



# MOVIE ASSISTANT CHATBOT

Data Mining | Software Engineering Master's Degree | Universidade do Minho

★ Abhimanyu Aryan PG51632 | Millena Santos PG54107 | Ricardo Oliveira PG54177 ★

# Table of contents

01

## Introduction

- Context
- Objectives
- Planning

02

## Technologies

03

## Implementation

- Data Extraction
- RAG pipeline

04

## Results

- User Interface
- Comparison with GPT4o

05

## Conclusion

- Limitations and Future Work



01

# Introduction

# Context

- Movies are a powerful medium of storytelling that have a profound impact on society. They entertain, educate, and inspire audiences, transcending geographical and cultural boundaries;
- Personalized recommendations help users discover movies that match their preferences, enhancing their viewing experience;
- Hollywood films often set trends in filmmaking and popular culture. The industry is renowned for producing a wide variety of genres, including action, drama, science fiction, and animation.
- Based in Mumbai, India, Bollywood is the largest film industry in the world in terms of the number of films produced and tickets sold.
- The cultural influence of Bollywood is immense, with its movies and stars being adored across South Asia and among the Indian diaspora worldwide.



# Objectives



**Better recommendation  
for movies based on  
more context**

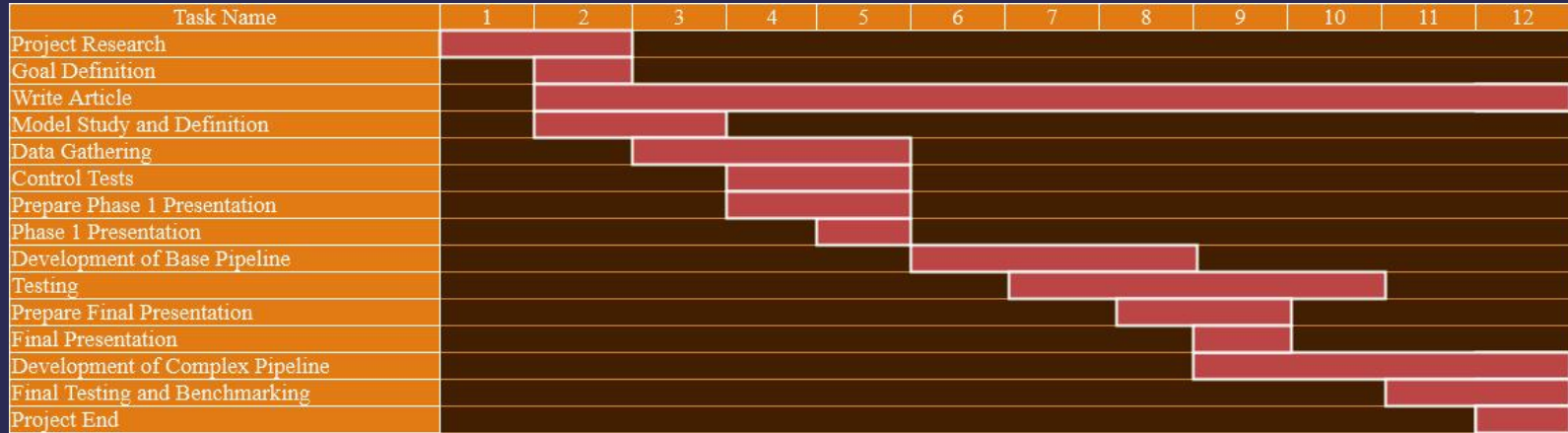


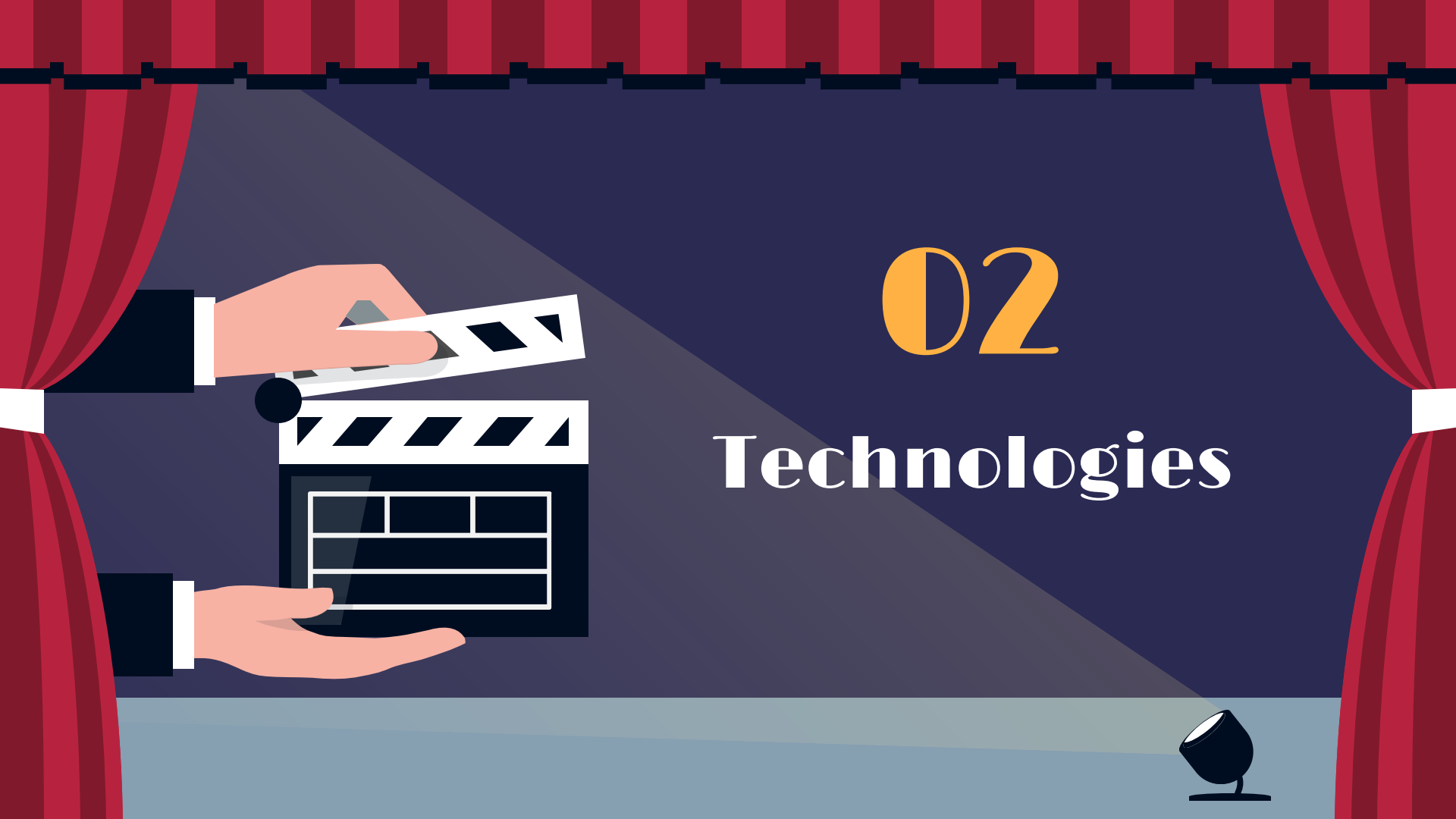
**Specific questions  
about movies**



**Up to date with  
the latest movies**

# Planning





02

Technologies

# Technologies



**Text-embedding  
ada-002 v2**



**CHATGPT 3.5  
Turbo 0125**



**Llama index**



The background is a stylized stage. At the top, there are red curtains with vertical stripes. A dark blue backdrop covers the upper part of the stage. A red podium with two microphones is positioned on the right. A spotlight from the bottom left illuminates the podium and the text. The floor is a mix of light blue and red. In the foreground, there are yellow stanchions with red ropes forming a path.

03

# Implementation

# Implementation **Process**



**Data Gathering**



**Information  
Indexing**



**Pipeline  
Definition**



**Test the LLM**



**Evaluate the  
Results**

# Data Gathering

- All movies from the 2020's section in the page *"Lists of Hindi films"*;
- All movies from the 2020's section in the page *"Lists of American films"*;
- Used Wikipedia API to extract all content from each movie's page;
- Contents saved as PDF.

# RAG Pipeline

- Prompt Engineering;
- Indexing and Embedding;
- Embedding Storage.

A stylized illustration of a woman with blonde hair, wearing a black off-the-shoulder gown and high heels, standing on a yellow star on a blue carpet. The scene is framed by red curtains on the left and right. A spotlight shines on the woman from the top left. Another spotlight is on the floor to the right. A stanchion with a red rope is visible in the background. The background is a dark blue gradient.

04

Results

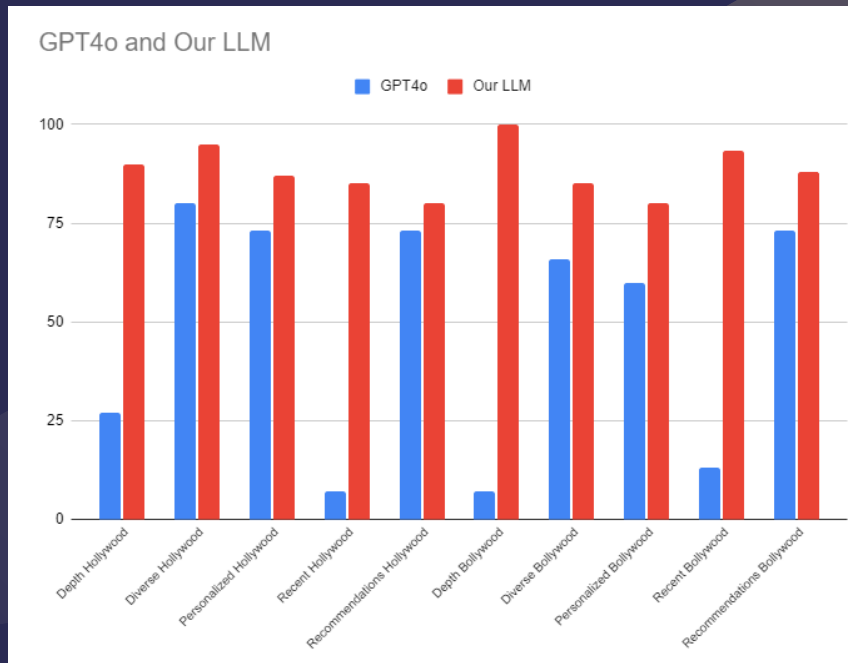
# User Interface

**Enter a description and get a list of movies!**

Example: A comedy movie set in new york city

Generate!

# Comparison with GPT-4o



# 05

## Conclusion





# Limitations and Future Work

- Difficulty in handling "tricky" questions;
- Evaluation Pipeline
  - Sentence-window retrieval;
  - Auto-merging retrieval.



# MOVIE ASSISTANT CHATBOT

Data Mining | Software Engineering Master's Degree | Universidade do Minho

★ Abhimanyu Aryan PG51632 | Millena Santos PG54107 | Ricardo Oliveira PG54177 ★