Movie Recommendation with Agentic AI

Đánh giá mô hình

Phiên bản <1.2>

Lịch sử thay đổi

|  |  |  |  |
| --- | --- | --- | --- |
| **Ngày** | **Phiên bản** | **Mô tả thay đổi** | **Tác giả** |
| 25/05/2025 | 1.0 | Viết các nội dung đầu tiên | Lê Châu Hữu Thọ |
| 27/05/2025 | 1.1 | Bổ sung phần đánh giá mô hình Chatbot |  |
| 06/06/2025 | 1.2 | Cập nhật các số liệu đánh giá |  |
|  |  |  |  |

Mục lục

[1. Giới thiệu 4](#_Toc200145783)

[2. Mô tả mô hình 4](#_Toc200145784)

[2.1 Hệ thống đề xuất phim 4](#_Toc200145785)

[2.2 Chatbot đề xuất phim 4](#_Toc200145786)

[3. Thiết kế 5](#_Toc200145787)

[3.1 Bộ dữ liệu 5](#_Toc200145788)

[3.2 Đánh giá 5](#_Toc200145789)

[3.3 Các chỉ số đánh giá 6](#_Toc200145790)

[3.4 Công cụ đánh giá 6](#_Toc200145791)

[3.4.1 Phần cứng 6](#_Toc200145792)

[3.4.2 Phần mềm 7](#_Toc200145793)

[4. Kết quả triển khai 7](#_Toc200145794)

[4.1 Recommendation System 7](#_Toc200145795)

[4.2 Chatbot 7](#_Toc200145796)

[5. Nhận xét 8](#_Toc200145797)

Đánh giá mô hình

# Giới thiệu

Tài liệu này đề cập đến việc đánh giá hiệu quả của các mô hình Học máy và LLM được triển khai trong hệ thống "Movie Recommendation with Agent AI". Cụ thể, nó tập trung vào việc đánh giá mô hình **Hệ thống đề xuất phim (Content-based Recommendation System)** và **Chatbot đề xuất phim**. Việc đánh giá không chỉ tập trung vào các chỉ số kỹ thuật như độ chính xác, phạm vi bao phủ, hoặc thời gian phản hồi, mà còn xem xét khả năng mở rộng, ổn định và tính linh hoạt khi áp dụng trong môi trường thực tế.

Tài liệu này nên được sử dụng bởi tất cả các bên liên quan, bao gồm các nhà phát triển, kiểm thử viên, quản lý dự án và các bên liên quan khác.

Các đặc điểm chính của tài liệu này bao gồm việc trình bày rõ ràng:

* Mô tả chi tiết về các mô hình ML được đánh giá.
* Thiết kế thử nghiệm, bao gồm các bộ dữ liệu được sử dụng, các mô hình/đường cơ sở để so sánh, các chỉ số hiệu suất được áp dụng và thiết lập môi trường.
* Các kết quả đánh giá được báo cáo dưới dạng bảng biểu hoặc hình ảnh cùng với mô tả chi tiết.
* Phần kết luận tóm tắt kết quả thử nghiệm và đưa ra các kết luận liên quan đến các mô hình được sử dụng, bao gồm các hạn chế và đề xuất cải tiến trong tương lai.

# Mô tả mô hình

## Hệ thống đề xuất phim

* **Tên mô hình (ban đầu):** Content-based Recommendation System
* **Cách thức hoạt động:** Đề xuất phim dựa trên sự tương đồng giữa các đặc trưng (tags) của phim và sở thích đã chọn của người dùng khi tạo tài khoản.
* **Lý do lựa chọn:** Phù hợp với giai đoạn ban đầu khi hệ thống chưa có lịch sử tương tác đủ lớn từ người dùng.
* **Ưu điểm:** Có thể hoạt động ngay cả khi người dùng mới (cold-start problem) vì không yêu cầu dữ liệu hành vi quá nhiều.
* **Hạn chế:** Có thể thiếu sự đa dạng trong đề xuất nếu sở thích người dùng quá hẹp; cần người kiểm thử thủ công để đánh giá ban đầu.
* **Triển khai:** Mô hình được tích hợp vào thành phần **Recommendation System** trong Backend.
* **Huấn luyện lại (Retraining):** Dữ liệu phim mới được thêm vào hàng tuần, quy trình huấn luyện động để phù hợp với thời gian thực. Sau khi ứng dụng phát hành và có đủ dữ liệu người dùng, có thể chuyển sang sử dụng Collaborative Filtering để tối ưu hơn.
* **Đánh giá và cập nhật:** Ban đầu sẽ cần kiểm thử thủ công. Sau khi phát hành, sẽ đánh giá dựa trên thực tế.

## Chatbot đề xuất phim

* **Tên mô hình:** Chatbot đề xuất phim (sử dụng kiến trúc Retrieval-Augmented Generation - RAG)
* **Cách thức hoạt động:** Chatbot sử dụng API từ **Groq** làm mô hình ngôn ngữ lớn (LLM) để hiểu và tạo ra phản hồi. Dữ liệu liên quan đến phim và các thông tin khác (chung với mô hình đề xuất phim) được **embedding và lưu trữ trong cơ sở dữ liệu vector Chroma**. Khi người dùng đưa ra yêu cầu bằng ngôn ngữ tự nhiên, chatbot sẽ truy vấn các embedding có liên quan từ Chroma, sau đó chuyển thông tin này cùng với câu hỏi của người dùng tới LLM của Groq để tạo ra câu trả lời và đề xuất phim phù hợp. Chatbot phân tích yêu cầu bằng ngôn ngữ tự nhiên của người dùng, truy vấn dữ liệu liên quan và đưa ra gợi ý phù hợp.
* **Lý do lựa chọn:** Tăng cường khả năng hiểu và phản hồi chính xác dựa trên ngữ cảnh dữ liệu thực tế; khả năng truy xuất dữ liệu nhanh chóng và linh hoạt, phù hợp với nhiều nguồn dữ liệu khác nhau (phim, lịch chiếu, thông tin rạp). Việc sử dụng Groq giúp tối ưu hóa tốc độ suy luận của LLM và tiết kiệm chi phí do Groq cung cấp API miễn phí.
* **Ưu điểm:** Tăng cường khả năng hiểu và phản hồi chính xác dựa trên ngữ cảnh dữ liệu thực tế; khả năng truy xuất dữ liệu nhanh chóng và linh hoạt, phù hợp với nhiều nguồn dữ liệu khác nhau. Tốc độ xử lý của Groq giúp cải thiện đáng kể thời gian phản hồi của chatbot. Ngoài ra nó còn phù hợp với quỹ vốn của dự án.
* **Hạn chế:** Phụ thuộc vào chất lượng của dữ liệu được embedding và truy xuất từ Chroma; có thể yêu cầu chi phí cho API của Groq; cần huấn luyện hoặc tinh chỉnh để hiểu rõ các ngữ cảnh liên quan đến phim và đặt vé.
* **Triển khai:** Mô hình Chatbot chạy trong hệ thống Backend (thành phần Agent Query, Prompts, và Tools), tích hợp API của Groq và cơ sở dữ liệu vector Chroma để truy xuất thông tin.
* **Huấn luyện lại (Retraining):** Điều chỉnh định kỳ mô hình ngôn ngữ (nếu có fine-tuning) và cập nhật cơ sở dữ liệu truy xuất (ví dụ: cập nhật thông tin phim, rạp chiếu, embedding mới).
* **Đánh giá và cập nhật:** Được kiểm tra dựa trên độ chính xác của câu trả lời, thời gian phản hồi và mức độ hài lòng của người dùng.

# Thiết kế

## Bộ dữ liệu

Dữ liệu cho hệ thống đề xuất phim và chatbot được thu thập từ hai nguồn chính:

* **Cơ sở dữ liệu MongoDB:** Chứa dữ liệu về phim, có nguồn gốc từ việc tiền xử lý (preprocessing) các tập dữ liệu phim lớn như IMDB movie dataset. Dữ liệu này bao gồm thông tin chi tiết về phim như tiêu đề, thể loại, diễn viên, đạo diễn, mô tả, v.v.. Đối với Chatbot, dữ liệu này cũng đóng vai trò là nguồn thông tin chính để truy xuất khi người dùng hỏi về phim, sau khi được embedding và lưu trữ trong Chroma.
* **Cơ sở dữ liệu MySQL:** Lưu trữ dữ liệu người dùng và các tương tác của họ với hệ thống và phim. Bao gồm thông tin tài khoản, lịch sử xem, đánh giá (vote sao, bình luận), phim yêu thích, lịch sử đặt vé, và các sở thích đã lưu của người dùng.
* **API Rạp chiếu phim (cho Chatbot):** Đối với việc đề xuất/tìm vé xem phim, chatbot sẽ truy xuất dữ liệu lịch chiếu và vé còn trống thông qua các API của rạp phim.

Quá trình chuẩn bị dữ liệu bao gồm làm sạch dữ liệu (xử lý các giá trị thiếu, không nhất quán), chuyển đổi định dạng nếu cần, và tích hợp dữ liệu từ các nguồn này để tạo ra tập dữ liệu huấn luyện và truy xuất đầy đủ.

## Đánh giá

* **Hệ thống đề xuất phim (Content-based):** Vì đây là giai đoạn phát triển ban đầu, tập trung vào việc xác thực khả năng đề xuất dựa trên nội dung, nên việc so sánh trực tiếp với các mô hình phức tạp hơn có thể chưa phù hợp. Thay vào đó, nhóm sẽ coi chính hệ thống Content-based là một **đường cơ sở (baseline)** cho các đánh giá nội bộ ban đầu, đặc biệt là thông qua **đánh giá thủ công** để xác định mức độ phù hợp của các đề xuất với các tag đã chọn ban đầu của người dùng. Sau này, khi có đủ dữ liệu người dùng, nhóm sẽ so sánh với các kỹ thuật **Collaborative Filtering** hoặc các mô hình **Neural Collaborative Filtering** đã được đề cập trong SAD như một mục tiêu phát triển tiếp theo.
* **Chatbot đề xuất phim:** Với bản chất phức tạp và sự kết hợp của mô hình ngôn ngữ lớn (LLM từ Groq) với hệ thống truy xuất dữ liệu động (từ Chroma), không có nhiều mô hình đường cơ sở trực tiếp để so sánh. Thay vào đó, nhóm sẽ đánh giá hiệu suất của hệ thống dựa trên các chỉ số thực tế như độ chính xác của câu trả lời, thời gian phản hồi và khả năng truy xuất thông tin liên quan, so với các hệ thống chatbot truyền thống không sử dụng phương pháp RAG

## Các chỉ số đánh giá

Để đánh giá hiệu quả của các mô hình Học máy trong nghiên cứu này, nhóm sẽ sử dụng các chỉ số hiệu suất được điều chỉnh cho từng loại mô hình và mục tiêu vấn đề:

* **Hệ thống đề xuất phim:**
  + **Đánh giá Offline (ban đầu, thủ công):**
    - **Precision và Recall (Top-N Recommendation):** Tỷ lệ phim được đề xuất chính xác so với các tag được cung cấp ban đầu.
    - **F1-Score:** Kết hợp Precision và Recall.
    - **Diversity:** Đánh giá sự đa dạng của các phim được đề xuất trong danh sách Top-N.
    - **Novelty:** Đánh giá mức độ "mới lạ" của các đề xuất (liệu có đề xuất những phim mà người dùng ít có khả năng tự tìm thấy không).
    - **Đánh giá thủ công:** Ước tính số tập đánh giá khoảng 100 - 200 bộ phim để kiểm tra độ phù hợp của đề xuất.
  + **Đánh giá dựa trên thực tế (Online Evaluation - sau khi phát hành):**
    - **Tỷ lệ nhấp (Click-Through Rate - CTR):** Tỷ lệ người dùng nhấp vào các phim được đề xuất.
    - **Thời gian xem/tương tác:** Thời gian người dùng dành cho các phim được đề xuất.
    - **Khảo sát người dùng:** Trực tiếp thu thập phản hồi từ người dùng về mức độ hài lòng với các đề xuất.
* **Chatbot đề xuất phim:**
  + **Độ chính xác (Answer Accuracy):** Tỷ lệ các câu trả lời của chatbot là đúng hoặc phù hợp với câu hỏi của người dùng.
  + **Thời gian phản hồi (Response Time):** Thời gian cần thiết để chatbot trả lời một câu hỏi, ảnh hưởng trực tiếp đến trải nghiệm người dùng.
  + **Mức độ hài lòng của người dùng (User Satisfaction):** Đánh giá định tính dựa trên phản hồi của người dùng trong quá trình sử dụng.

Việc sử dụng đa dạng các chỉ số đánh giá giúp đảm bảo các mô hình được đánh giá toàn diện từ các khía cạnh về độ chính xác, hiệu quả và trải nghiệm thực tế.

## Công cụ đánh giá

### Phần cứng

|  |  |
| --- | --- |
| **Thành phần** | **Yêu cầu** |
| CPU | Intel Core i3 là tối thiểu |
| RAM | 16-32 GB, để lưu trữ và xử lý dữ liệu vừa phải. |
| GPU | NVIDIA (ví dụ: GTX 1060/RTX 2060 hoặc tương đương) để tăng tốc huấn luyện và suy luận các mô hình học sâu và mô hình ngôn ngữ lớn (LLM) trong chatbot. |
| SSD | 512 GB hoặc hơn, đảm bảo truy cập dữ liệu nhanh và ổn định. |

### Phần mềm

|  |  |
| --- | --- |
| **Phần mềm/Công cụ** | **Yêu cầu/Mục đích** |
| Hệ điều hành | Windows 10/11 hoặc Ubuntu 24.04, tùy thuộc vào môi trường phát triển chính. |
| Python | Phiên bản 3.9+ |
| PyTorch | Cho việc xây dựng và huấn luyện mô hình học sâu (nếu được áp dụng sau này). |
| scikit-learn | Cho các tác vụ tiền xử lý dữ liệu, hoặc các mô hình đơn giản. |
| API Client cho Groq | Để tích hợp và sử dụng API của Groq cho chatbot. |
| ChromaDB Client (chromadb) | Để tương tác với cơ sở dữ liệu vector Chroma. |
| Môi trường ảo (virtualenv hoặc conda) | Để quản lý các thư viện và phiên bản phần mềm. |

# Kết quả triển khai

Trong quá trình đánh giá ban đầu, nhóm đã triển khai và đánh giá **Hệ thống đề xuất phim dựa trên nội dung (Content-based Recommendation System)** và **Chatbot đề xuất phim**.

## Recommendation System

|  |  |
| --- | --- |
| **Chỉ số** | **Kết quả** |
| Precision@10 | 0.65 |
| Recall@10 | 0.58 |
| F1-Score | 0.61 |
| Diversity | 0.70 |
| Novelty | 0.55 |

## Chatbot

|  |  |
| --- | --- |
| **Chỉ số** | **Kết quả** |
| Accuracy | 0.75 |
| Response Time (ms) | 200 |
| User Satisfaction | - |

# Nhận xét

Qua quá trình đánh giá ban đầu, cả **Hệ thống đề xuất phim dựa trên nội dung** **(Content-based Recommendation System)** và **Chatbot đề xuất phim** đã được triển khai và đánh giá. Mô hình đề xuất phim cho thấy tiềm năng trong việc cá nhân hóa trải nghiệm người dùng dựa trên sở thích. Mô hình chatbot chứng tỏ khả năng phản hồi chính xác và nhanh chóng, cải thiện tương tác.

Tuy nhiên, vẫn còn một số hạn chế: mô hình recommend và chatbot hiện tại **chưa đạt được các chỉ số kỳ vọng** đã đặt ra trong tài liệu SAD. Đối với chatbot, dữ liệu hiện tại chỉ mới được nhúng các thành phần cơ bản như **tên sản phẩm, thể loại và mô tả**.

Trong tương lai, nhóm sẽ tiếp tục tối ưu hóa quy trình huấn luyện, cải thiện hiệu suất đề xuất và mở rộng dữ liệu để tăng độ chính xác. Đặc biệt, khi dữ liệu hành vi người dùng đủ lớn, nhóm sẽ chuyển sang triển khai và đánh giá các mô hình Collaborative Filtering hoặc Neural Collaborative Filtering. Đối với chatbot, nhóm cần tiếp tục mở rộng và cải thiện dữ liệu cũng như các thành phần khác trong tương lai để nâng cao hiệu suất và đáp ứng tốt hơn mục tiêu đề ra.