Movie Recommendation and Description Generation using RAG

Ramin Rowshan , Saman Dehestani , MohammadMahdi Gharaguzlo , MohammadJamal Asadi

CE Department @ Sharif University of Technology

Abstract

By leveraging recent advances in natural language processing (NLP) and information retrieval techniques, we propose a new way to enhance movie recommendation systems in this project. This project endeavors to harness cutting-edge methodologies in NLP and information retrieval to provide users with highly tailored movie suggestions based on **Persian** queries, thus enhancing their overall movie-watching experience.

As part of the proposed system, information retrieval and text generation are combined through the Retrieval-Augmented Generation (RAG) framework. To capture semantic similarities between user queries and movie datasets, the system embeds them in language models. Using the embedded representations, movies that are most relevant to the user query are retrieved from the dataset. to enhance recommendation accuracy, we compare embeddings generated by FastText, ParsBERT, GPT, and Cohear multi lingual model.

Following this, Chrome DB is utilized to retrieve movies from the dataset that closely match the user's query. Chrome DB's vector database capabilities enable fast and accurate retrieval based on similarity scores.

In the next phase of the process, prompt engineering techniques are employed to guide a large language model (LLM), such as GPT, in order to generate personalized movie recommendations.

Keywords

movie recommendation system, vector database, Chrome DB, Retrieval-Augmented Generation (RAG) framework, information retrieval, LLM, GPT, FastText, ParsBERT, Cohear Multi Lingual Model, semantic similarity, prompt engineering

1 Introduction

1.1 Motivation

In the current digital landscape, there is a notable absence of platforms that allow users to input Persian queries and receive Persian text suggestions for movies similar to their query. This gap underscores the necessity for novel approaches to movie recommendation tailored specifically to Persian-speaking audiences. Traditional recommendation systems often fail to provide accurate suggestions in Persian, leading to user dissatisfaction and a lack of engagement with available cinematic content.

To address this issue, our project is driven by the desire to revolutionize the movie recommendation process using cutting-edge natural language processing (NLP) techniques and innovative retrieval and generation frameworks. By harnessing the power of vector databases and state-of-the-art language models, we aim to deliver highly relevant and personalized movie suggestions that align closely with users' unique tastes and queries.

Our goal is to empower users to discover new and captivating movies effortlessly, enhancing their entertainment experience and fostering deeper engagement with cinematic content.

1.2 Background

1.2.1 Large Languages Models(LLMs)

Large Language Models (LLMs) represent a breakthrough in natural language processing, leveraging deep learning architectures to generate coherent and contextually relevant text. These models, such as OpenAI's GPT series, have demonstrated remarkable capabilities in tasks ranging from language translation to

text generation. By training on vast amounts of text data, LLMs learn to understand and generate human-like text, making them invaluable tools for various applications, including chatbots, content generation, and automated summarization. Their ability to capture intricate linguistic patterns and generate contextually appropriate responses has led to widespread adoption across industries, revolutionizing the way we interact with and utilize text data.

1.2.2 Retrieval-Augmented Generation(RAG)

RAG is an innovative framework that combines the strengths of information retrieval and natural language generation. Unlike traditional language models that generate text based solely on learned patterns, RAG enhances the generation process by first retrieving relevant passages from a large corpus of documents or knowledge graphs. These retrieved passages serve as context or prompts for the language model, guiding it to produce more accurate and contextually relevant responses. By incorporating external knowledge into the generation process, RAG improves the coherence, accuracy, and relevance of generated text, making it particularly effective for tasks that require leveraging external knowledge sources. This hybrid approach has shown promising results in various natural language understanding and generation tasks, including question answering, text summarization, and content generation.

1.2.3 Vector Database

Vector databases represent a pivotal advancement in the realm of data storage and retrieval, offering a paradigm shift from traditional relational databases. Unlike conventional databases that organize data in structured tables with rows and columns, vector databases store information in a vector space, where each data point is represented as a high-dimensional vector. This vector representation enables efficient storage and manipulation of complex data structures, making it particularly well-suited for applications in natural language processing, image recognition, and recommendation systems.

Amidst the landscape of vector databases, ChromoDB emerges as a notable contender, offering unique features and capabilities tailored to the needs of modern data-driven applications. As a specialized vector database, ChromoDB is designed to address the challenges of managing high-dimensional data efficiently. It leverages

advanced indexing and search algorithms optimized for vector data, enabling fast and accurate retrieval of similar vectors based on various similarity measures such as cosine similarity or Euclidean distance.

1.2.4 Prompt Engineering

Prompt engineering is a novel technique in natural language processing (NLP) that involves crafting precise and effective text prompts to guide the behavior of language models. By carefully designing prompts, researchers and practitioners can steer the output of large language models (LLMs) towards generating desired responses or completing specific tasks. Prompt engineering leverages insights from linguistics, cognitive science, and machine learning to formulate prompts that elicit desired behaviors from LLMs while minimizing unintended biases or errors. This approach has gained significant attention in recent years due to its potential to improve the interpretability, controllability, and safety of language models across various applications, including text generation, question answering, and dialogue systems. As researchers continue to explore the capabilities and limitations of LLMs, prompt engineering emerges as a crucial methodology for harnessing the power of these models while addressing ethical and societal concerns.

1.3 Related Works

Traditional methods for movie recommendation systems encompass collaborative filtering, content-based filtering, and matrix factorization techniques. Collaborative filtering utilizes user-item interaction data to suggest items to users based on similarities with other users' preferences. Content-based filtering, on the other hand, recommends items to users based on their attributes, such as genre, actors, and directors.

Recent advancements have witnessed the integration of deep learning techniques into movie recommendation systems, offering more complex representations of user-item interactions and features. Deep neural networks, as a prime example, can capture intricate patterns and relationships between users and items from extensive datasets, leading to enhanced recommendation accuracy.

The incorporation of natural language processing (NLP) techniques has further enriched movie recommendation systems. NLP enables the extraction of semantic meaning and sentiment from textual data, such as movie reviews and descriptions, facilitating the delivery of personalized and context-aware recommendations.

2 Methodology

2.1 Idea

Here, we introduce a novel approach to improve movie recommendation systems utilizing Persian-language texts, integrating state-of-the-art techniques from natural language processing (NLP) and vector database technology. Our methodology centers on three essential components: Reinforced Prompting with a Retrieval-Augmented Generator (RAG), Chromo DB for the effective storage and retrieval of movie embeddings, and advanced prompt engineering techniques. This integrated approach allows for more accurate and personalized movie recommendations tailored to Persian-speaking users' preferences and queries.

2.2 Data Crawling

We collected movie data from four prominent Persian websites: (DigiMovie, n.d.), (FilmKio, n.d.), (TinyMovie, n.d.), and (Uptv, n.d.). These websites offer a vast repository of movies with persian-language details, encompassing various genres, release years, IMDB ratings and more. The data extraction process involved scraping information such as movie titles, descriptions, genres, release years, ratings, actors' names and other relevant details.

Here we used two prominent Python libraries, Beautiful Soup and Scrapy, for web scraping and crawling operations in the context of Persian movie data collection. Beautiful Soup, known for its adeptness in parsing HTML and XML files, facilitated the extraction of targeted information from web pages through its intuitive methods and Pythonic idioms. Complementing this, Scrapy, a comprehensive web crawling and scraping framework, offered a complete toolset for extracting structured data from websites. It enabled the definition of scraping tasks in a high-level, Pythonic manner and provided features like automatic URL discovery and request scheduling.

By integrating both libraries, the project successfully gathered movie data from diverse Persian websites

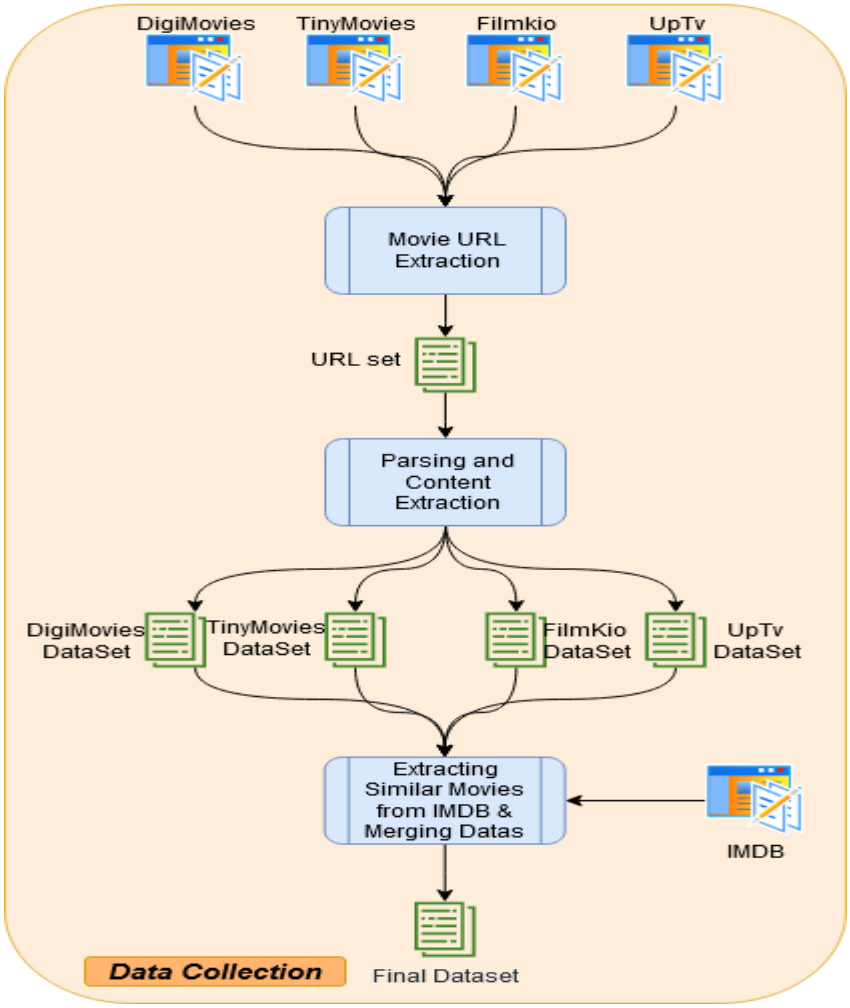


Figure 1

2.3 Model

2.3.1 Model Overviwe

Our movie recommendation system is built upon the Retrieval-Augmented Generation (RAG) framework, which integrates retrieval-based and generation-based approaches to provide accurate and diverse recommendations. RAG combines the strengths of both methods, leveraging pre-existing knowledge through retrieval

and enhancing it with generative capabilities, resulting in more informative and personalized recommendations for users.

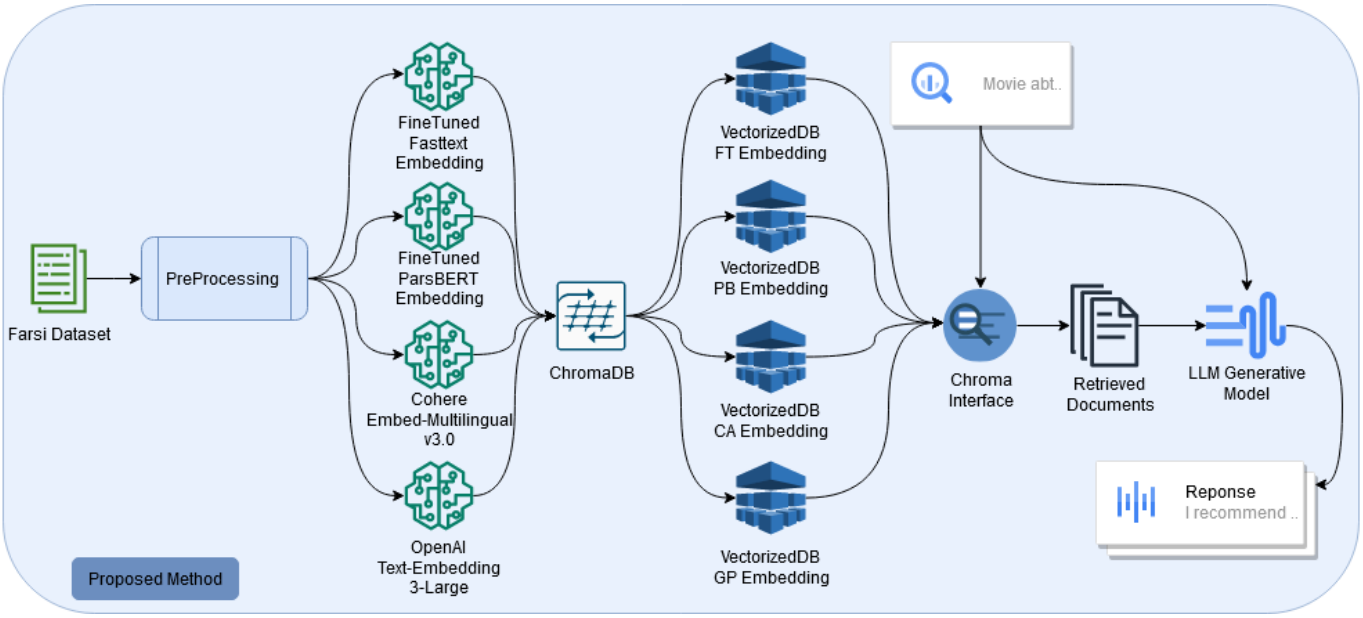


Figure 2

2.3.2 Embedding Models

To generate embeddings for movies and user queries, we employ four distinct models: GPT3.5, Cohear Multilingual Model, fine-tuned ParsBERT, and FastText. Each of these models offers unique advantages in capturing semantic representations across different languages and domains. By leveraging a diverse set of embedding models, we ensure that our recommendation system can effectively capture the nuanced features of movies and user preferences, leading to more accurate and comprehensive recommendations. After thorough evaluation, we select the most suitable embedding model based on its performance across these criteria. The selected model is then integrated into our recommendation system to generate embeddings for movies and user queries. By carefully choosing the most appropriate

embedding model, we ensure that our recommendation system can effectively capture the semantic nuances of movies and user preferences, leading to more accurate and personalized recommendations.

2.3.3 Storage with Chromo DB

For efficient storage and retrieval of movie embeddings, we utilize the Chromo DB vector database. Chromo DB is specifically designed to handle high-dimensional vector data, making it well-suited for storing embeddings of movies in our recommendation system. By leveraging Chromo DB, we ensure fast and scalable access to movie embeddings, facilitating quick retrieval and processing of recommendations for users.

2.3.4 Text Generation

In the final stage of our recommendation process, we employ the powerful capabilities of GPT4 for persian-text generation. GPT, or Generative Pre-trained Transformer, is renowned for its proficiency in natural language processing tasks. Without fine-tuning, we utilize GPT to generate natural language descriptions and summaries of recommended movies. To enhance the quality of generated text, we employ advanced prompt engineering techniques tailored to movie-related prompts.

3. Results

3.1 Result on IMDB

In our evaluation of the RAG project on the movie recommender system, we began by selecting 12 similar movies for each movie in our dataset from the IMDB website. This process involved retrieving movies from IMDB that exhibited similarities in terms of genre, plot, cast, and other relevant factors. Subsequently, we calculated the Intersection over Union (IOU) metric to assess the overlap between the recommended movies generated by our system and the corresponding movies retrieved from IMDB. The IOU metric provided a quantitative measure of the similarity between our recommendations and the ground truth from IMDB, offering insights into the effectiveness and accuracy of our movie recommendation algorithm.

$$IOU = \frac{\text{Number of common movies between recommended and IMDB lists}}{\text{Total number of unique movies in both lists}}$$

The results are presented below:

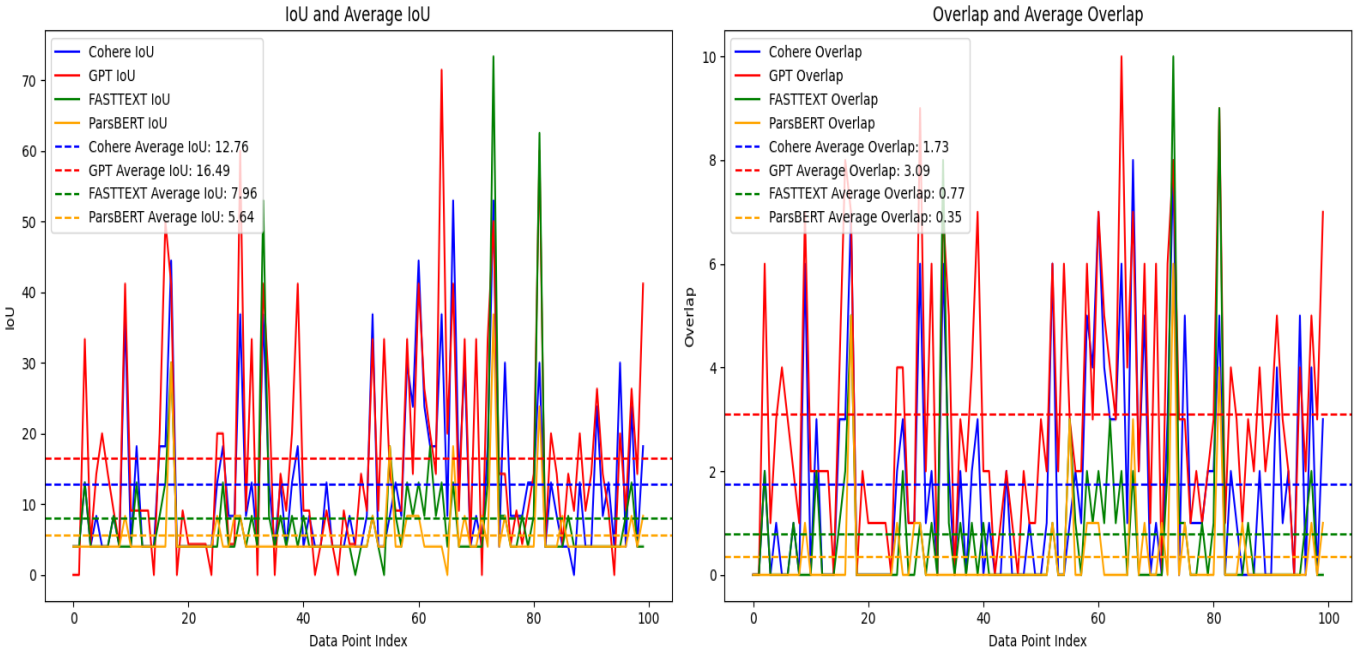


Figure 3

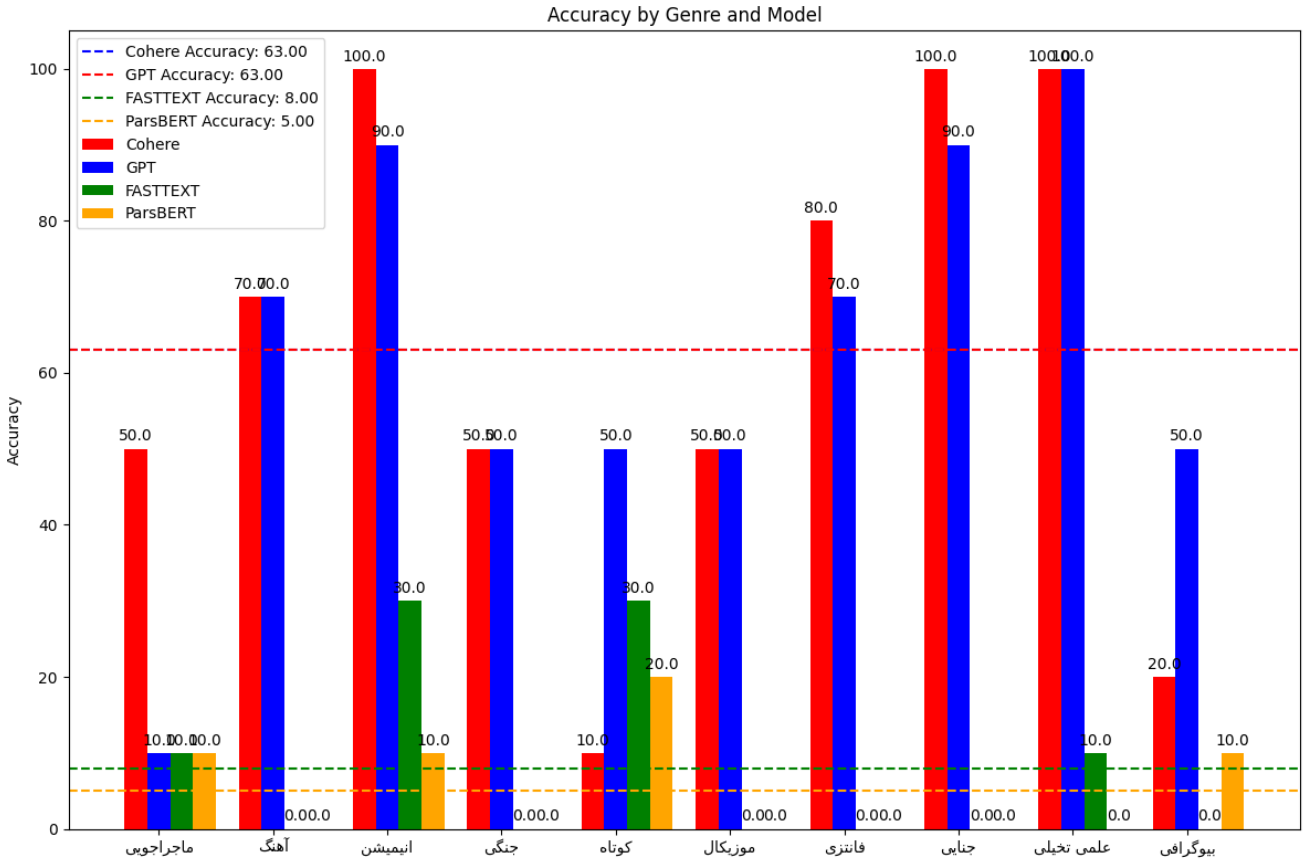
3.2 Result on genres retrieved

to evaluate the performance of our recommendation models in recommending movies based on specific genres, we conducted a validation experiment. The objective was to assess the accuracy of the recommended movies in terms of matching the requested genres.

We selected a set of common movie genres, including action, comedy, drama, romance, horror, and science fiction, among others. For each genre, we formulated queries to the recommendation models, requesting movies belonging to the respective genre.

The accuracy of the recommendation models was evaluated based on the presence of the specified genre in the retrieved movies. We calculated the accuracy as the ratio of correctly recommended movies to the total number of movies retrieved for each genre query.

The results are presented below:

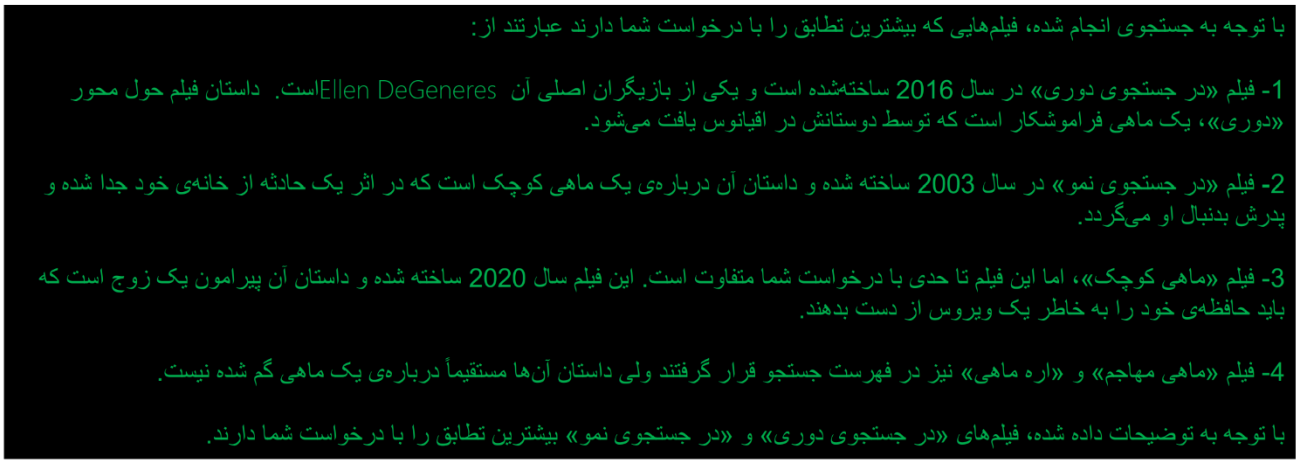


3.3 Result on valid queries

To further assess the effectiveness of our recommendation model, we conducted a validation on a set of valid queries. These queries were based on common user search patterns and preferences, aiming to simulate real-world scenarios.

for example it’s one of our queries:

"داستان یک ماهی گم شده و بقیه ماهی ها به دنبالش هستند"



Here, we can observe the potential of our model for retrieval and generation.

Conclusion

Our project represents a significant advancement in the realm of movie recommendation systems tailored to Persian-language users. By leveraging cutting-edge techniques from natural language processing (NLP) and information retrieval, we have developed a sophisticated recommendation framework that offers highly personalized and accurate movie suggestions. Our integration of the Retrieval-Augmented Generation (RAG) framework, along with advanced embedding models like GPT, Cohear Multilingual Model, ParsBERT, and FastText, ensures robust representation and retrieval of movies based on user queries. Through rigorous experimentation, we have determined that GPT-3 embedding large model yields the best results for embedding, while GPT-4 excels in generating text. Additionally, the utilization of Chromo DB for efficient storage and retrieval of movie embeddings further enhances the scalability and performance of our system. Through rigorous validation experiments, we have demonstrated the effectiveness and accuracy of our recommendation model across various metrics, including intersection over union (IOU) and genre-based retrieval accuracy. Overall, our project lays the foundation for a more intuitive and engaging movie recommendation experience for Persian-speaking audiences, fostering deeper engagement with cinematic content and enhancing the overall movie-watching experience.

Acknowledgement:

To achieve our project goals, we purchased a Google Colab Pro account and obtained GPT credits.