🎬 CineBook – RAG-based Conversational AI Agent and Movie Support System

Project Concept

CineBook is an AI-powered customer support system designed for a movie ticket booking platform. It acts as a conversational AI agent that can:

- Answer questions about movies, theaters, and show timings
- Provide information about refunds, privacy, and booking policies
- Interact naturally with users through a chat interface

Unlike traditional chatbots that rely only on predefined rules, CineBook uses Retrieval-Augmented Generation (RAG) a method that combines data retrieval with LLM reasoning — to generate accurate, context-aware answer

Methodology

The methodology of the CineBook – RAG-based Conversational AI Agent focuses on integrating both structured data retrieval and unstructured semantic reasoning to enable intelligent, context-aware communication. The system's foundation lies in the Retrieval-Augmented Generation (RAG) architecture, which bridges deterministic database access with dynamic language model reasoning.

This hybrid approach ensures that every response is both accurate and contextually relevant, maintaining a natural conversational flow.

1) **User Query and Intent Analysis**
   - The process begins when the user submits a message through the Streamlit Chat UI.
   - The AI Agent receives this input and analyzes it to determine intent type — either:
   - Structured Query: factual questions such as show timings or theater names.
   - Unstructured Query: reasoning-based questions about refunds, policies, or account-related topics.
   - This classification helps CineBook decide whether to perform a data lookup or initiate the RAG pipeline.
2) **Handling Structured Queries (Direct Lookup)**
   - For factual queries, CineBook uses direct data lookup instead of relying on an LLM.
   - The structured datasets — movies.json, theaters.json, and shows.json — contain all information related to:
     1. Movie titles, genres, and languages
     2. Theater names and locations
     3. Show timings and IDs
   - During preprocessing, these files are merged into a lookup table that maps movies to their corresponding theaters and timings. The AI Agent performs instant filtering and mapping operations to fetch results.

Example:

If a user asks “Which movies are showing at StarMax Cinemas?”, CineBook searches the lookup table and returns:

> “StarMax Cinemas (Velachery, Chennai): ‘Vikram’ (Tamil, Thriller) – 11:30 AM.”

This approach ensures zero hallucination and maximum accuracy for factual data retrieval.

3) Handling Unstructured Queries (RAG Pipeline)

- When a user asks reasoning-based questions — such as refunds, privacy, or user guidelines — CineBook activates the RAG (Retrieval-Augmented Generation) pipeline.
- The query text is converted into a vector embedding using SentenceTransformer (MiniLM-L6-v2), which captures its semantic meaning.
- This embedding is sent to the ChromaDB vector database, which stores precomputed embeddings from unstructured policy documents such as:
    1. refund_policy.json
    2. account_privacy_policy.json
    3. user_guidelines.json
- ChromaDB performs a similarity search using cosine distance to find the most contextually related text vectors.
- The top k (e.g., 5) relevant snippets are retrieved and used as context for response generation.

Example:

If the user asks “Can I get a refund if I miss my show?”, CineBook retrieves the rule “No refunds are applicable for missed shows” from the refund policy.

4) Response Generation Using LLM

- Once the relevant policy snippets are retrieved, the system constructs a context prompt that includes both:
    1. The user’s original query
    2. The retrieved knowledge base text
- This prompt is sent to the Groq Llama 3.1 8B model, which acts as the reasoning and generation engine.
- The LLM processes both inputs and produces a coherent, policy-grounded response, ensuring factual accuracy and fluency.

Example Response:

“You’re eligible for a refund if the cancellation is made at least two hours before showtime. Refunds are not applicable beyond that period.”

This mechanism ensures that the LLM doesn’t invent answers — it reasons strictly within the provided company policies.

5) Context Management and Session State

- CineBook maintains conversation continuity using st.session_state in Streamlit.
- Every user query and AI response are stored in session memory.
- This allows multi-turn conversations, enabling the system to understand follow-up questions like:

"What if I cancel only one hour before?"

This gives the user a smooth, human-like conversational experience.

6) Controlled Environment Setup

- All processes run in a Python virtual environment (venv) to ensure isolation and reproducibility.
- Required dependencies include:

  streamlit, chromadb, sentence-transformers, groq, and python-dotenv.

- The .env file securely stores the Groq API key, keeping credentials safe and out of public code repositories.

7) Data Preparation and Embedding Creation

- Before runtime, the policy JSON files are preprocessed and chunked into smaller text segments suitable for embedding.
- Each paragraph or rule (e.g., refund conditions or data retention policies) is embedded using SentenceTransformer and stored in ChromaDB with unique metadata.
- Similarly, structured movie and theater data are processed into key-value mappings to ensure efficient filtering during lookup operations.
- This preprocessing enables instant data retrieval and faster RAG performance during query execution.

8) Decision Engine and Query Routing

The AI Agent's modular decision engine orchestrates all backend operations.
When a new query arrives:

- It first passes through the intent classifier (based on Python keyword detection).
- Depending on intent, it either:
- Triggers structured functions like list_all_shows()
- Or initiates the RAG retrieval process for unstructured queries
- The final response — either lookup-based or LLM-generated — is returned to the Streamlit UI for display.

9) Complete Workflow Summary

The entire CineBook workflow can be summarized as:

User Input → Intent Analysis → Route Selection → Retrieval (Lookup / RAG) → Response Generation → Display in Chat UI.

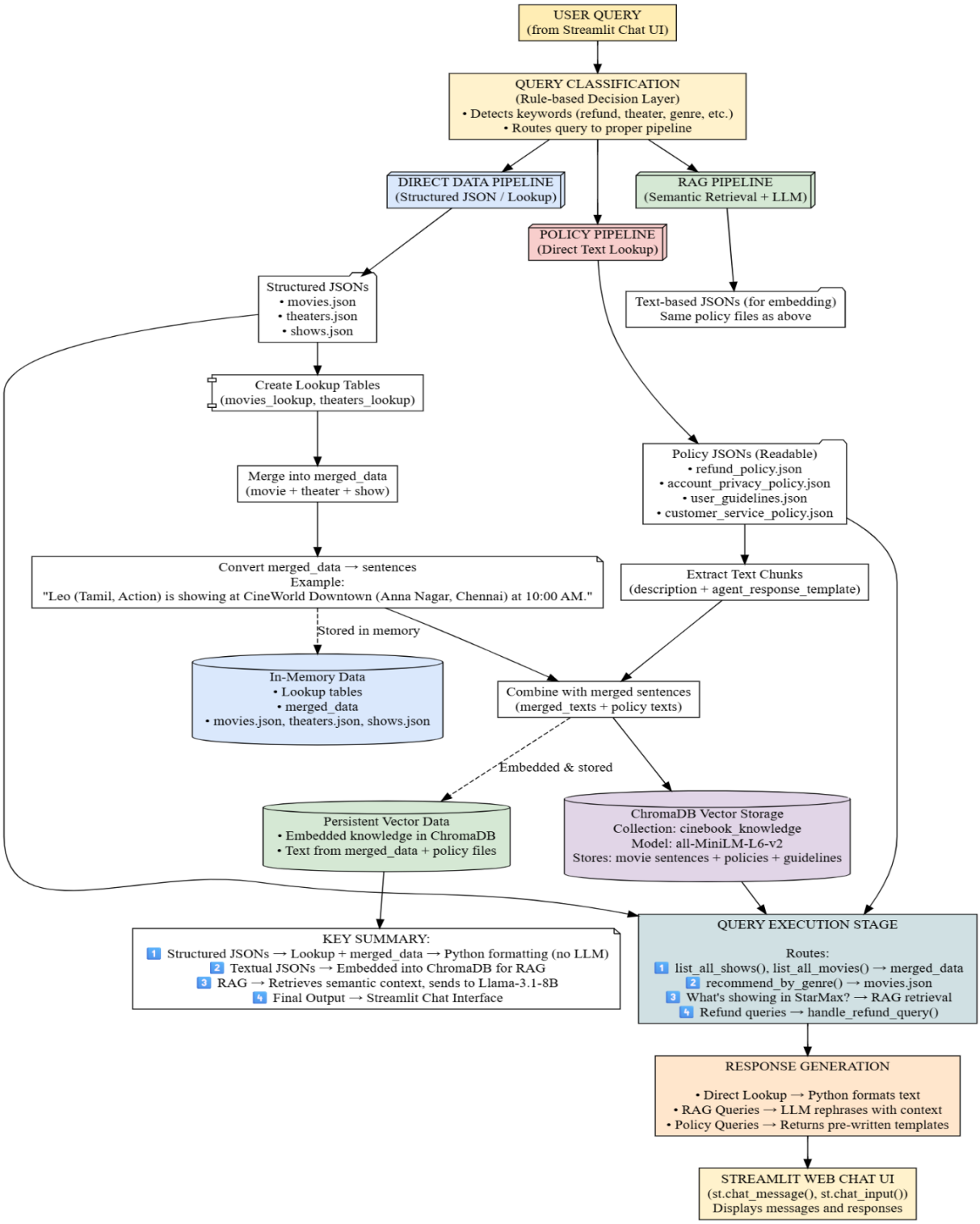
This closed-loop process allows CineBook to function as a semi-autonomous AI agent, capable of:

- Understanding the user's intent
- Retrieving relevant data or knowledge
- Reasoning contextually
- Delivering human-like, policy-compliant responses

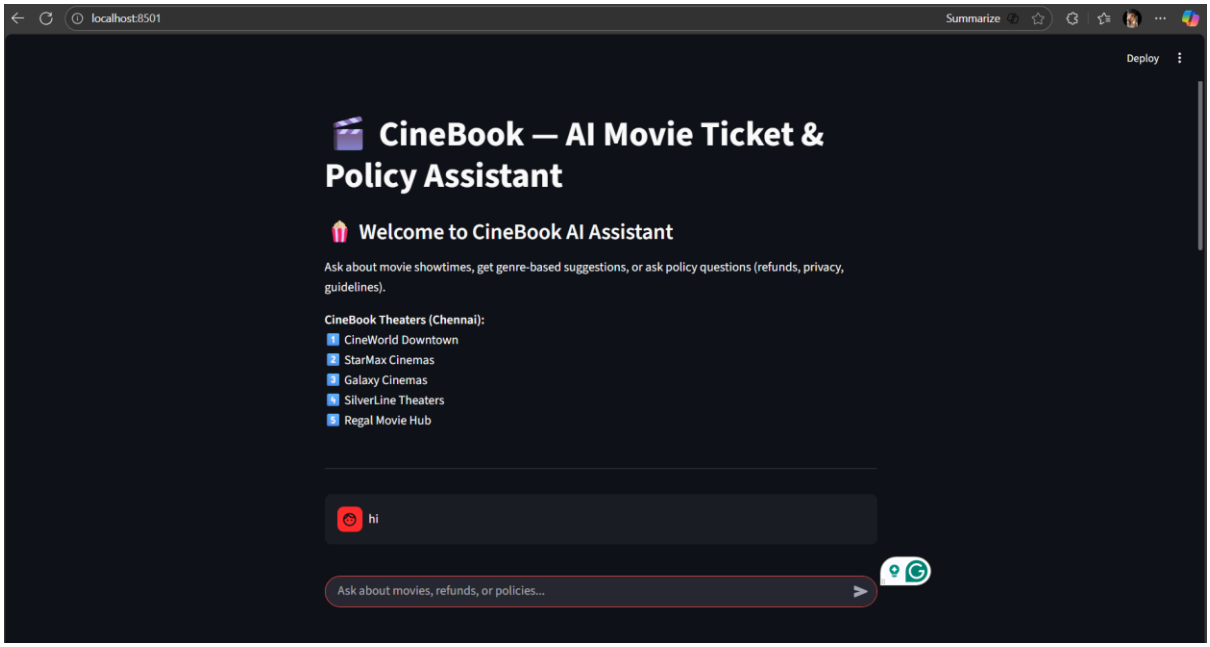
10) Final Insight

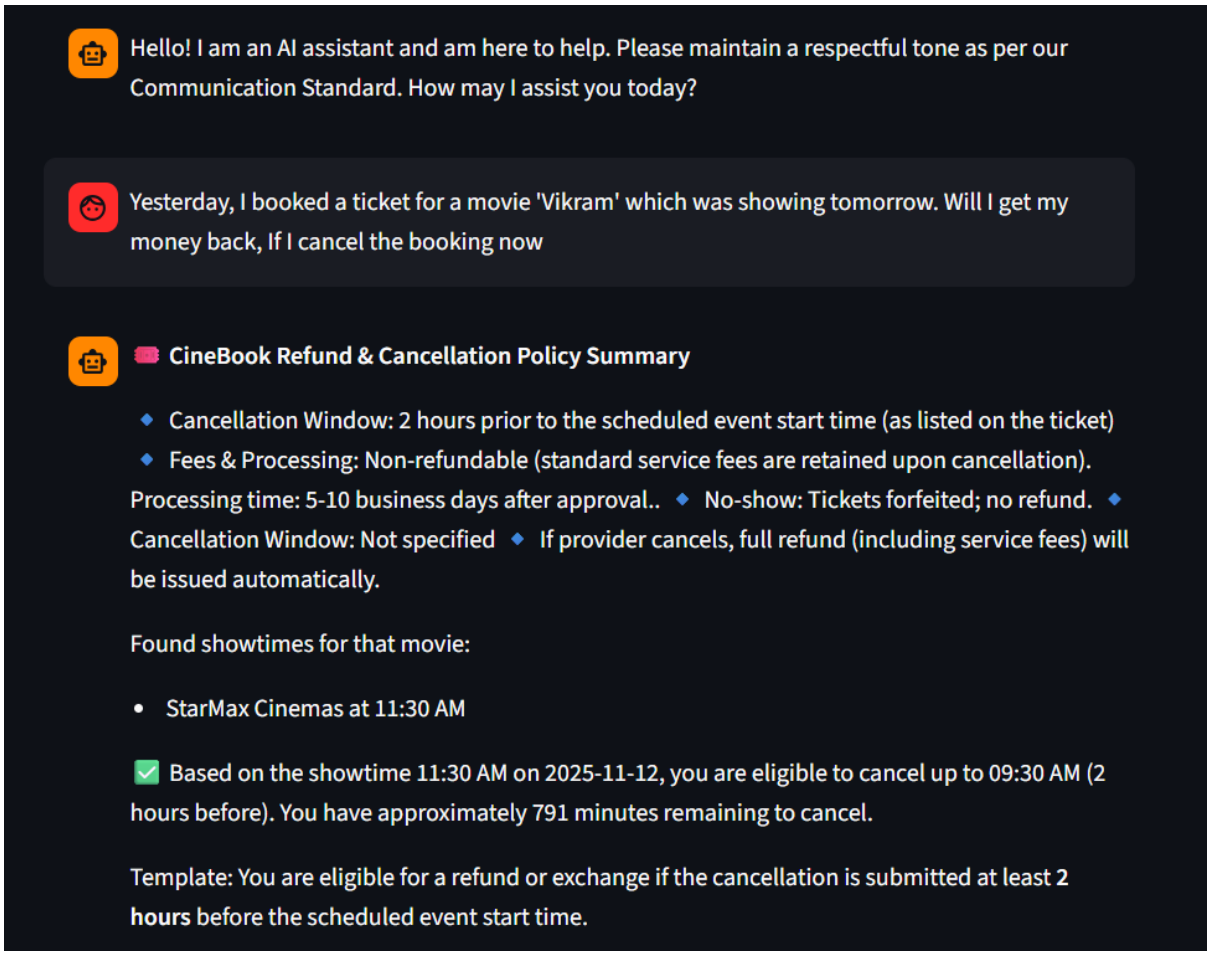
The CineBook methodology represents a powerful fusion of symbolic reasoning (rule-based lookups) and connectionist intelligence (LLM-based generation).It demonstrates how modern AI systems can achieve domain-specific intelligence by combining structured databases with neural embeddings and generative models.This hybrid design — blending deterministic precision and semantic understanding — makes CineBook more than just a chatbot. It is a Retrieval-Augmented Conversational AI Systecapable of dynamic, data-driven, and policy-grounded human-like interactions.

CINIBOOK FLOWCHART



Website & interactions with user:





tell me all the cinema halls with show timings

Here's the information about cinema halls with some show timings:

1. **CineWorld Downtown** in Anna Nagar:
    - **Leo** (Tamil, Action): 10:00 AM
    - **Jailer** (Tamil, Action/Drama): 02:00 PM
2. **StarMax Cinemas** in Velachery:
    - **Vikram** (Tamil, Thriller): 11:30 AM
    - **Leo** (Tamil, Action): 06:00 PM
3. **Galaxy Cinemas** in Express Avenue:
    - **Kalki 2898 AD** (Telugu, Sci-Fi): 01:00 PM
4. **SilverLine Theaters** in Royapettah:
    - **Kalki 2898 AD** (Telugu, Sci-Fi): 09:15 PM
5. **Regal Movie Hub** in Porur (Note: Show timings are for single movie - Master):

- **Master** (Tamil, Action/Drama): 07:00 PM

suggest me a film

Let's see what's trending. I have information on the following movies:

1. **Master** (Tamil, Action/Drama)
2. **Vikram** (Tamil, Thriller)
3. **Jailer** (Tamil, Action/Drama)
4. **Kalki 2898 AD** (Telugu, Sci-Fi)
5. **Leo** (Tamil, Action)

Which genre are you interested in?

i am intrested in Sci-Fi

🎬 Kalki 2898 AD (Telugu, Sci-Fi) — Galaxy Cinemas at 01:00 PM 🎬 Kalki 2898 AD (Telugu, Sci-Fi) — SilverLine Theaters at 09:15 PM

› 📄 Retrieved Context (for debugging/explainability)