GenAI-Powered Data Pipeline Project Report

Date: March 31, 2025

Student: Vignesh Jetty Ravi(Assumed from screenshots)

Student ID: 018286970

1. Project Overview

This project focuses on building a **cloud-based GenAI-powered data pipeline** using **Langflow** and **AstraDB**.

Key Components:

- **Data Ingestion**: Unstructured FAQ dataset is processed and ingested.
- **Vectorization**: The dataset is transformed using **OpenAI embeddings**.
- **Retrieval-Augmented Generation (RAG)**: The system enables intelligent query responses by leveraging vectorized data.

This architecture enhances the ability to provide accurate, context-aware answers based on the ingested FAQs.

2. Architecture Overview

The system is structured into two primary workflows in **Langflow**:

- **Load Data Flow**:
    - Ingests CSV data
    - Splits text into manageable chunks
    - Generates embeddings using **OpenAI**
    - Stores the vectorized data in **AstraDB**
- **Retriever Flow**:
    - Processes user queries

- Embeds the input question
- Retrieves relevant text chunks from **AstraDB**
- Constructs a prompt and generates a response using **GPT-3.5**

This streamlined approach ensures efficient data handling and accurate retrieval-based answering.

3. Tools & Technologies

- **Langflow (Cloud Version)** – Orchestrates the data pipeline
- **AstraDB** – Stores vectorized data with **vector search enabled**
- **OpenAI Embeddings** – Utilizes **text-embedding-3-small** for vectorization
- **OpenAI GPT-3.5 Turbo** – Powers the Retrieval-Augmented Generation (RAG) process
- **Dataset** – **faq_bulk_150_records.csv**, an unstructured FAQ dataset

4. Key Design Decisions

- **Text Chunking**: Set to **100 tokens** with a **20-token overlap** to enhance context retrieval.
- **Embedding Model**: Used **`text-embedding-3-small`** for faster and efficient vectorization.
- **Retrieval Optimization**: Limited results to **5 chunks** to minimize LLM context size and improve response speed.
- **LLM Selection**: Chose **GPT-3.5 and 4o mini** for its balance of speed and response quality.
- **Dynamic Prompting**: Implemented **prompt templating** with variables for optimized **RAG-based queries**.

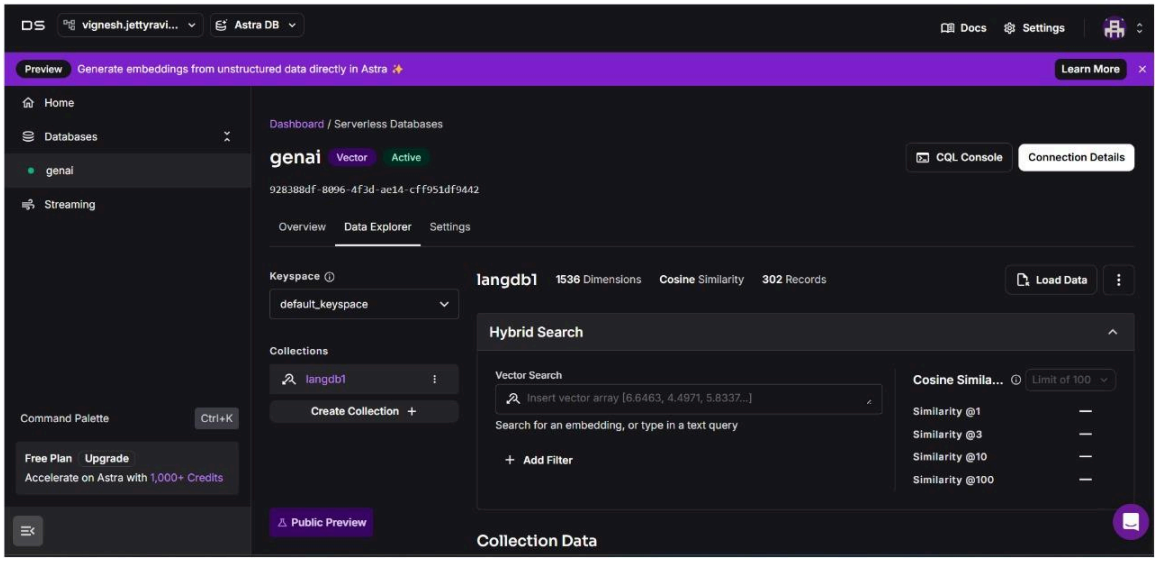
5. Challenges & Solutions

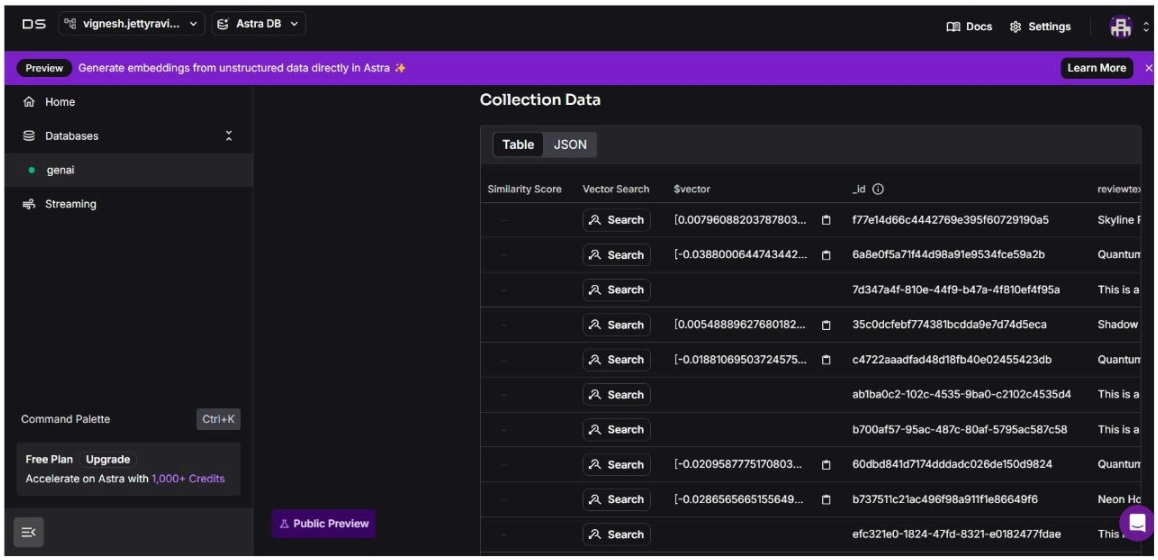
-ERROR building Component Astra DB: Error adding documents to AstraDBVectorStore: Length of vector parameter different from declared '$vector' dimension: root cause = (InvalidQueryException) Failed to insert document with _id 979313aa79e74367a333c98102224431: Unexpected 2048 extraneous bytes after vector<float, 1024> . I had to adjust the vector dimension from 1024 to 1536.

-ERROR during langflow run: solved via careful step-by-step testing and visual flow debugging and refreshing the page for re-run.

6. Results & Screenshots

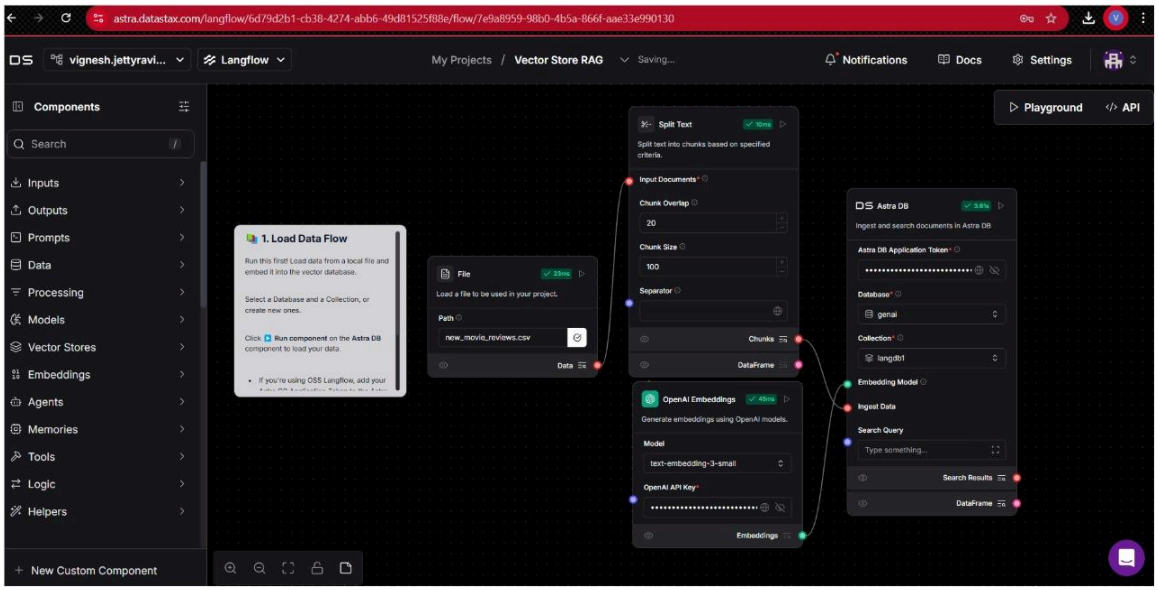
Astra DB



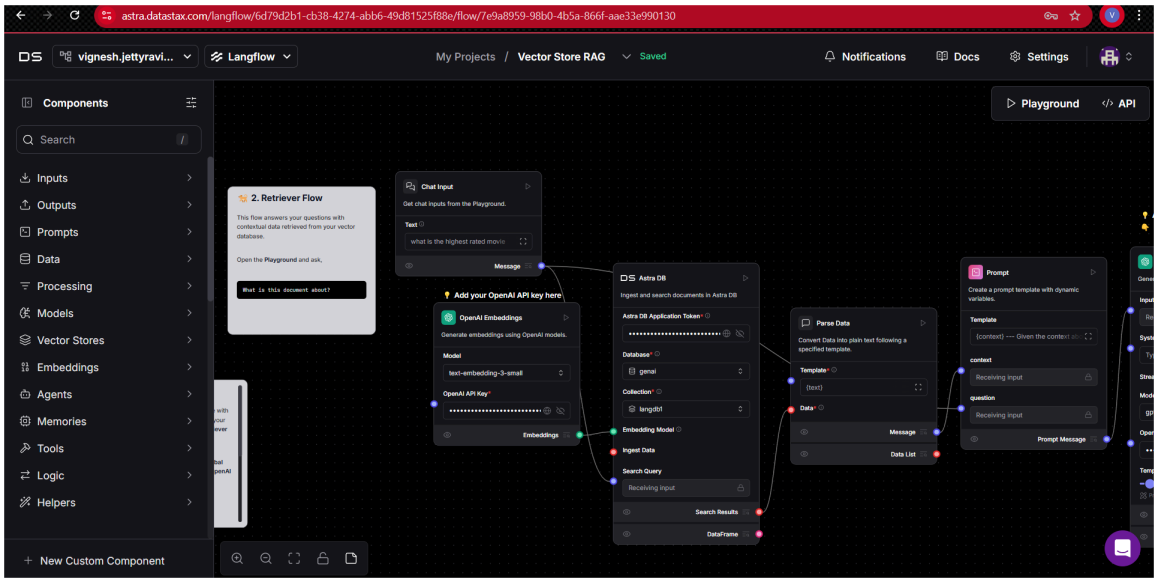


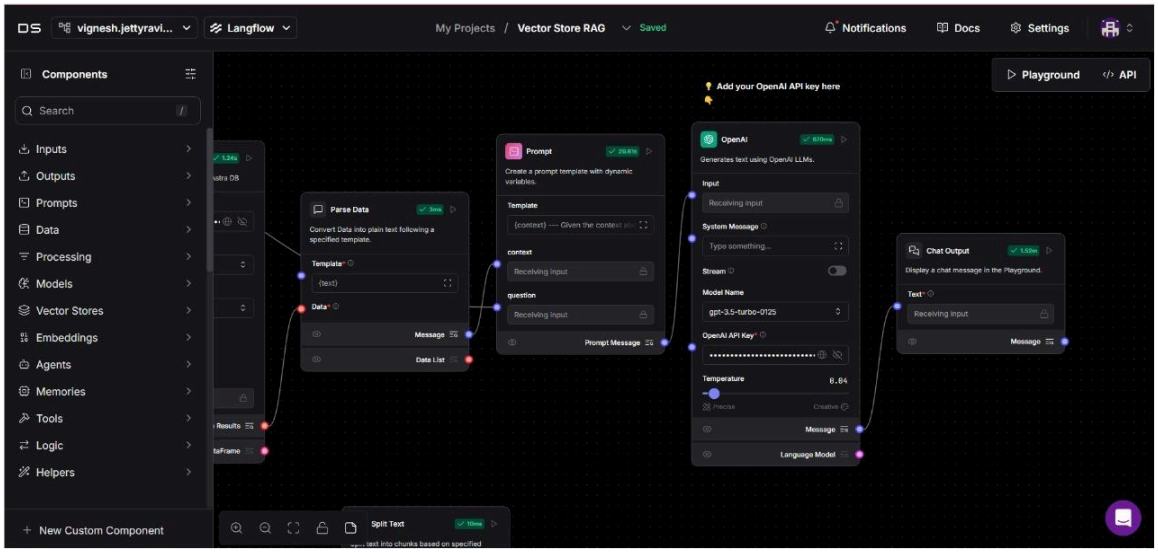
Langflow

1.load data flow

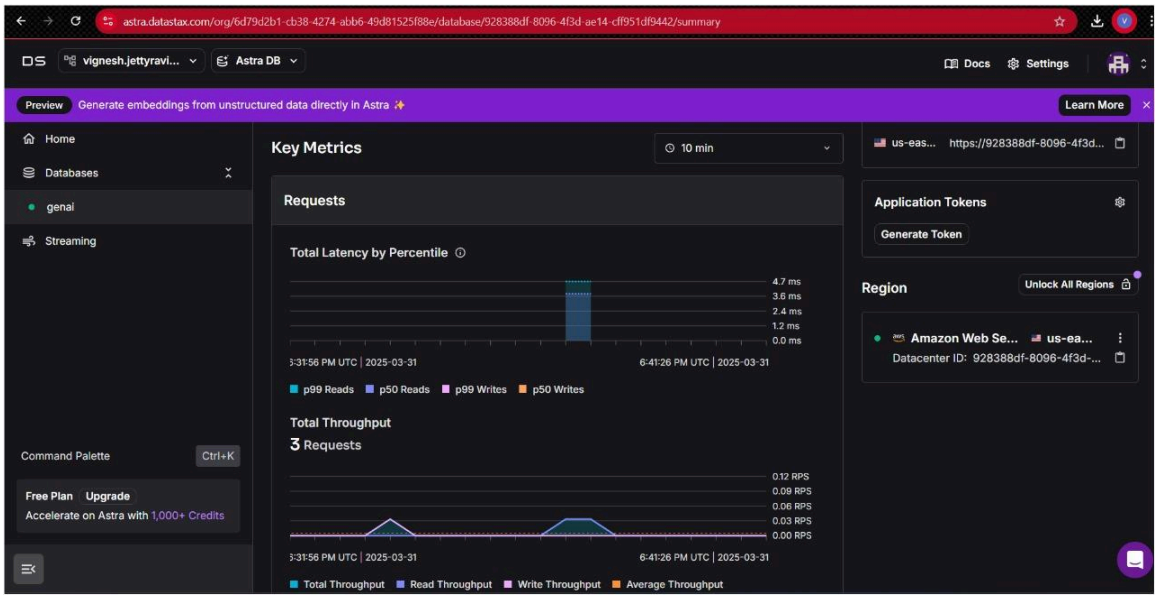


2. retriever flow

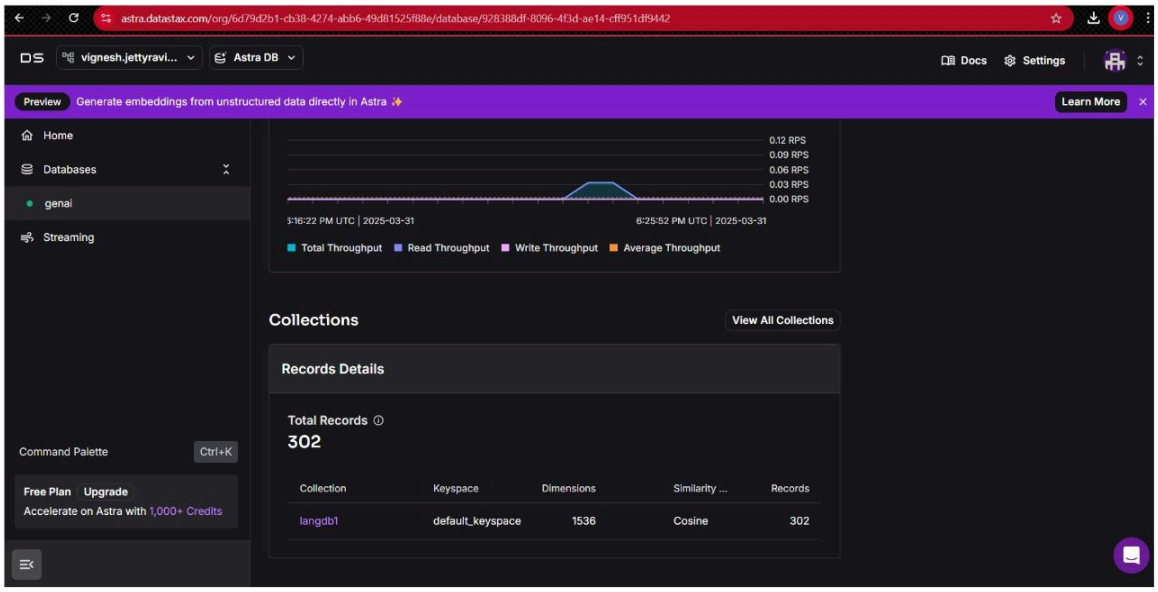




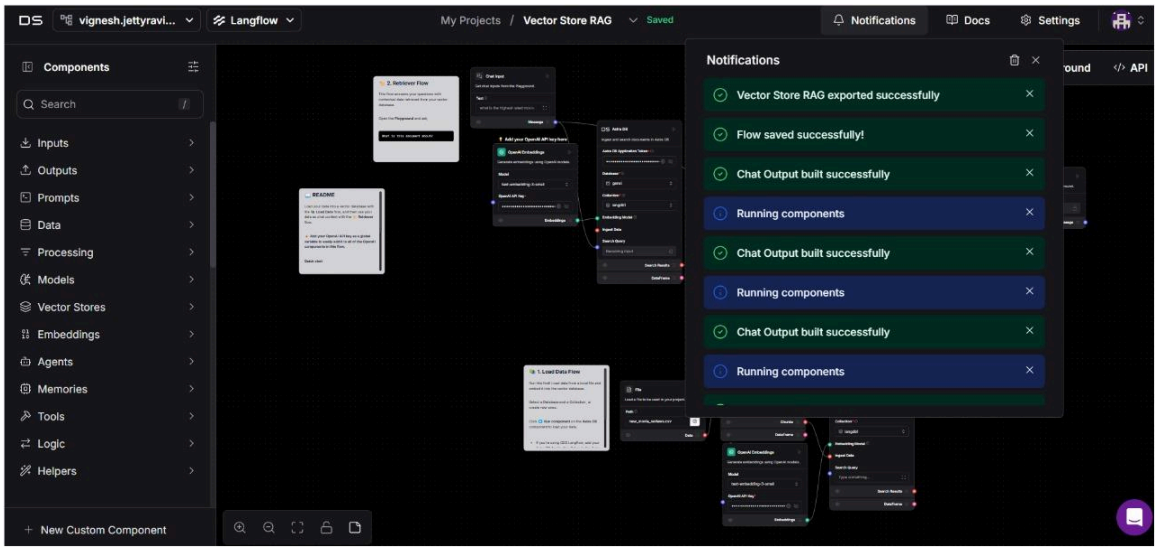
3.key metrics



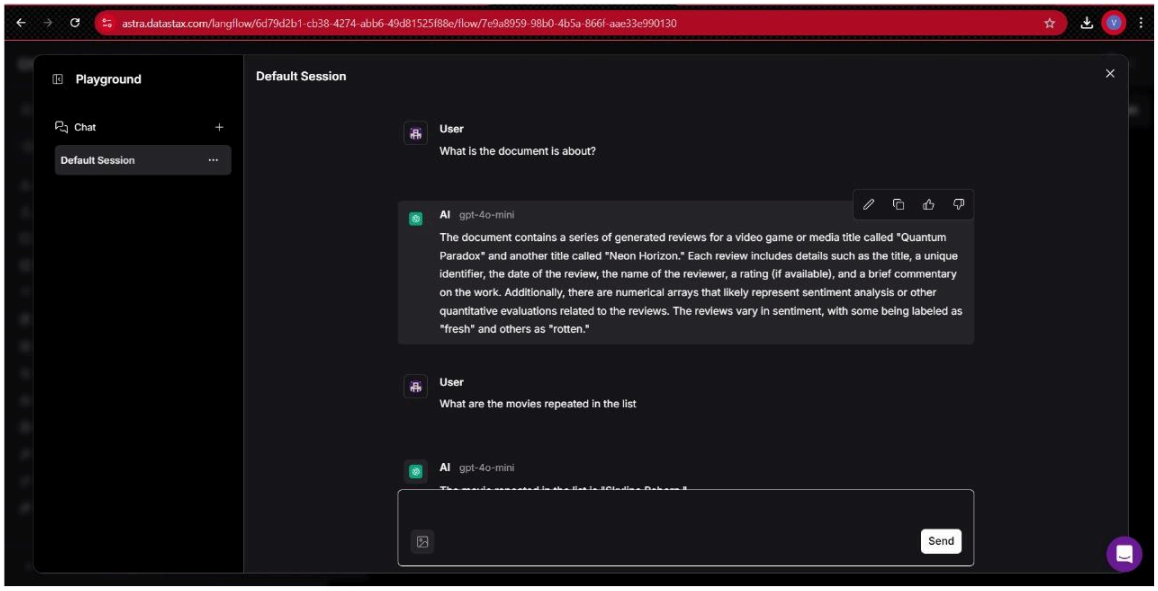
4. collections

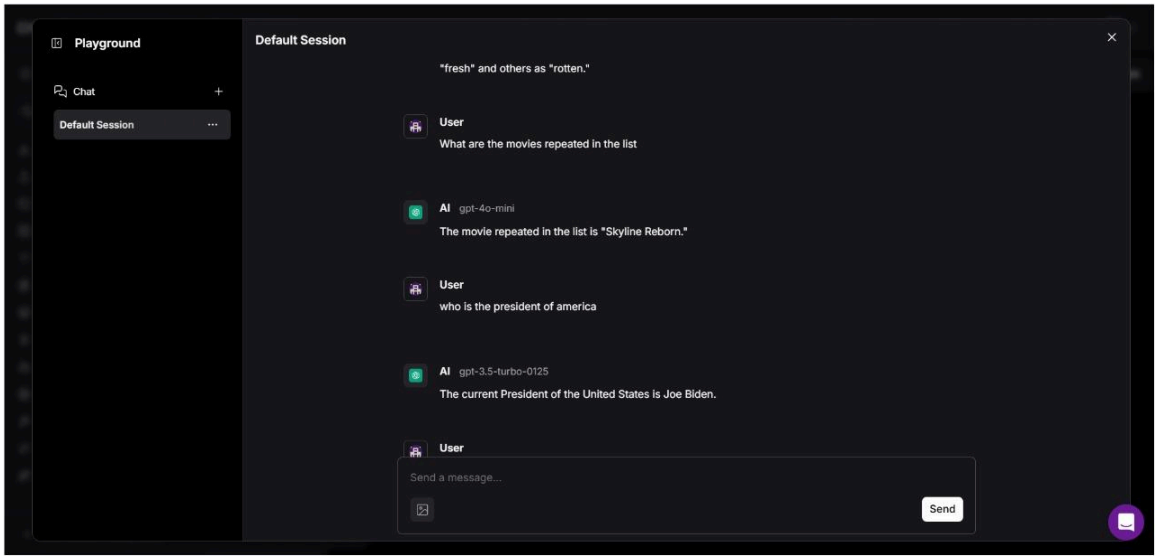


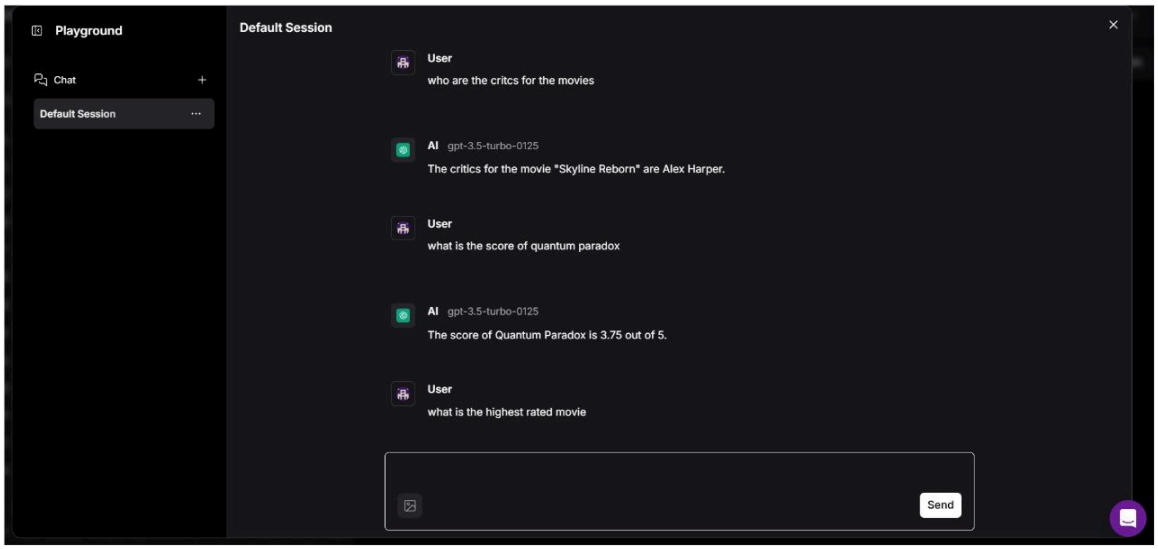
5. notifications of completed prompts

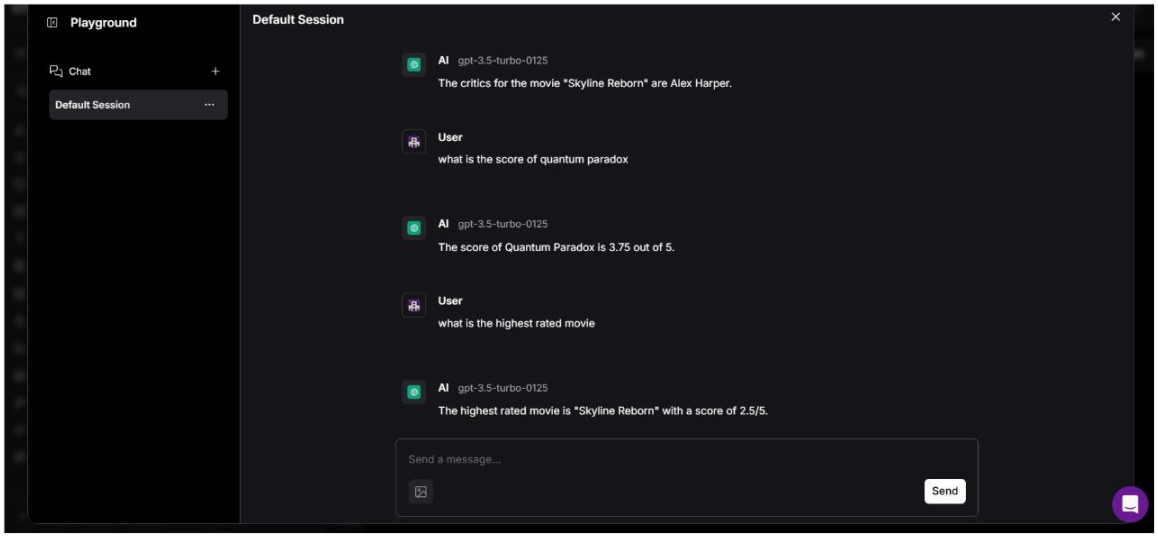


6.prompts and session results

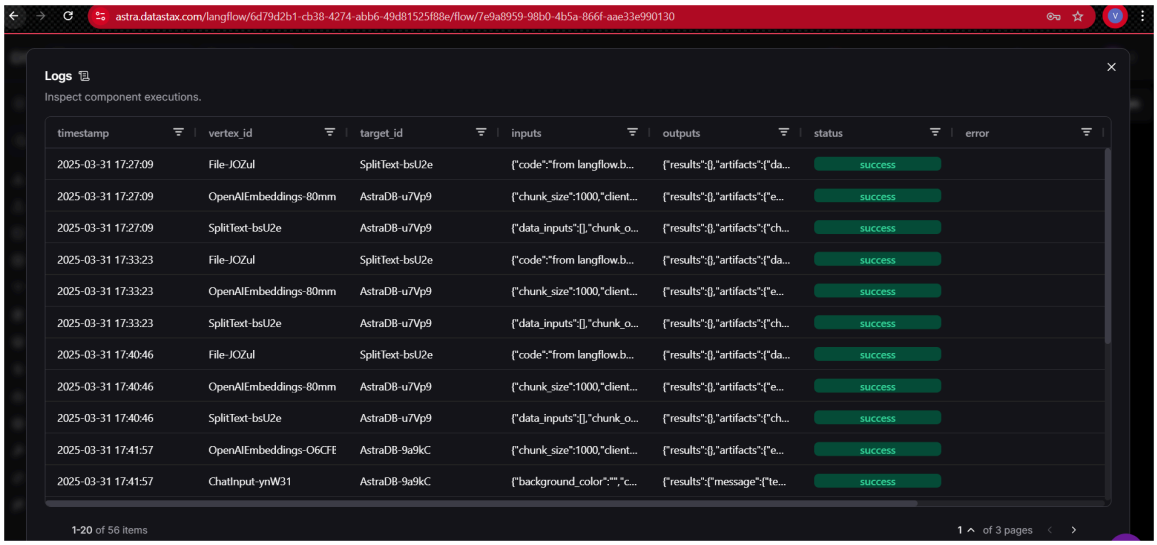

astra.datastax.com/langflow/6d79d2b1-cb38-4274-abb6-49d81525f88e/flow/7e9a8959-98b0-4b5a-866f-aae33e990130
Playground
Chat
Default Session
Default Session
User
What is the document is about?
AI gpt-4o-mini
The document contains a series of generated reviews for a video game or media title called "Quantum Paradox" and another title called "Neon Horizon." Each review includes details such as the title, a unique identifier, the date of the review, the name of the reviewer, a rating (if available), and a brief commentary on the work. Additionally, there are numerical arrays that likely represent sentiment analysis or other quantitative evaluations related to the reviews. The reviews vary in sentiment, with some being labeled as "fresh" and others as "rotten."
User
What are the movies repeated in the list
AI gpt-4o-mini
Send


Playground
Chat
Default Session
Default Session
"fresh" and others as "rotten."
User
What are the movies repeated in the list
AI gpt-4o-mini
The movie repeated in the list is "Skyline Reborn."
User
who is the president of america
AI gpt-3.5-turbo-0125
The current President of the United States is Joe Biden.
User
Send a message...
Send

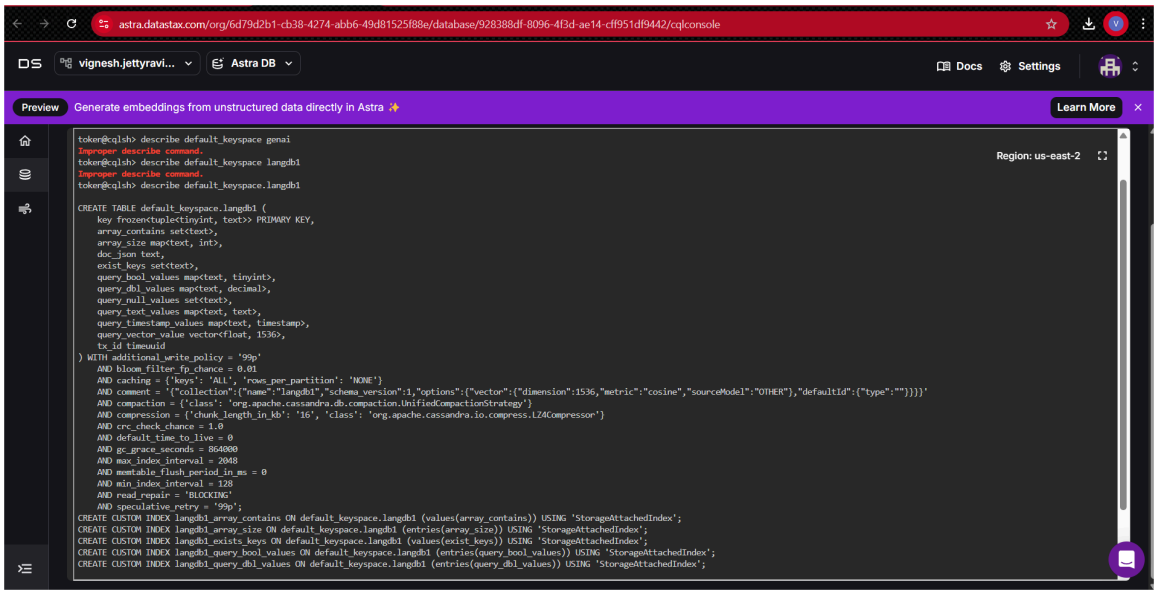

Playground
Chat
Default Session
Default Session
User
who are the critcs for the movies
AI gpt-3.5-turbo-0125
The critics for the movie "Skyline Reborn" are Alex Harper.
User
what is the score of quantum paradox
AI gpt-3.5-turbo-0125
The score of Quantum Paradox is 3.75 out of 5.
User
what is the highest rated movie
Send

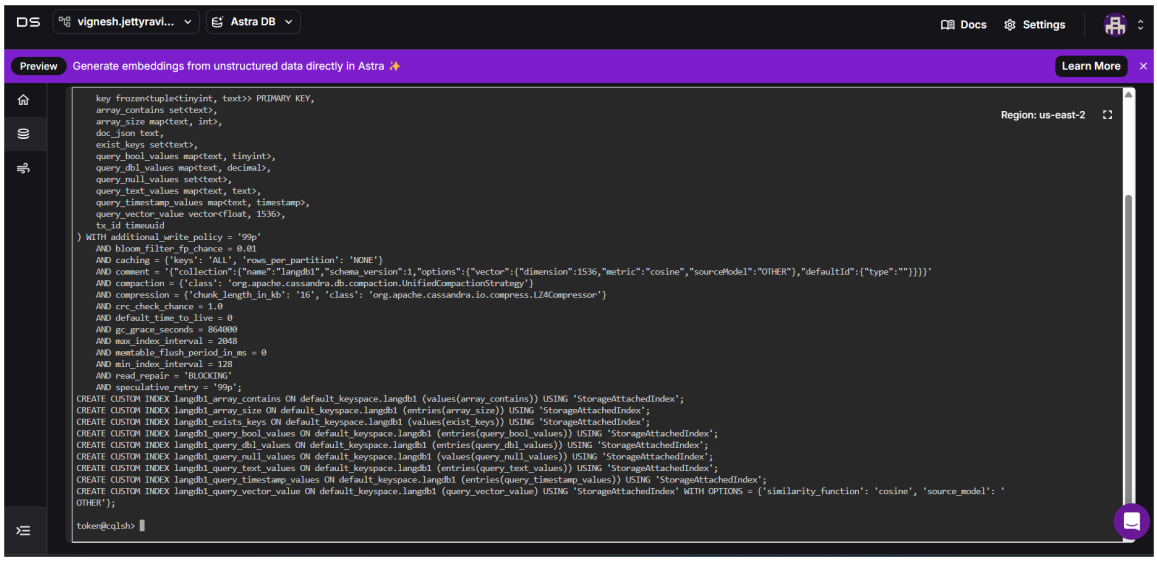

Playground
Chat
Default Session
Default Session
AI gpt-3.5-turbo-0125
The critics for the movie "Skyline Reborn" are Alex Harper.
User
what is the score of quantum paradox
AI gpt-3.5-turbo-0125
The score of Quantum Paradox is 3.75 out of 5.
User
what is the highest rated movie
AI gpt-3.5-turbo-0125
The highest rated movie is "Skyline Reborn" with a score of 2.5/5.
Send a message...
Send

7. Logs


astra.datastax.com/langflow/6d79d2b1-cb38-4274-abb6-49d81525f88e/flow/7e9a8959-98b0-4b5a-866f-aae33e990130
Logs
Inspect component executions.
timestamp | vertex_id | target_id | inputs | outputs | status | error
2025-03-31 17:27:09 | File-JOZul | SplitText-bsU2e | {"code":"from langflow.b... | {"results":{},"artifacts":{"da... | success
2025-03-31 17:27:09 | OpenAIEmbeddings-80mm | AstraDB-u7Vp9 | {"chunk_size":1000,"client... | {"results":{},"artifacts":{"e... | success
2025-03-31 17:27:09 | SplitText-bsU2e | AstraDB-u7Vp9 | {"data_inputs":[],"chunk_o... | {"results":{},"artifacts":{"ch... | success
2025-03-31 17:33:23 | File-JOZul | SplitText-bsU2e | {"code":"from langflow.b... | {"results":{},"artifacts":{"da... | success
2025-03-31 17:33:23 | OpenAIEmbeddings-80mm | AstraDB-u7Vp9 | {"chunk_size":1000,"client... | {"results":{},"artifacts":{"e... | success
2025-03-31 17:33:23 | SplitText-bsU2e | AstraDB-u7Vp9 | {"data_inputs":[],"chunk_o... | {"results":{},"artifacts":{"ch... | success
2025-03-31 17:40:46 | File-JOZul | SplitText-bsU2e | {"code":"from langflow.b... | {"results":{},"artifacts":{"da... | success
2025-03-31 17:40:46 | OpenAIEmbeddings-80mm | AstraDB-u7Vp9 | {"chunk_size":1000,"client... | {"results":{},"artifacts":{"e... | success
2025-03-31 17:40:46 | SplitText-bsU2e | AstraDB-u7Vp9 | {"data_inputs":[],"chunk_o... | {"results":{},"artifacts":{"ch... | success
2025-03-31 17:41:57 | OpenAIEmbeddings-O6CFE | AstraDB-9a9kC | {"chunk_size":1000,"client... | {"results":{},"artifacts":{"e... | success
2025-03-31 17:41:57 | ChatInput-ynW31 | AstraDB-9a9kC | {"background_color":"","c... | {"results":{"message":{"te... | success
1-20 of 56 items
1 of 3 pages

8.dbschema


astra.datastax.com/org/6d79d2b1-cb38-4274-abb6-49d81525f88e/database/928388df-8096-4f3d-ae14-cff951df9442/cqlconsole
vignesh.jettyravi...
Astra DB
Docs
Settings
Preview
Generate embeddings from unstructured data directly in Astra ✨
Learn More
Region: us-east-2
token@cqlsh> describe default_keyspace genai
Improper describe command.
token@cqlsh> describe default_keyspace langdb1
Improper describe command.
token@cqlsh> describe default_keyspace.langdb1
CREATE TABLE default_keyspace.langdb1 (
key frozen<tuple<tinyint, text>> PRIMARY KEY,
array_contains set<text>,
array_size map<text, int>,
doc_json text,
exist_keys set<text>,
query_bool_values map<text, tinyint>,
query_dbl_values map<text, decimal>,
query_null_values set<text>,
query_text_values map<text, text>,
query_timestamp_values map<text, timestamp>,
query_vector_value vector<float, 1536>,
tx_id timeuuid
) WITH additional_write_policy = '99p'
AND bloom_filter_fp_chance = 0.01
AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
AND comment = '{"collection":{"name":"langdb1","schema_version":1,"options":{"vector":{"dimension":1536,"metric":"cosine","sourceModel":"OTHER"},"defaultId":{"type":""}}}}'
AND compaction = {'class': 'org.apache.cassandra.db.compaction.UnifiedCompactionStrategy'}
AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
AND crc_check_chance = 1.0
AND default_time_to_live = 0
AND gc_grace_seconds = 864000
AND max_index_interval = 2048
AND memtable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';
CREATE CUSTOM INDEX langdb1_array_contains ON default_keyspace.langdb1 (values(array_contains)) USING 'StorageAttachedIndex';
CREATE CUSTOM INDEX langdb1_array_size ON default_keyspace.langdb1 (entries(array_size)) USING 'StorageAttachedIndex';
CREATE CUSTOM INDEX langdb1_exists_keys ON default_keyspace.langdb1 (values(exist_keys)) USING 'StorageAttachedIndex';
CREATE CUSTOM INDEX langdb1_query_bool_values ON default_keyspace.langdb1 (entries(query_bool_values)) USING 'StorageAttachedIndex';
CREATE CUSTOM INDEX langdb1_query_dbl_values ON default_keyspace.langdb1 (entries(query_dbl_values)) USING 'StorageAttachedIndex';

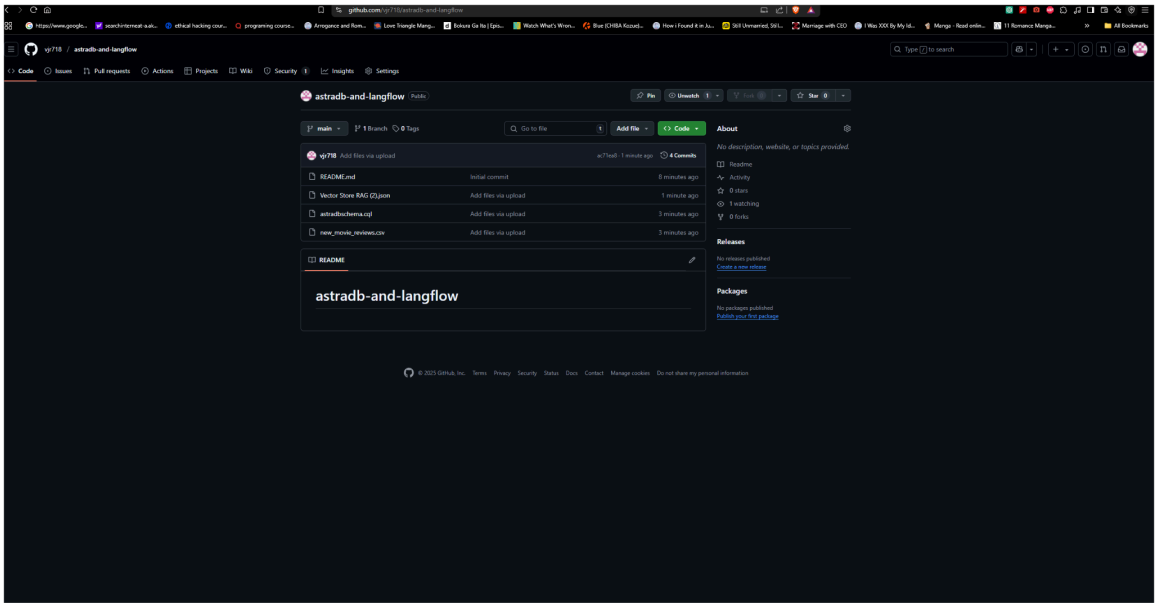


7. Future Enhancements

- **Implement a feedback loop** to refine and improve answers over time.
- **Deploy as a public chatbot** using **Streamlit** or **Gradio** for wider accessibility.
- **Integrate LangSmith** to enhance observability and monitoring.
- **Leverage a fine-tuned LLM** for improved domain-specific responses.
- **Explore Hugging Face open-source LLMs** as a cost-effective alternative.

Github repository for the assignment

https://github.com/vjr718/astradb-and-langflow



->Due to the recent update we can’t share the deployment model of langflow based genai pipeline.

so I have given the link here
https://astra.datastax.com/org/6d79d2b1-cb38-4274-abb6-49d81525f88e/database/928388df-8096-4f3d-ae14-cff951df9442/data-explorer

https://astra.datastax.com/langflow/6d79d2b1-cb38-4274-abb6-49d81525f88e/flow/7e9a8959-98b0-4b5a-866f-aae33e990130

——--END——--