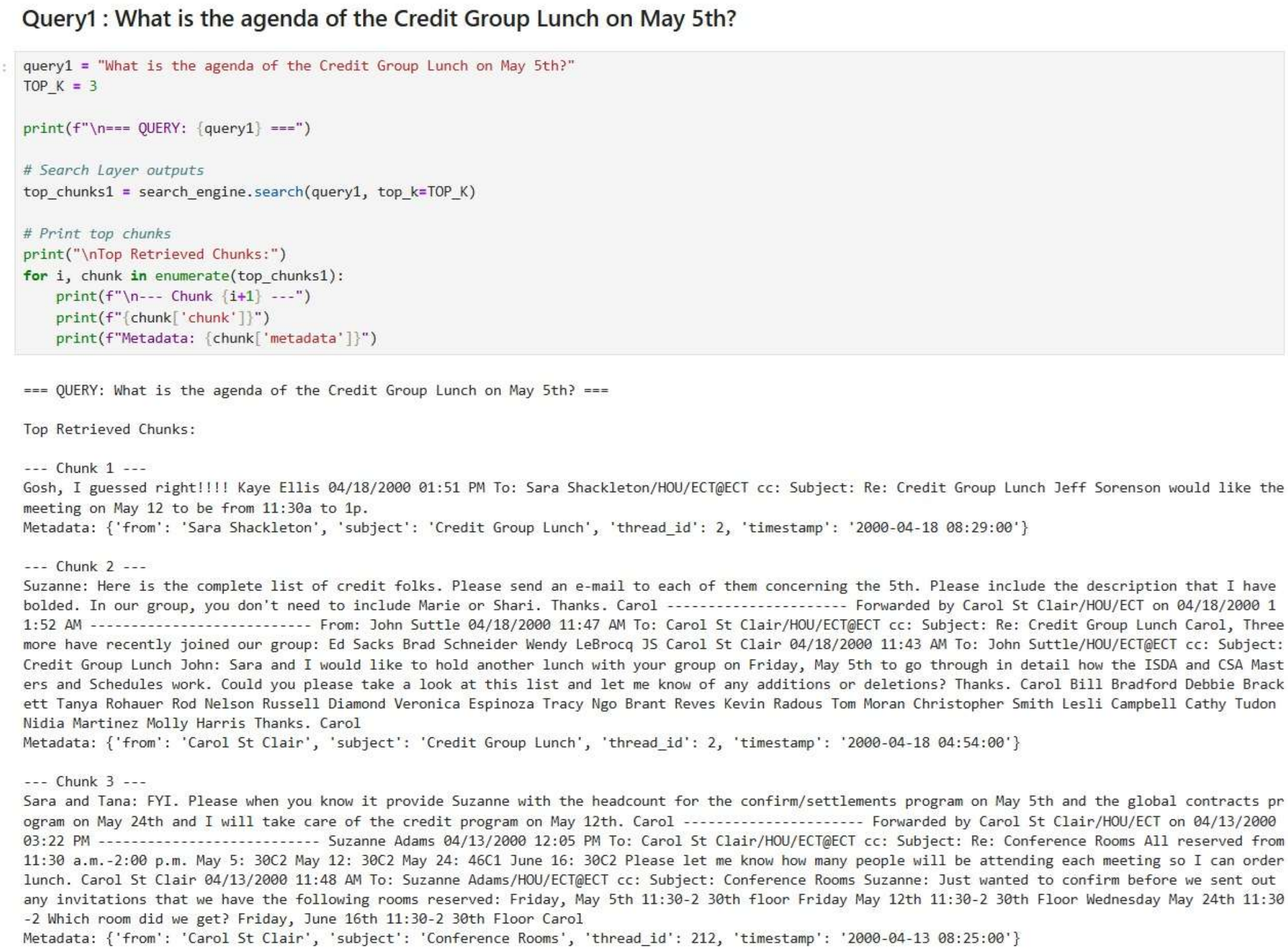
Search Layer



Generation Layer



Query1 – Top3 Answers

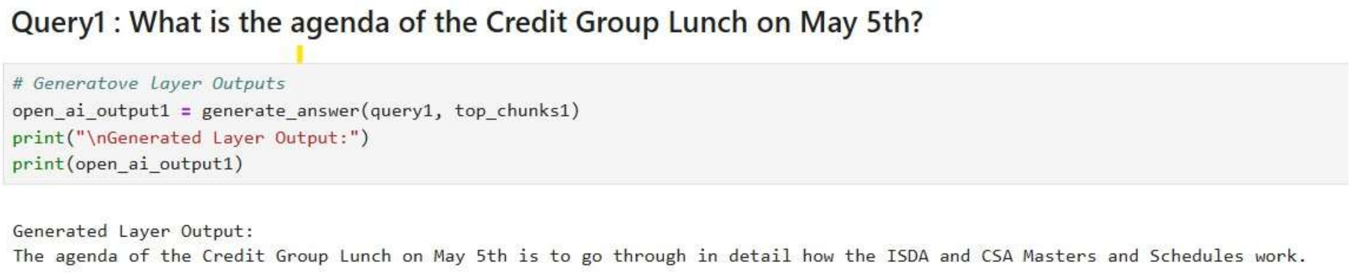


Query1 : What is the agenda of the Credit Group Lunch on May 5th?

```
query1 = "What is the agenda of the Credit Group Lunch on May 5th?"
TOP_K = 3

print(f"\n=== QUERY: {query1} ===")

# Search Layer outputs
top_chunks1 = search_engine.search(query1, top_k=TOP_K)

# Print top chunks
print("\nTop Retrieved Chunks:")
for i, chunk in enumerate(top_chunks1):
    print(f"\n--- Chunk {i+1} ---")
    print(f"{chunk['chunk']}")
    print(f"Metadata: {chunk['metadata']}")
```

```
=== QUERY: What is the agenda of the Credit Group Lunch on May 5th? ===

Top Retrieved Chunks:

--- Chunk 1 ---
Gosh, I guessed right!!!! Kaye Ellis 04/18/2000 01:51 PM To: Sara Shackleton/HOU/ECT@ECT cc: Subject: Re: Credit Group Lunch Jeff Sorenson would like the
meeting on May 12 to be from 11:30a to 1p.
Metadata: {'from': 'Sara Shackleton', 'subject': 'Credit Group Lunch', 'thread_id': 2, 'timestamp': '2000-04-18 08:29:00'}

--- Chunk 2 ---
Suzanne: Here is the complete list of credit folks. Please send an e-mail to each of them concerning the 5th. Please include the description that I have
bolded. In our group, you don't need to include Marie or Shari. Thanks. Carol --------------------- Forwarded by Carol St Clair/HOU/ECT on 04/18/2000 1
1:52 AM --------------------------- From: John Suttle 04/18/2000 11:47 AM To: Carol St Clair/HOU/ECT@ECT cc: Subject: Re: Credit Group Lunch Carol, Three
more have recently joined our group: Ed Sacks Brad Schneider Wendy LeBrocq JS Carol St Clair 04/18/2000 11:43 AM To: John Suttle/HOU/ECT@ECT cc: Subject:
Credit Group Lunch John: Sara and I would like to hold another lunch with your group on Friday, May 5th to go through in detail how the ISDA and CSA Mast
ers and Schedules work. Could you please take a look at this list and let me know of any additions or deletions? Thanks. Carol Bill Bradford Debbie Brack
ett Tanya Rohauer Rod Nelson Russell Diamond Veronica Espinoza Tracy Ngo Brant Reves Kevin Radous Tom Moran Christopher Smith Lesli Campbell Cathy Tudon
Nidia Martinez Molly Harris Thanks. Carol
Metadata: {'from': 'Carol St Clair', 'subject': 'Credit Group Lunch', 'thread_id': 2, 'timestamp': '2000-04-18 04:54:00'}

--- Chunk 3 ---
Sara and Tana: FYI. Please when you know it provide Suzanne with the headcount for the confirm/settlements program on May 5th and the global contracts pr
ogram on May 24th and I will take care of the credit program on May 12th. Carol --------------------- Forwarded by Carol St Clair/HOU/ECT on 04/13/2000
03:22 PM --------------------------- Suzanne Adams 04/13/2000 12:05 PM To: Carol St Clair/HOU/ECT@ECT cc: Subject: Re: Conference Rooms All reserved from
11:30 a.m.-2:00 p.m. May 5: 30C2 May 12: 30C2 May 24: 46C1 June 16: 30C2 Please let me know how many people will be attending each meeting so I can order
lunch. Carol St Clair 04/13/2000 11:48 AM To: Suzanne Adams/HOU/ECT@ECT cc: Subject: Conference Rooms Suzanne: Just wanted to confirm before we sent out
any invitations that we have the following rooms reserved: Friday, May 5th 11:30-2 30th floor Friday May 12th 11:30-2 30th Floor Wednesday May 24th 11:30
-2 Which room did we get? Friday, June 16th 11:30-2 30th Floor Carol
Metadata: {'from': 'Carol St Clair', 'subject': 'Conference Rooms', 'thread_id': 212, 'timestamp': '2000-04-13 08:25:00'}
```

Query1 – Final Answer

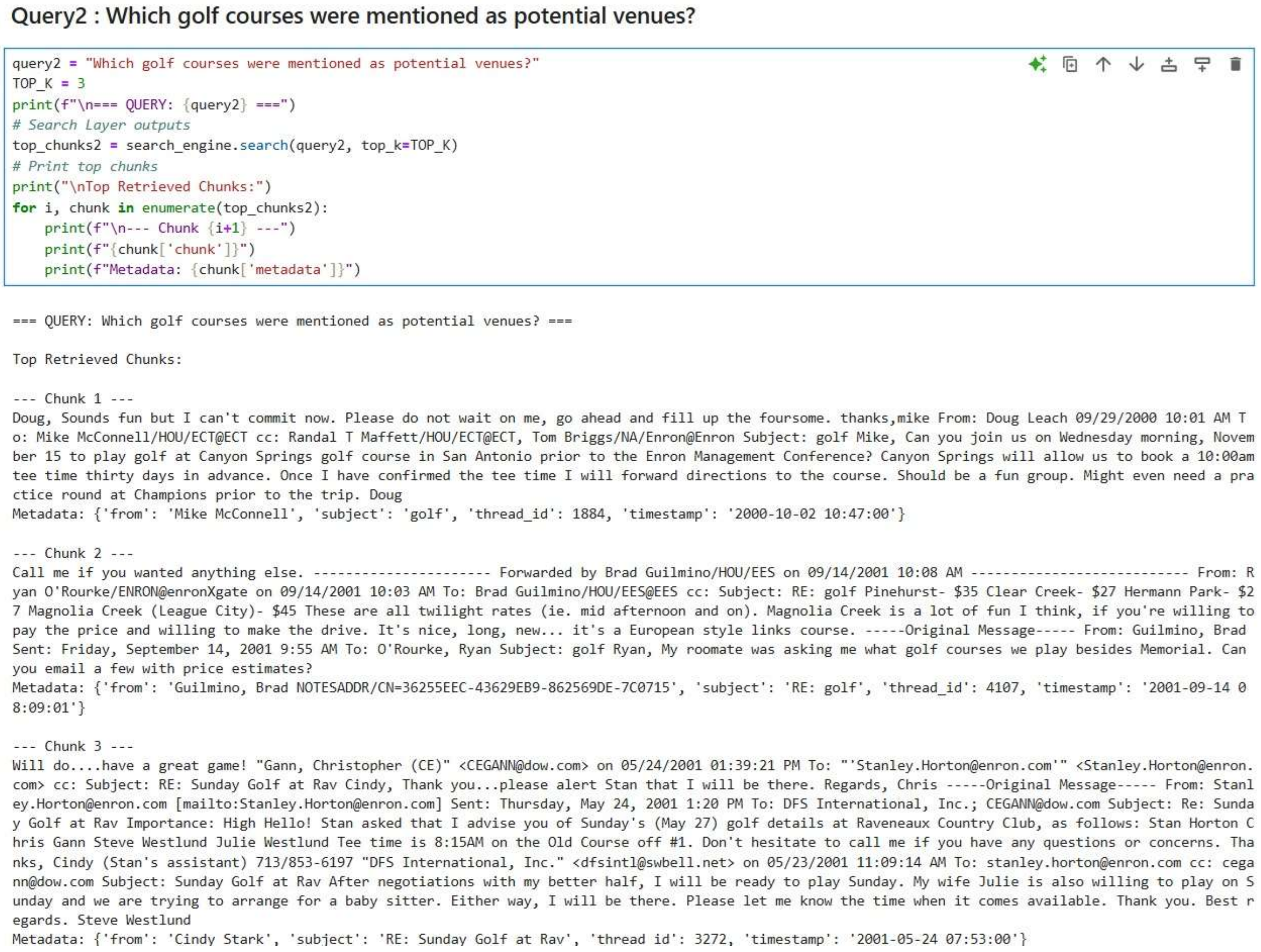


Query1 : What is the agenda of the Credit Group Lunch on May 5th?

```
# Generatove layer Outputs
open_ai_output1 = generate_answer(query1, top_chunks1)
print("\nGenerated Layer Output:")
print(open_ai_output1)
```

```
Generated Layer Output:
The agenda of the Credit Group Lunch on May 5th is to go through in detail how the ISDA and CSA Masters and Schedules work.
```

Query2 – Top 3 Answers

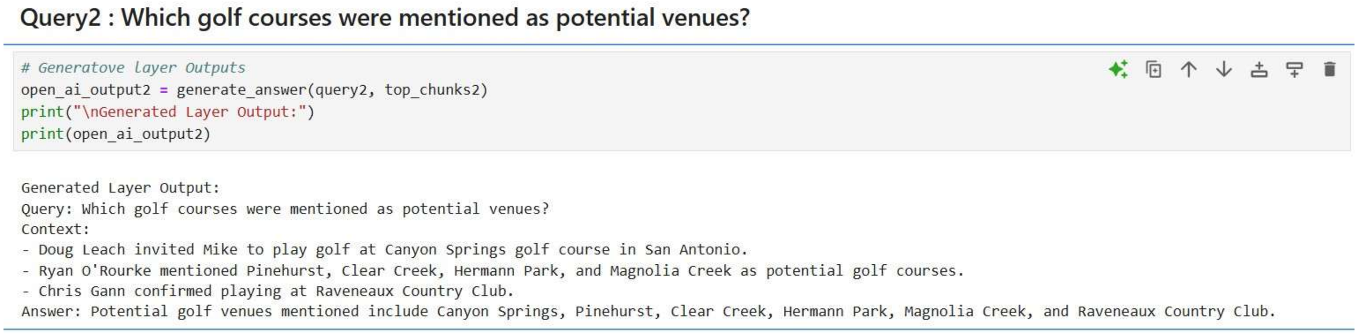


Query2 : Which golf courses were mentioned as potential venues?

```python
query2 = "Which golf courses were mentioned as potential venues?"
TOP_K = 3
print(f"\n=== QUERY: {query2} ===")
# Search Layer outputs
top_chunks2 = search_engine.search(query2, top_k=TOP_K)
# Print top chunks
print("\nTop Retrieved Chunks:")
for i, chunk in enumerate(top_chunks2):
    print(f"\n--- Chunk {i+1} ---")
    print(f"{chunk['chunk']}")
    print(f"Metadata: {chunk['metadata']}")
```

```
=== QUERY: Which golf courses were mentioned as potential venues? ===

Top Retrieved Chunks:

--- Chunk 1 ---
Doug, Sounds fun but I can't commit now. Please do not wait on me, go ahead and fill up the foursome. thanks,mike From: Doug Leach 09/29/2000 10:01 AM T
o: Mike McConnell/HOU/ECT@ECT cc: Randal T Maffett/HOU/ECT@ECT, Tom Briggs/NA/Enron@Enron Subject: golf Mike, Can you join us on Wednesday morning, Novem
ber 15 to play golf at Canyon Springs golf course in San Antonio prior to the Enron Management Conference? Canyon Springs will allow us to book a 10:00am
tee time thirty days in advance. Once I have confirmed the tee time I will forward directions to the course. Should be a fun group. Might even need a pra
ctice round at Champions prior to the trip. Doug
Metadata: {'from': 'Mike McConnell', 'subject': 'golf', 'thread_id': 1884, 'timestamp': '2000-10-02 10:47:00'}

--- Chunk 2 ---
Call me if you wanted anything else. ---------------------- Forwarded by Brad Guilmino/HOU/EES on 09/14/2001 10:08 AM --------------------------- From: R
yan O'Rourke/ENRON@enronXgate on 09/14/2001 10:03 AM To: Brad Guilmino/HOU/EES@EES cc: Subject: RE: golf Pinehurst- $35 Clear Creek- $27 Hermann Park- $2
7 Magnolia Creek (League City)- $45 These are all twilight rates (ie. mid afternoon and on). Magnolia Creek is a lot of fun I think, if you're willing to
pay the price and willing to make the drive. It's nice, long, new... it's a European style links course. -----Original Message----- From: Guilmino, Brad
Sent: Friday, September 14, 2001 9:55 AM To: O'Rourke, Ryan Subject: golf Ryan, My roomate was asking me what golf courses we play besides Memorial. Can
you email a few with price estimates?
Metadata: {'from': 'Guilmino, Brad NOTESADDR/CN=36255EEC-43629EB9-862569DE-7C0715', 'subject': 'RE: golf', 'thread_id': 4107, 'timestamp': '2001-09-14 0
8:09:01'}

--- Chunk 3 ---
Will do....have a great game! "Gann, Christopher (CE)" <CEGANN@dow.com> on 05/24/2001 01:39:21 PM To: "'Stanley.Horton@enron.com'" <Stanley.Horton@enron.
com> cc: Subject: RE: Sunday Golf at Rav Cindy, Thank you...please alert Stan that I will be there. Regards, Chris -----Original Message----- From: Stanl
ey.Horton@enron.com [mailto:Stanley.Horton@enron.com] Sent: Thursday, May 24, 2001 1:20 PM To: DFS International, Inc.; CEGANN@dow.com Subject: Re: Sunda
y Golf at Rav Importance: High Hello! Stan asked that I advise you of Sunday's (May 27) golf details at Raveneaux Country Club, as follows: Stan Horton C
hris Gann Steve Westlund Julie Westlund Tee time is 8:15AM on the Old Course off #1. Don't hesitate to call me if you have any questions or concerns. Tha
nks, Cindy (Stan's assistant) 713/853-6197 "DFS International, Inc." <dfsintl@swbell.net> on 05/23/2001 11:09:14 AM To: stanley.horton@enron.com cc: cega
nn@dow.com Subject: Sunday Golf at Rav After negotiations with my better half, I will be ready to play Sunday. My wife Julie is also willing to play on S
unday and we are trying to arrange for a baby sitter. Either way, I will be there. Please let me know the time when it comes available. Thank you. Best r
egards. Steve Westlund
Metadata: {'from': 'Cindy Stark', 'subject': 'RE: Sunday Golf at Rav', 'thread id': 3272, 'timestamp': '2001-05-24 07:53:00'}
```

Query2 – Final Answer

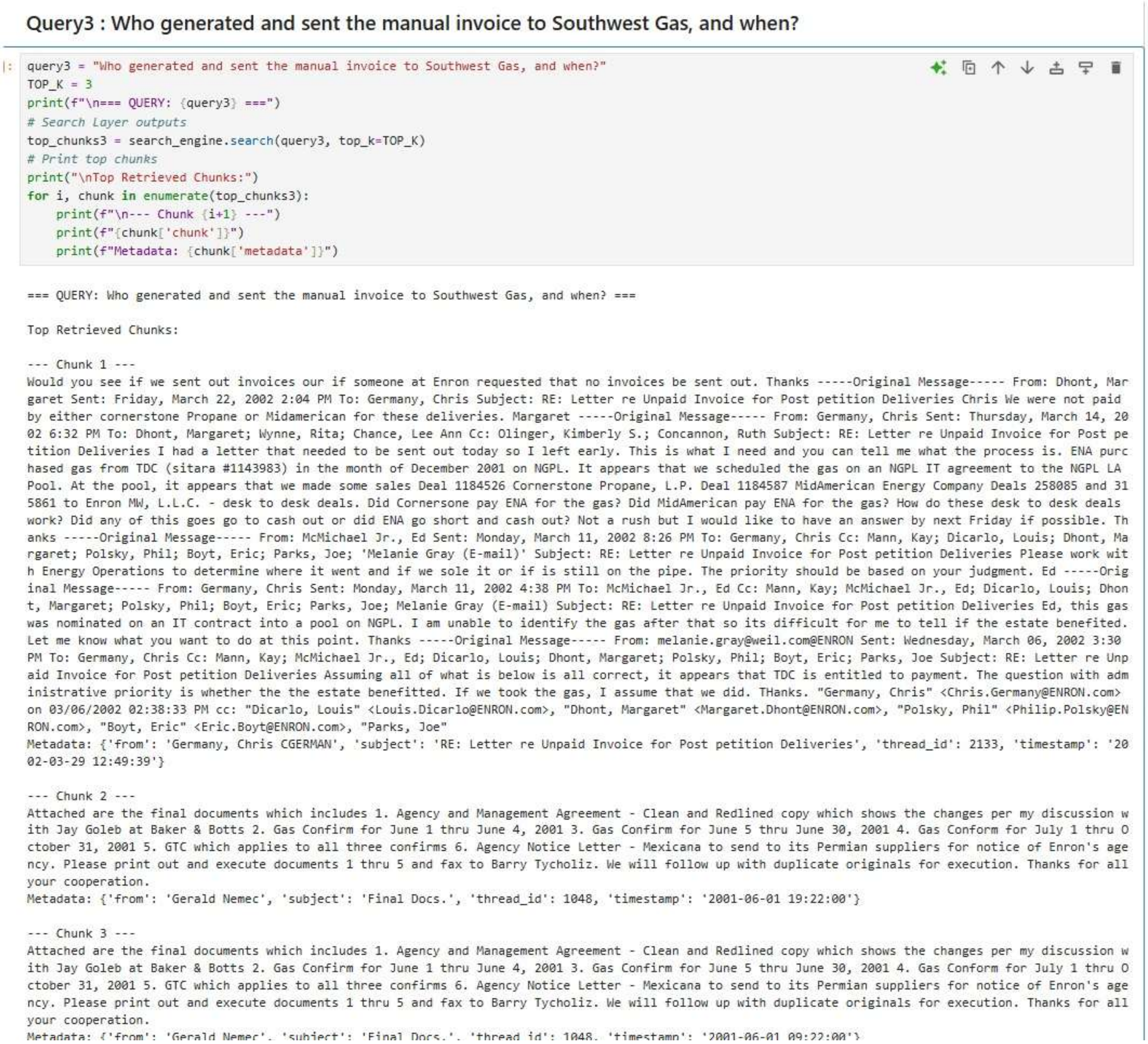


Query2 : Which golf courses were mentioned as potential venues?

```python
# Generatove layer Outputs
open_ai_output2 = generate_answer(query2, top_chunks2)
print("\nGenerated Layer Output:")
print(open_ai_output2)
```

```
Generated Layer Output:
Query: Which golf courses were mentioned as potential venues?
Context:
- Doug Leach invited Mike to play golf at Canyon Springs golf course in San Antonio.
- Ryan O'Rourke mentioned Pinehurst, Clear Creek, Hermann Park, and Magnolia Creek as potential golf courses.
- Chris Gann confirmed playing at Raveneaux Country Club.
Answer: Potential golf venues mentioned include Canyon Springs, Pinehurst, Clear Creek, Hermann Park, Magnolia Creek, and Raveneaux Country Club.
```

Query3 – Top3 Answers

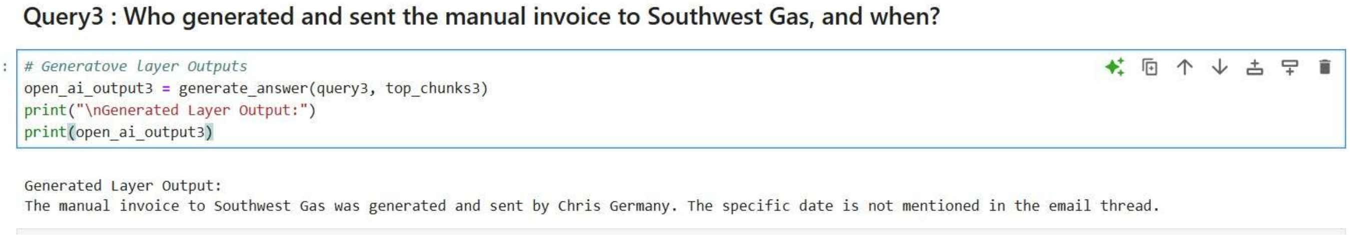


Query3 : Who generated and sent the manual invoice to Southwest Gas, and when?

```
query3 = "Who generated and sent the manual invoice to Southwest Gas, and when?"
TOP_K = 3
print(f"\n=== QUERY: {query3} ===")
# Search Layer outputs
top_chunks3 = search_engine.search(query3, top_k=TOP_K)
# Print top chunks
print("\nTop Retrieved Chunks:")
for i, chunk in enumerate(top_chunks3):
    print(f"\n--- Chunk {i+1} ---")
    print(f"{chunk['chunk']}")
    print(f"Metadata: {chunk['metadata']}")
```

```
=== QUERY: Who generated and sent the manual invoice to Southwest Gas, and when? ===

Top Retrieved Chunks:

--- Chunk 1 ---
Would you see if we sent out invoices our if someone at Enron requested that no invoices be sent out. Thanks -----Original Message----- From: Dhont, Mar
garet Sent: Friday, March 22, 2002 2:04 PM To: Germany, Chris Subject: RE: Letter re Unpaid Invoice for Post petition Deliveries Chris We were not paid
by either cornerstone Propane or Midamerican for these deliveries. Margaret -----Original Message----- From: Germany, Chris Sent: Thursday, March 14, 20
02 6:32 PM To: Dhont, Margaret; Wynne, Rita; Chance, Lee Ann Cc: Olinger, Kimberly S.; Concannon, Ruth Subject: RE: Letter re Unpaid Invoice for Post pe
tition Deliveries I had a letter that needed to be sent out today so I left early. This is what I need and you can tell me what the process is. ENA purc
hased gas from TDC (sitara #1143983) in the month of December 2001 on NGPL. It appears that we scheduled the gas on an NGPL IT agreement to the NGPL LA
Pool. At the pool, it appears that we made some sales Deal 1184526 Cornerstone Propane, L.P. Deal 1184587 MidAmerican Energy Company Deals 258085 and 31
5861 to Enron MW, L.L.C. - desk to desk deals. Did Cornersone pay ENA for the gas? Did MidAmerican pay ENA for the gas? How do these desk to desk deals
work? Did any of this goes go to cash out or did ENA go short and cash out? Not a rush but I would like to have an answer by next Friday if possible. Th
anks -----Original Message----- From: McMichael Jr., Ed Sent: Monday, March 11, 2002 8:26 PM To: Germany, Chris Cc: Mann, Kay; Dicarlo, Louis; Dhont, Ma
rgaret; Polsky, Phil; Boyt, Eric; Parks, Joe; 'Melanie Gray (E-mail)' Subject: RE: Letter re Unpaid Invoice for Post petition Deliveries Please work wit
h Energy Operations to determine where it went and if we sole it or if is still on the pipe. The priority should be based on your judgment. Ed -----Orig
inal Message----- From: Germany, Chris Sent: Monday, March 11, 2002 4:38 PM To: McMichael Jr., Ed Cc: Mann, Kay; McMichael Jr., Ed; Dicarlo, Louis; Dhon
t, Margaret; Polsky, Phil; Boyt, Eric; Parks, Joe; Melanie Gray (E-mail) Subject: RE: Letter re Unpaid Invoice for Post petition Deliveries Ed, this gas
was nominated on an IT contract into a pool on NGPL. I am unable to identify the gas after that so its difficult for me to tell if the estate benefited.
Let me know what you want to do at this point. Thanks -----Original Message----- From: melanie.gray@weil.com@ENRON Sent: Wednesday, March 06, 2002 3:30
PM To: Germany, Chris Cc: Mann, Kay; McMichael Jr., Ed; Dicarlo, Louis; Dhont, Margaret; Polsky, Phil; Boyt, Eric; Parks, Joe Subject: RE: Letter re Unp
aid Invoice for Post petition Deliveries Assuming all of what is below is all correct, it appears that TDC is entitled to payment. The question with adm
inistrative priority is whether the the estate benefitted. If we took the gas, I assume that we did. THanks. "Germany, Chris" <Chris.Germany@ENRON.com>
on 03/06/2002 02:38:33 PM cc: "Dicarlo, Louis" <Louis.Dicarlo@ENRON.com>, "Dhont, Margaret" <Margaret.Dhont@ENRON.com>, "Polsky, Phil" <Philip.Polsky@EN
RON.com>, "Boyt, Eric" <Eric.Boyt@ENRON.com>, "Parks, Joe"
Metadata: {'from': 'Germany, Chris CGERMAN', 'subject': 'RE: Letter re Unpaid Invoice for Post petition Deliveries', 'thread_id': 2133, 'timestamp': '20
02-03-29 12:49:39'}

--- Chunk 2 ---
Attached are the final documents which includes 1. Agency and Management Agreement - Clean and Redlined copy which shows the changes per my discussion w
ith Jay Goleb at Baker & Botts 2. Gas Confirm for June 1 thru June 4, 2001 3. Gas Confirm for June 5 thru June 30, 2001 4. Gas Conform for July 1 thru O
ctober 31, 2001 5. GTC which applies to all three confirms 6. Agency Notice Letter - Mexicana to send to its Permian suppliers for notice of Enron's age
ncy. Please print out and execute documents 1 thru 5 and fax to Barry Tycholiz. We will follow up with duplicate originals for execution. Thanks for all
your cooperation.
Metadata: {'from': 'Gerald Nemec', 'subject': 'Final Docs.', 'thread_id': 1048, 'timestamp': '2001-06-01 19:22:00'}

--- Chunk 3 ---
Attached are the final documents which includes 1. Agency and Management Agreement - Clean and Redlined copy which shows the changes per my discussion w
ith Jay Goleb at Baker & Botts 2. Gas Confirm for June 1 thru June 4, 2001 3. Gas Confirm for June 5 thru June 30, 2001 4. Gas Conform for July 1 thru O
ctober 31, 2001 5. GTC which applies to all three confirms 6. Agency Notice Letter - Mexicana to send to its Permian suppliers for notice of Enron's age
ncy. Please print out and execute documents 1 thru 5 and fax to Barry Tycholiz. We will follow up with duplicate originals for execution. Thanks for all
your cooperation.
Metadata: {'from': 'Gerald Nemec', 'subject': 'Final Docs.', 'thread_id': 1048, 'timestamp': '2001-06-01 09:22:00'}
```

Query3 – Final Answer



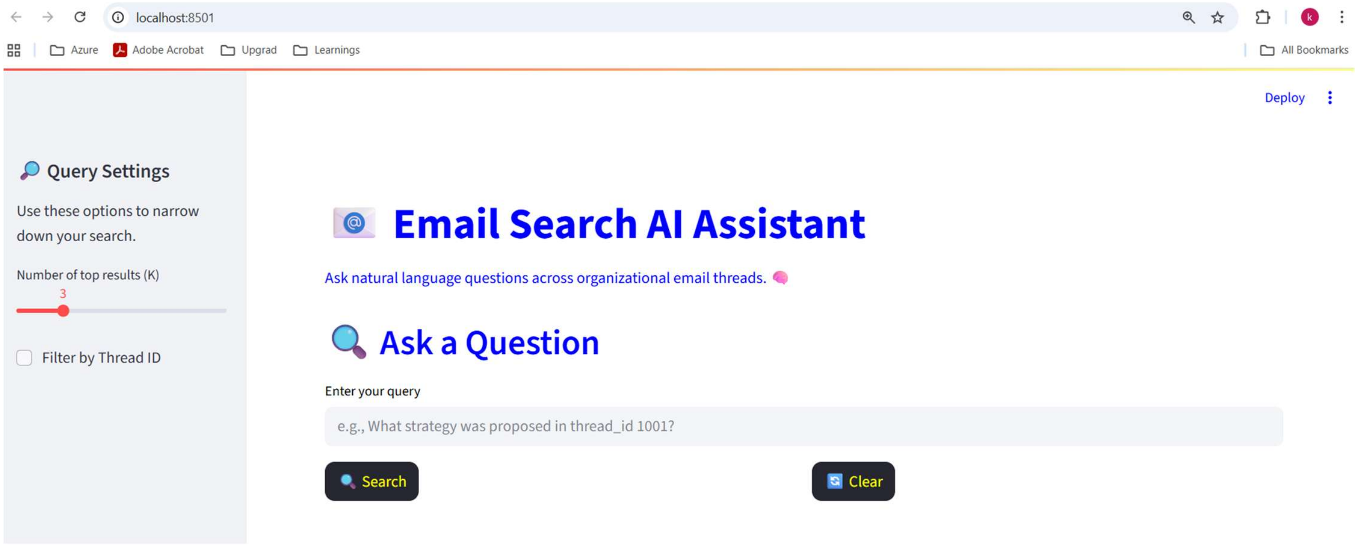
Query3 : Who generated and sent the manual invoice to Southwest Gas, and when?

```
# Generatove Layer Outputs
open_ai_output3 = generate_answer(query3, top_chunks3)
print("\nGenerated Layer Output:")
print(open_ai_output3)
```

```
Generated Layer Output:
The manual invoice to Southwest Gas was generated and sent by Chris Germany. The specific date is not mentioned in the email thread.
```

Stream lint UI Screenshots:

Home Page:



Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

☐ Filter by Thread ID

Email Search AI Assistant

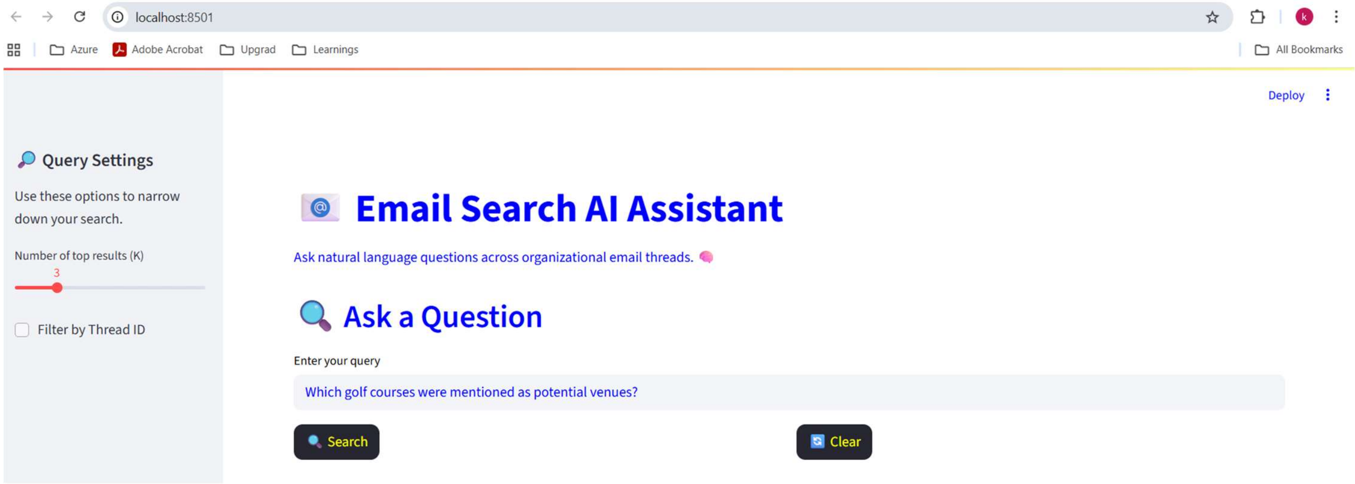
Ask natural language questions across organizational email threads.

Ask a Question

Enter your query

e.g., What strategy was proposed in thread_id 1001?

Ask Question:



Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

☐ Filter by Thread ID

Email Search AI Assistant

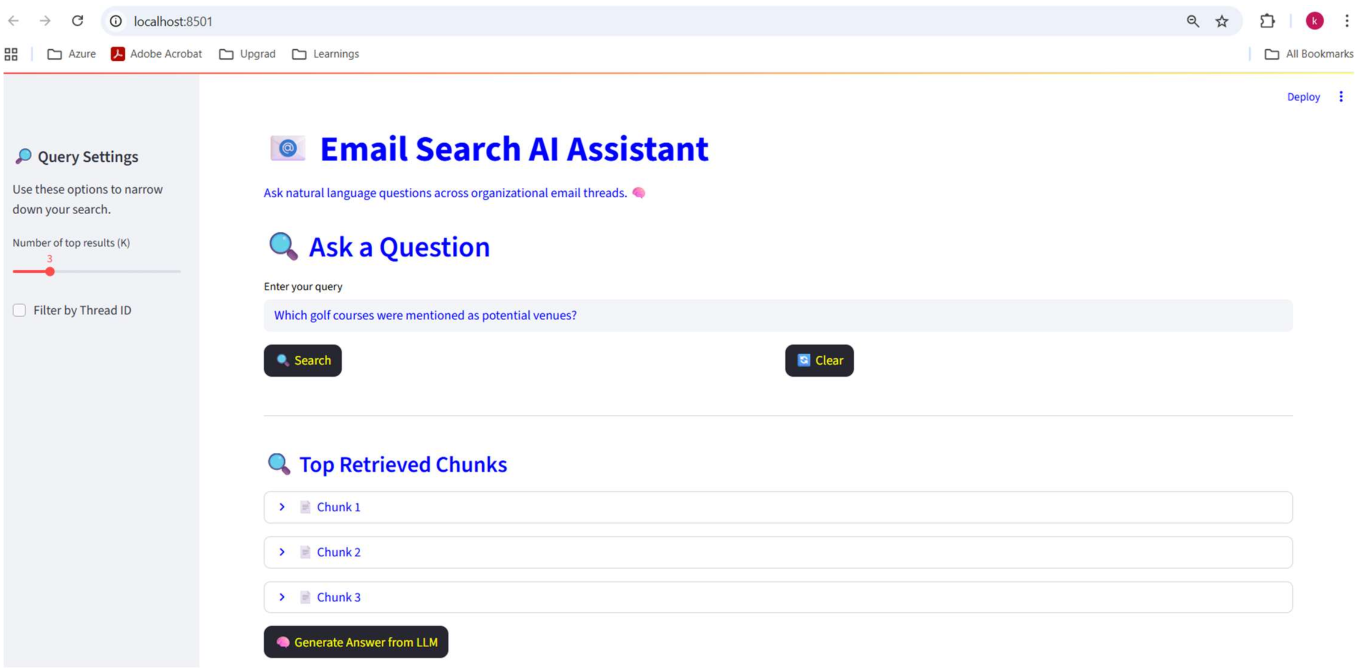
Ask natural language questions across organizational email threads.

Ask a Question

Enter your query

Which golf courses were mentioned as potential venues?

Search results with Top 3 chunks:



Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

□ Filter by Thread ID

Email Search AI Assistant

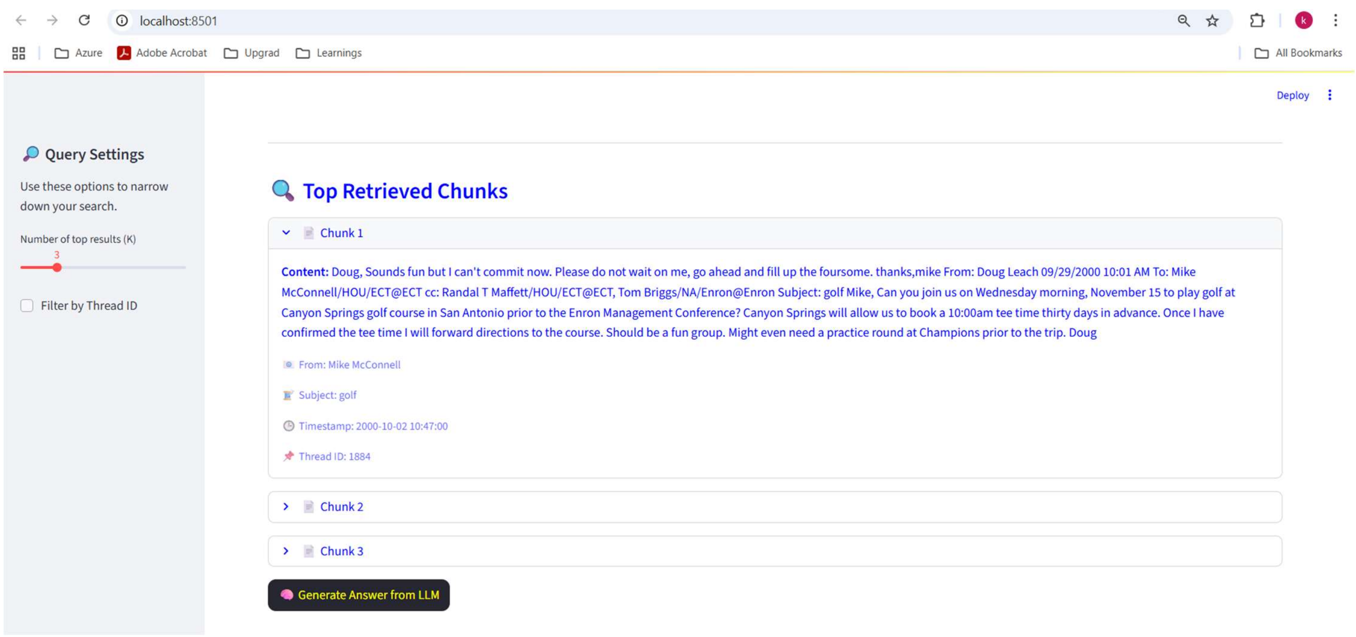
Ask natural language questions across organizational email threads.

Ask a Question

Enter your query: Which golf courses were mentioned as potential venues?

Top Retrieved Chunks

Top1 Chunk Output:



Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

□ Filter by Thread ID

Top Retrieved Chunks

Chunk 1

Content: Doug, Sounds fun but I can't commit now. Please do not wait on me, go ahead and fill up the foursome. thanks,mike From: Doug Leach 09/29/2000 10:01 AM To: Mike McConnell/HOU/ECT@ECT cc: Randal T Maffett/HOU/ECT@ECT, Tom Briggs/NA/Enron@Enron Subject: golf Mike, Can you join us on Wednesday morning, November 15 to play golf at Canyon Springs golf course in San Antonio prior to the Enron Management Conference? Canyon Springs will allow us to book a 10:00am tee time thirty days in advance. Once I have confirmed the tee time I will forward directions to the course. Should be a fun group. Might even need a practice round at Champions prior to the trip. Doug

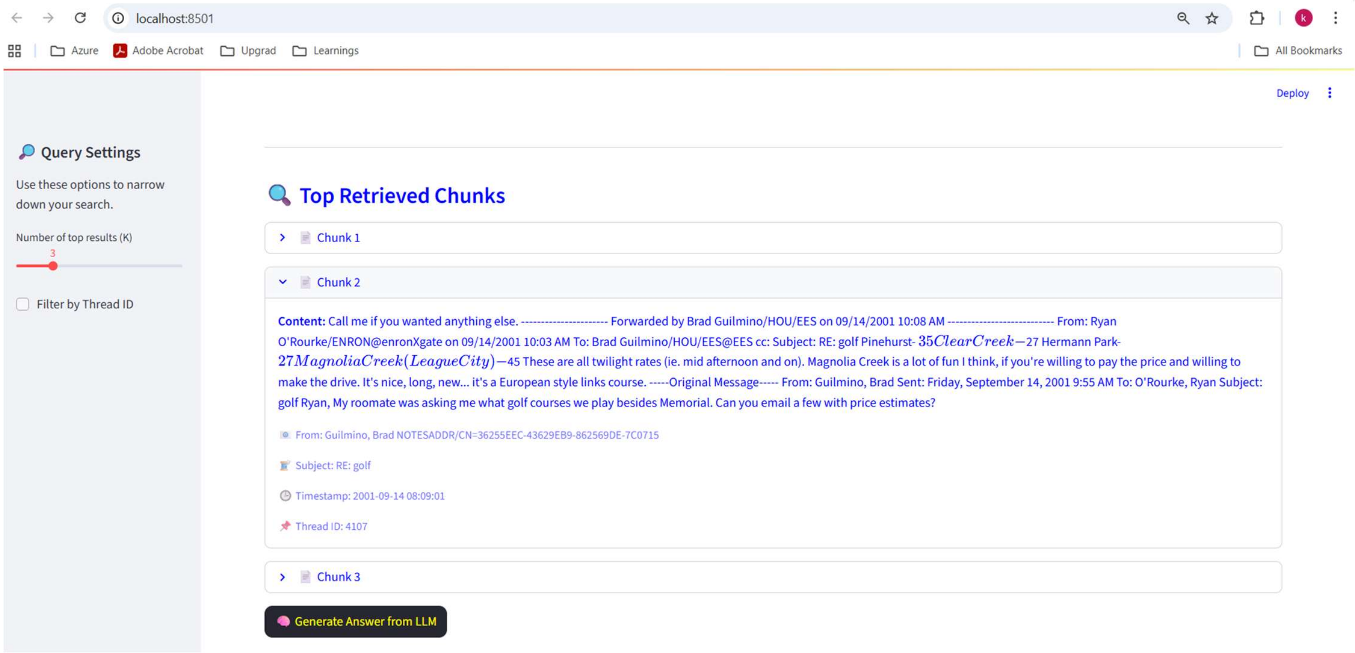
From: Mike McConnell

Subject: golf

Timestamp: 2000-10-02 10:47:00

Thread ID: 1884

Top2 Chunk Output:



🔎 Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

□ Filter by Thread ID

🔍 Top Retrieved Chunks

› Chunk 1

⌄ Chunk 2

Content: Call me if you wanted anything else. --------------------- Forwarded by Brad Guilmino/HOU/EES on 09/14/2001 10:08 AM --------------------------- From: Ryan O'Rourke/ENRON@enronXgate on 09/14/2001 10:03 AM To: Brad Guilmino/HOU/EES@EES cc: Subject: RE: golf Pinehurst- $35ClearCreek—$27 Hermann Park- $27MagnoliaCreek(LeagueCity)—$45 These are all twilight rates (ie. mid afternoon and on). Magnolia Creek is a lot of fun I think, if you're willing to pay the price and willing to make the drive. It's nice, long, new... it's a European style links course. -----Original Message----- From: Guilmino, Brad Sent: Friday, September 14, 2001 9:55 AM To: O'Rourke, Ryan Subject: golf Ryan, My roomate was asking me what golf courses we play besides Memorial. Can you email a few with price estimates?

From: Guilmino, Brad NOTESADDR/CN=36255EEC-43629EB9-862569DE-7C0715

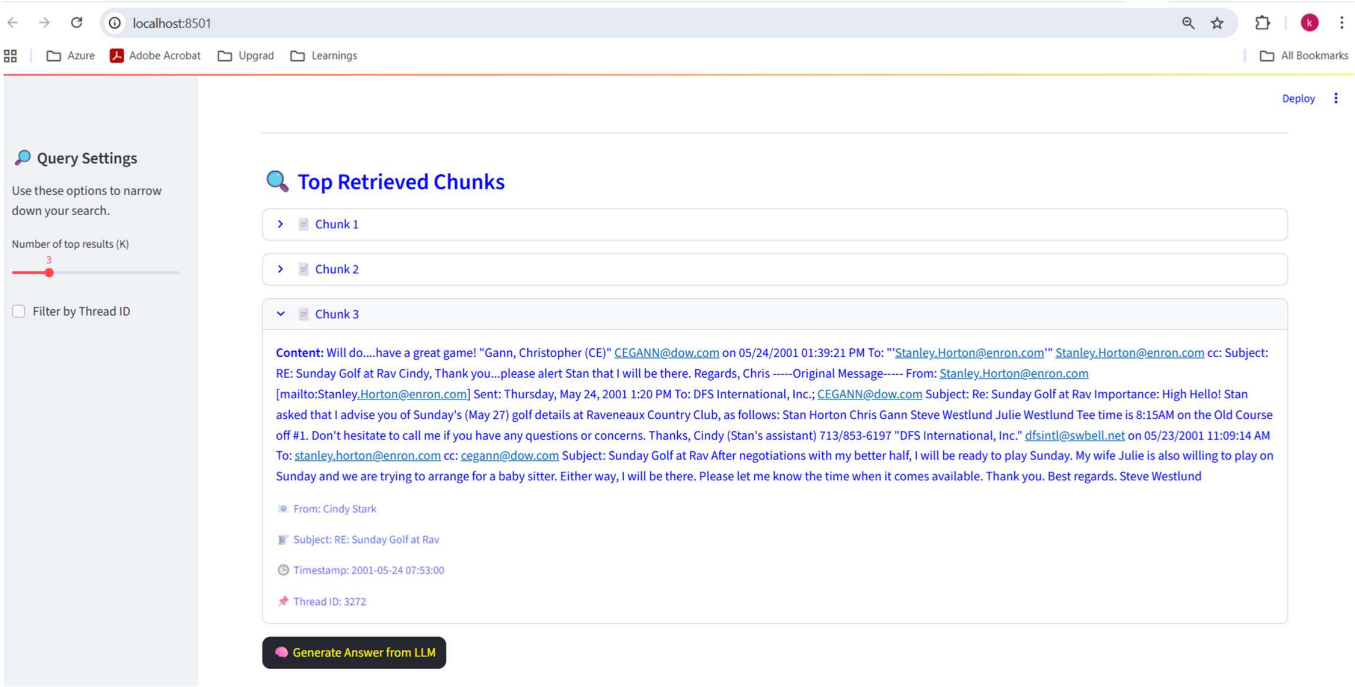
Subject: RE: golf

Timestamp: 2001-09-14 08:09:01

Thread ID: 4107

› Chunk 3

Top3 Chunk Output:



🔎 Query Settings

Use these options to narrow down your search.

Number of top results (K): 3

□ Filter by Thread ID

🔍 Top Retrieved Chunks

› Chunk 1

› Chunk 2

⌄ Chunk 3

Content: Will do....have a great game! "Gann, Christopher (CE)" CEGANN@dow.com on 05/24/2001 01:39:21 PM To: "'Stanley.Horton@enron.com'" Stanley.Horton@enron.com cc: Subject: RE: Sunday Golf at Rav Cindy, Thank you...please alert Stan that I will be there. Regards, Chris -----Original Message----- From: Stanley.Horton@enron.com [mailto:Stanley.Horton@enron.com] Sent: Thursday, May 24, 2001 1:20 PM To: DFS International, Inc.; CEGANN@dow.com Subject: Re: Sunday Golf at Rav Importance: High Hello! Stan asked that I advise you of Sunday's (May 27) golf details at Raveneaux Country Club, as follows: Stan Horton Chris Gann Steve Westlund Julie Westlund Tee time is 8:15AM on the Old Course off #1. Don't hesitate to call me if you have any questions or concerns. Thanks, Cindy (Stan's assistant) 713/853-6197 "DFS International, Inc." dfsintl@swbell.net on 05/23/2001 11:09:14 AM To: stanley.horton@enron.com cc: cegann@dow.com Subject: Sunday Golf at Rav After negotiations with my better half, I will be ready to play Sunday. My wife Julie is also willing to play on Sunday and we are trying to arrange for a baby sitter. Either way, I will be there. Please let me know the time when it comes available. Thank you. Best regards. Steve Westlund

From: Cindy Stark

Subject: RE: Sunday Golf at Rav

Timestamp: 2001-05-24 07:53:00

Thread ID: 3272

Future Enhancements

Embedding Layer Enhancements:

- Parallelize or batch chunking and embedding for large datasets.
- Support multilingual email embedding using a multilingual transformer model (e.g., distiluse-base-multilingual-cased).
- Add logging and error handling during embedding and chunking.
- Deduplicate similar chunks before storing in the vector DB to reduce redundancy.
- Store additional metadata (e.g., department, priority) to enable advanced filtering during search.

Cahce Layer Enhancements:

- Replace JSON with Redis or SQLite for faster lookup and persistence in multi-user environments.
- Add cache eviction policy (e.g., LRU) to avoid unlimited growth.
- Track cache hit/miss stats for performance analytics.
- Encrypt cache contents if storing sensitive queries or responses.

Search Layer Enhancements:

- Improve reranking with better cross-encoders like bge-reranker or cohere models.
- Add semantic filters beyond thread_id (e.g., date, sender, topic).
- Support multi-query or follow-up query handling (e.g., thread-based QA).
- Paginate results and allow sorting based on relevance, timestamp, etc.
- Expose the search as an API with configurable parameters.

Generation Layer Enhancements:

- Use function calling / structured output instead of plain text (for automation).
- Support custom prompt templates per use case (summarization, classification, etc.).
- Switch to a self-hosted model (e.g., LLaMA 3, Mistral) for cost and privacy control.
- Limit token count dynamically to avoid truncation of large prompts.

Overall Architecture Enhancements

- Centralized logging and monitoring (e.g., using logging, Sentry, or Prometheus).
- Unit and integration tests for all layers to ensure robustness.
- Add retry mechanisms for external API calls (OpenAI, Chroma).
- Implement role-based access control (RBAC) if deployed in an enterprise environment.
- Deploy as a containerized microservice (Docker + FastAPI) with endpoints for embedding, search, and generation.