CineMatch: A Hybrid LLM and Relational Approach for Personalized Movie Recommendation

Abdulrahman Alamoudi
Computer Science and Applications
Virginia Tech
aramoudi@vt.edu

Gary Young
Computer Science and Applications
Virginia Tech
garyyoung@vt.edu

Jack Lyons
Computer Science and Applications
Virginia Tech
jackl25@vt.edu

Yazeed Alharthi
Computer Science and Applications
Virginia Tech
yazeena@vt.edu

Yordanos Tessema
Computer Science and Applications
Virginia Tech
yordanost@vt.edu

***Abstract*—CineMatch is a hybrid movie recommendation system where each component handles the part of the problem it is suited for. XGBoost takes care of all candidate scoring and ranking. A Large Language Model handles three specific tasks: enriching the movie database through a PostgreSQL trigger pipeline, converting natural language queries into SQL, and writing short explanations for why each movie was recommended. PostgreSQL manages all structured storage, filtering, and aggregation. Apache AGE handles graph traversal for related-title discovery inside the same PostgreSQL cluster. A Python service owns the ML pipeline. User preference data is collected at signup through a short onboarding conversation so the system can make personalized recommendations from the very first session.**

***Index Terms*—Recommendation Systems, Large Language Models, XGBoost, Text-to-SQL, PostgreSQL, Apache AGE, Graph Databases, RAG, Trigger-Based Enrichment, Personalization**

I. Introduction

A. Motivation

Most recommendation systems work the same way: watch what users do, accumulate history, and slowly build up a taste model. This is fine when there is enough data, but it runs into two problems that do not get enough attention.

The first is cold-start. When someone creates an account, there is nothing to go on, so the system recommends whatever is popular. Some platforms add a preference quiz at signup, but those answers rarely reach the actual ranking model. The user gets generic results until watch history builds up.

The second is mood. Even users with years of history do not always want the same kind of movie they usually watch. A model trained on behavioral history can tell you what a person usually likes but not what they feel like watching right now. Mood shifts between sessions and a system that cannot account for it keeps getting things partially wrong.

CineMatch collects structured preference data at signup so recommendations are personalized from day one. Mood is treated as a real feature in the user profile and gets updated as users leave feedback.

B. Problem Statement

There are three main problems this project addresses. The first is the gap between how users express what they want and what a database and ML model can work with. A search like "something like Interstellar but less confusing" has real signal in it: sci-fi genre, preference for linear narrative, similar runtime. The system needs to extract that and turn it into a SQL query and scored features reliably. A rough similarity match is not enough.

The second is enriching the movie database without doing it by hand. The corpus has around 100,000 movies. Going through each one to assign mood labels and thematic keywords is not realistic. The system needs a pipeline that handles this automatically when a new movie is inserted.

The third is explainability. A ranked list with no context is hard to trust and hard to give feedback on. Recommendations should come with a reason.

C. Significance

The usual framing in recommendation research is that the database is just storage and the model does the real work. CineMatch questions that. The database here is not passive. It gets enriched automatically by the LLM through a trigger pipeline and handles filtering, aggregation, and graph traversal directly. The LLM is not the recommendation engine. It handles the parts involving natural language and nothing else. XGBoost does the ranking using structured features the database and onboarding flow have already prepared. This makes the system easier to inspect since any recommendation traces back to specific feature values, a specific SQL query, and a specific logged LLM output.

II. Related Work

A few platforms are worth examining to understand where CineMatch fits.

Letterboxd [1] is probably the best social film discovery tool available. Users log what they watch, write reviews, and follow each other. The community lists are genuinely useful. The problem is there is no personalized recommendation engine behind it. What you find depends entirely on who you follow. There is no ranking algorithm, no cold-start handling, and no preference model.

Rotten Tomatoes [2] aggregates critic and audience scores and is useful as a quality signal, but it does not try to model individual taste at all. Every user sees the same scores for every film.

IMDb [3] has the most complete structured movie data that exists, with cast, crew, and ratings at a depth nothing else matches. Its "More Like This" feature uses shallow genre overlap with no underlying preference model. It is a reference database, not a recommendation system.

TMDB [4] has a well-documented REST API that updates daily and covers a wide range of international cinema. It is the main external data source for CineMatch.

None of these platforms solve cold-start through onboarding that connects to a ranking model. None run LLM-based semantic enrichment as a live part of the data pipeline. None use graph traversal to find thematic connections between films. CineMatch puts all three together.

III. Proposed Approach

The design comes down to three ideas: collect real preference data at signup so the system works from session one; automatically enrich the movie database using an LLM trigger pipeline; and use XGBoost for all ranking so the LLM never touches the scoring path.

A. Interactive Onboarding

At signup, a short questionnaire builds a structured preference profile before the user has watched anything. It collects genre preferences with free-text notes, preferred runtime, decade range, language openness, and content tolerance stored as numeric values. The most useful part is the mood anchor: the user names three films they liked and three they did not. The LLM reads all six and extracts attributes like pacing, emotional tone, and narrative complexity, stored as a JSONB column. People are much better at naming films they have opinions about than describing those opinions in the abstract. The extracted attributes feed into XGBoost from the very first request.

B. LLM-Triggered Database Enrichment

Every time a new row gets inserted into the `movies` table, a PostgreSQL trigger fires and queues an enrichment job. The job sends the movie's title, genre list, and plot summary to Gemini Flash, which returns semantic tags covering mood, pacing, narrative complexity, and thematic keywords, each with a confidence score between 0 and 1. Those tags get written back into `movie_tags` with `source = 'llm'`. There is no batch job to schedule. Enrichment happens as a side effect of insertion.

The same pipeline builds the graph in Apache AGE. After tagging, the LLM scores thematic similarity between the new movie and existing movies that share genre or keyword overlap. An edge is created only when the similarity score clears a set threshold, which keeps the graph from just connecting everything in the same genre. Within minutes of being inserted, a movie is tagged and graph-connected.

C. Ranking, Feedback, and Architecture

When a user opens the home feed, the Go API loads their profile from PostgreSQL and calls the Python scoring service. That service filters out movies that fail hard constraints then runs XGBoost over the rest. The feature vector includes genre weights, mood profile attributes, the movie's assigned tags, average rating, runtime, release year, and cast-and-crew overlap with the user's stated preferences. The top 20 come back to the Go API. For the top 5, Gemini Flash generates a one-sentence explanation produced after ranking with no effect on the order. Feedback closes the loop: star ratings and not-interested signals update weights via exponential moving average, rejected titles are excluded permanently, and the model retrains weekly on accumulated data.

The system runs as four Docker containers via `docker-compose`: a **Go API server** for routing, JWT auth, and coordination; a **Python analytics service** (FastAPI) exposing `/score`, `/train`, `/enrich`, and `/rag`; **PostgreSQL 16 with Apache AGE** as the single source of truth; and a **React and TypeScript frontend** that only talks to the Go API.

IV. System Features

Personalized Home Feed. Twenty candidates ranked by XGBoost with one-sentence LLM explanations for the top five, generated after ranking and cached per session.

Conversational Movie Search. Users type queries in plain English. The Go API sends the query with the relevant schema to Gemini Flash, which returns a parameterized SQL `SELECT`. The query is checked against a whitelist before running: read-only, `LIMIT` required, no unsafe keywords. The generated SQL can be shown to the user as a transparency option. Filters include genre, MPAA rating, score range, year, runtime, language, director, actor, and mood tags.

Mood-Based Quick Pick. A mood selector offers six options: Feel-Good, Tense, Thought-Provoking, Funny, Romantic, and Epic. Picking one filters `movie_tags` and re-ranks the results against the user's XGBoost score. The mood tags were written by the trigger pipeline when each movie was inserted.

Related Movies via Graph. Each movie detail page has a "You might also like" section from Apache AGE traversal. Edges represent shared director, shared lead actor, shared genre, and LLM-scored thematic similarity. Going two or three hops out finds movies with multiple indirect connections,

a stronger signal than a flat genre match. AGE runs inside PostgreSQL so no separate infrastructure is needed.

City-Level Activity Map. A live map shows the most-watched movie per city and state. A background Go service generates simulated activity written to `activity_events` with a `city_state` field. A materialized view refreshes every 60 seconds over a rolling 10-minute window. The frontend renders this with ECharts using city and state GeoJSON.

Analytics Dashboards. ECharts charts backed by pre-computed PostgreSQL views cover genre trends over time, rating distribution by decade, top directors by recommendation frequency, a user activity heatmap, cold-start versus returning-user accuracy, mood tag popularity by rating, and language diversity. Free-text analytics questions get turned into SQL by the LLM and rendered as a chart.

Explicit Feedback Loop. Star ratings and not-interested signals update genre and mood weights through exponential moving average. Rejected titles are permanently removed from the candidate set between retraining cycles.

V. Datasets

The corpus is built from four sources merged on TMDB ID. The **TMDB Movies Dataset** [6] is the main source with about 100,000 titles from a collection of over one million daily-updated entries. The **IMDb Non-Commercial Datasets** [7] add crew data and audience ratings across 1.6 million titles. The **Netflix Reviews Dataset** [8] seeds the activity simulation and provides sentiment signals for feature engineering. The **Kaggle Movies Dataset** [5] verifies the ingestion pipeline is working correctly. Poster images and daily updates come from the TMDB [4] and OMDB [9] APIs.

VI. Checkpoint I

In this project checkpoint, we revised and clarified the features of our project based on feedback, including the schema, tech stack, and LLM scope. We have also redefined and reallocated group member roles in accordance with our expected project timeline to ensure an even distribution of work across our remaining time. With these tasks completed, we have finalized the design of our application and the schema of our database, and have begun building our backend.

LLM Guardrails. One of our concerns when enabling an LLM to dynamically construct SQL queries is the possibility to perform destructive actions on our database. As indicated by our conversational movie search feature, we intend for the LLM to parse plain text into a parameterized SELECT statement, and as such we wish to whitelist certain behavior that could be possible from user input. To ensure the integrity of our database, we are imposing limitations on the types of SQL statements the LLM is able to perform by checking the queries it constructs for specific keywords. We intend to implement a filter that checks all LLM generated queries and blocks potentially harmful ones from being executed on our database. This filter will be implemented in our backend to integrate with the LLM such that we can provide sample queries during testing. All statements capable of altering the

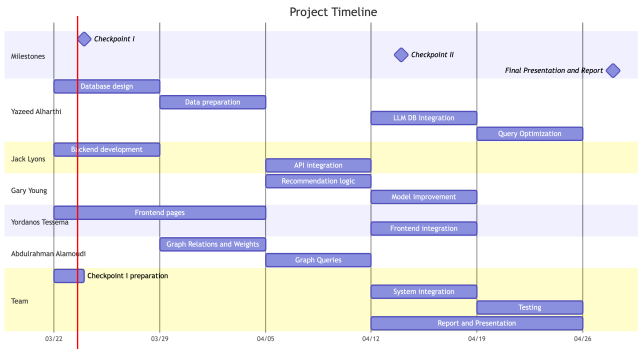

Fig. 1. Project Timeline

rows in our movie dataset will be forbidden by our filter, this includes INSERT, UPDATE, DELETE, and TRUNCATE. Additionally, we want to prevent any statements that could alter our database schema, including CREATE, ALTER, and DROP. The primary goal is to only allow for queries formatted as a SELECT statement and return an error for all others.

Users and Spatial Relations. The design of our database schema includes data for user locations as another means to provide accurate movie recommendations. As part of the initial user questionnaire intended to collect data for recommendations, the user can choose to provide location data. If provided, users will be able to toggle recommendations based on movie data from users in their area, improving accuracy by limiting preferences to what is popular in the area. This feature is intended to be simple to implement, using our same recommendation model on a subset of the full user dataset based on location data collected from the user during our initial questionnaire. We also intend for this feature to interact with our city-level activity map such that the user can visualize where their recommendations are derived.

Project Roles and Timeline. Since our project proposal, we have redefined each member's role in the project and have constructed a timeline of when each member should expect to complete each task they're assigned (See Fig. 1). This timeline is expected to be a guide, so the exact dates and members assigned to a specific task are subject to change if the amount of time and resources required is different from what we're projecting. We expect all members to contribute to system testing and integration to ensure the functionality of all components. All members will also contribute to the final report and presentation so we can accurately discuss all parts of the system.

Current Tasks and Milestones We have reached several goals at this stage of our project, and our project timeline indicates the tasks that are currently in progress. We completed the design for our database schema for both our users and our movie datasets, and we have a system to pull the relevant datasets from Kaggle. We have set up the required tools for our backend, including running Postgres in Docker. Our current tasks are developing our python scoring service and

questionnaire data insertion in the backend and developing the dashboard for our React frontend. Our reports on individual team member contributions reflect the changes in task allocation and our timeline.

Yazeed Alharthi

Yazeed is tasked with database design, data preparation, LLM DB integration, and query optimization. These tasks focus on determining our database schema and managing how the LLM interacts with it. For this checkpoint, he has set up Postgres, Docker, and Go API for our project.

Jack Lyons

Jack is tasked with backend development and API integration, focusing on the underlying logic and user features of the application. For this checkpoint, Jack primarily worked on expanding and formatting the report for this checkpoint to reflect the new changes to project goals, work allocation, and milestones. He is currently developing the questionnaire to collect initial user data and create entries in our user database.

Gary Young

Gary is tasked with recommendation logic and model improvement, focusing on designing and improving our core recommendation model. Thus, his plan is to consider parameters on recommendations for others. He also plans on supporting the team if there are problems. To simulate real user behavior, he plans to create dozens of virtual users with different locations and movie-watching activity within a limited area. The goal is to provide a sense of a more realistic user experience.

Yordanos Tessema

Yordanos' responsibility is to implement the frontend pages and integration, which entails a great deal of UX/UI work related to our application and our data. She works on the full visual stack of CineMatch, including the implementation of our React/TS frontend, such as our mood selector, our conversational search input, our movie pages, and our explicit feedback features.

Regarding data visualization, she integrates ECharts to render our analytics dashboards based on pre-computed PostgreSQL views, such as our genre trends, rating distributions, and our city-level activity map. She also works on ensuring our frontend remains responsive, as well as coordinating integration testing with our backend team.

Abdulrahman Alamoudi

Abdulrahman is responsible for designing the graph relationships, weights, and queries. His work focuses on implementing a graph structure to map movie interests and support our recommendation algorithm. He will prepare the data as a graph and build queries based on users' watched movies to find similar movies, as well as identify what other people in the same area are watching using PostGIS and graph technology.

Database Schema Overview

The schema splits into two groups. On the movie side, `movies` holds core metadata including title, year, runtime, language, ratings, and plot summary. Genres link through `movie_genres` and semantic tags from the LLM pipeline live in `movie_tags` with a float confidence score and a source field. Cast and crew go into `movie_crew` with a role column. On the user side, `user_preferences` stores genre weights and tolerance thresholds as JSONB, and `user_mood_profile` holds the onboarding-extracted mood attributes. Watch history, explicit feedback, and every recommendation made by the system each have their own table, forming the training data for XGBoost and the evaluation record for tracking model improvement over time.

Task Assignment

TABLE I
TASK ASSIGNMENT BY MEMBER

Task	Member
PostgreSQL schema design, migrations, and query optimization	Yazeed Alharthi
Data preparation and corpus ingestion pipeline	Yazeed Alharthi
LLM–database integration and query interaction	Yazeed Alharthi
Backend development and API integration	Jack Lyons
User questionnaire flow and backend data insertion	Jack Lyons
Recommendation logic and model improvement	Gary Young
Virtual user simulation for location-based movie activity	Gary Young
React/TypeScript frontend pages and UI integration	Yordanos Tessema
ECharts dashboards and city-level activity map integration	Yordanos Tessema
Responsive frontend design and UX/UI implementation	Yordanos Tessema
Apache AGE graph relationships, weights, and traversal queries	Abdulrahman Alamoudi
Spatial recommendation logic and local activity analysis using PostGIS	Abdulrahman Alamoudi
Integration, testing, final report, and presentation	All members

APPENDIX

REFERENCES

[1] Letterboxd. https://letterboxd.com/
[2] Rotten Tomatoes. https://www.rottentomatoes.com/
[3] IMDb. https://www.imdb.com/
[4] The Movie Database API. https://developer.themoviedb.org/docs/getting-started
[5] R. Banik, "The Movies Dataset," Kaggle. https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
[6] A. Saniczka, "TMDB Movies Dataset 2023," Kaggle. https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies
[7] IMDb Non-Commercial Datasets. https://datasets.imdbws.com/
[8] A. Kumarak, "Netflix Reviews," Kaggle. https://www.kaggle.com/datasets/ashishkumarak/netflix-reviews-playstore-daily-updated
[9] OMDB API. https://www.omdbapi.com/