Movie Recommender Chat-bot

CS-583 Deep Learning

Ashwin Dhanasamy
adhanasa@stevens.edu

Shaun Mendes
smendes@stevens.edu

December 2023

Abstract

This paper presents the design and development of an advanced, responsive chatbot, specifically engineered to effectively respond to user inquiries with precise movie details. In addition to providing accurate information, the chatbot demonstrates exceptional capability in delivering personalized movie recommendations, offering a maximum of three suggestions tailored to an individual's movie preference. The paper elaborates on the intricate process involved in constructing our fine-tuning datasets utilizing the PaLM 2(LLM). Furthermore, we provide a comprehensive explanation of our training process, inclusive of the training arguments and various methodologies employed to streamline the training procedure.

Github: https://github.com/smendes2901/movie_chat_bot.git

1 Introduction

Hedonic media products, which are primarily consumed to fulfill emotional needs and encompass items such as movies and music, are increasingly being accessed via streaming services. These platforms offer a vast array of content to users. To mitigate the issue of information overload and the ensuing selection difficulties, media streaming services like Netflix for films and Spotify for music employ algorithmic recommender systems to generate personalized content suggestions. These recommendations are predominantly based on the user's historical content consumption and the preferences of users with similar tastes.

However, the specific needs of users concerning hedonic media products are not solely reliant on historical preferences. They are particularly influenced by the user's current mood, emotions, and situation. Unlike traditional video or music stores where customer decisions were typically guided by sales staff, media streaming services lack interpersonal interactions. Text-based conversational agents, which interact with users through natural language and provide personalized content recommendations, may offer innovative solutions to these challenges.

Our objective was to construct a movie recommendation chat-bot capable of responding to user's natural language queries about various movie details, such as genre and storyline, available in a given context (e.g., IMDB Description of the movie). The chat-bot is also designed to recommend movies to the user when provided with a movie of interest to the user. To accomplish this, we developed a model proficient in executing these two distinct yet interconnected tasks.

2 Contribution

2.1 Ashwin Dhanasamy

- Prompt-Engineering to build the data corpus for train, validation and test sets for LLM and Retrieval Model
- Engineering model metrics
- Ran experiments with different training and LoRA hyper-parameters
- Presentation Slides and Report

2.2 Shaun Mendes

- Generating training and testing scripts for LLM and Retrieval Model
- Consolidating inference scripts into a single E2E workflow
- Ran experiments with different training and LoRA hyper-parameters
- Presentation Slides and Report

3 Method

We exclusively utilized movies from the IMDB datasets for this project. Specific data, including the movie title, runtime, genre, and release year, were extracted from the title.base.tsv file, while the rating and number of votes were obtained from the title.rating.tsv file. Due to the limited information available for each movie in the provided datasets, we supplemented our dataset by scraping the IMDB database to extract the description and plot summary for each movie.

Following the data crawling process, a data blob, encompassing the aforementioned details, was integrated into prompts and supplied to Generative Pretrained Transformer (GPT), Palm2, and Command to generate question-answer pairs for each movie. These pairs were subsequently partitioned into training, validation, and testing sets, which were utilized to fine-tune the Language Model (LLM) following the training format delineated in the Stanford Alpaca Repository. To enhance training efficiency, the LLM was trained using QLoRA on a

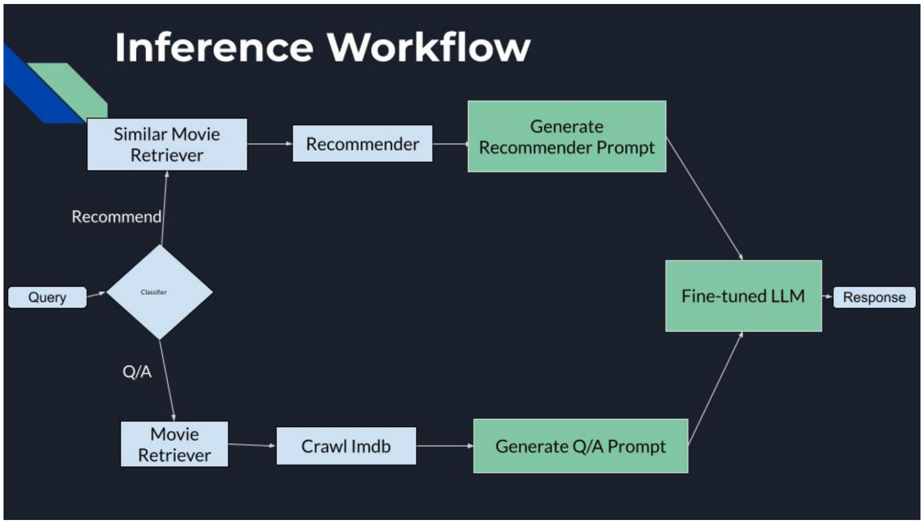

Figure 1: Inference Workflow

V100 GPU on Colab for a duration of 30 minutes. The output of the LLM was subsequently validated through human annotations.

Furthermore, we employed Retrieval Augmented Generation to improve the model's generalization beyond its training dataset and enhance its accuracy. The retriever was fine-tuned to optimize the extraction of reference text for the LLM. The output of the retriever was evaluated using cosine similarity and F1 score metrics.

4 Datasets

We utilized the IMDB Non-Commercial Dataset(https://datasets.imdbws.com) to obtain details of movies released between 2019 and 2022. Additionally, we performed web scraping on the IMDB website to extract movie descriptions. This comprehensive dataset, encompassing various aspects of each movie, enabled us to generate question-answer pairs for the purpose of fine-tuning our models.

In order to optimize the dataset for training, we applied a filter to exclude data from before 2017 and after 2020. We extracted the following columns from the IMDB datasets: movie title, runtime, genre, and release year from title.base.tsv, and rating and number of votes from title.rating.tsv. Due to the lack of sufficient information for each movie in the provided datasets, we further scraped the IMDB database to extract the description and plot summary for each movie.

The training process involved three distinct models.

- A retrieval model that provides 20 recommended movies to the Language

Model (LLM) prompt.

- A retrieval model that maps a query to its corresponding IMDB movie.
- Fine-tuned LLM to answer user queries.

4.1 Question and Answer(QnA) Dataset

To generate this dataset, we used the PaLM API to pass information about the movie to the PaLM2 LLM to generate question and answer pairs. We provide an example of question answer pairs and combine the IMDB information into a well crafted prompt and provide it to PaLM2.

```
Instruction: Create a maximum of 3 question and answer pairs that capture all the
information in the given movie description, including the name of the movie
in each question.
Movie: Movies: Mr. Fukyô vs eiga-tachi
Description: A rental video shop, whose customers are few and afar, is visited by
a mysterious man, Mr. Fukyo (meaning Mr. Recession). This man recommends closing
the business. To protect the beloved shop, Tatsuya, the manager's son and
successor, confronts the villain. He is joined by Mr. Violence, Mr. Human Drama,
Mr. Horror and Miss Love Story; spirits of each movie genre. Mr. Fukyo,
unable to compete against the spirits, brainwashes Mr. Violence into
eradicating the movies. To rescue Mr. Violence from Mr. Fukyo's brainwashing,
Tatsuya and the spirits will need to fight off Mr. Fukyo's forces. Can Tatsuya
and the 'movie gurus' save their beloved shop? These spirits now appear 120
years after the creation of the first movie! Protect the video store!!!
Q&A pairs:
[
{
     "question":"What is the main conflict in the movie \Mr. Fukyô vs eiga-tachi"?",
     "answer":"The main conflict in the movie \Mr. Fukyô vs eiga-tachi" is between
     Tatsuya, the manager's son and successor of a rental video shop, and a
     mysterious man named Mr. Fukyo who recommends closing the business. Tatsuya is
     joined by spirits of each movie genre to protect the shop."
},
{
     "question":"Who are the characters that join Tatsuya in his fight against Mr.
     Fukyo in the movie \Mr. Fukyô vs eiga-tachi"?",
     "answer":"In the movie \Mr. Fukyô vs eiga-tachi", Tatsuya is joined
     by Mr. Violence, Mr. Human Drama, Mr. Horror, and Miss Love Story, who are
     spirits of each movie genre."
{
     "question":"What is the climax of the movie \Mr. Fukyô vs eiga-tachi"?",
     "answer":"The climax of the movie \Mr. Fukyô vs eiga-tachi" is when
     Tatsuya and the spirits fight off Mr. Fukyo's forces to rescue Mr. Violence
     from Mr. Fukyo's brainwashing and save their beloved shop."
```

```
},
]
Instruction: Create a maximum of 3 question and answer pairs that capture all
the information in the given movie description, include the name of the
movie in each question.
Movie: Thelivu
Description: Description: Gouri, an orphan who marries truck driver Khalid
and seems to have meaning in her life. However, she is forced to kill the
owner of the orphanage she works in after realising that he was trying to
rape one of the inmates. The story takes off from there with Gouri and Khalid
lives changing overnight.
Release Year: 2019
Runtime(in minutes): 109
Genre: Crime,Drama,Family
Rating: 5.4
Votes: 76.0
Q&A pairs:
```

The prompt utilized incorporates an example with the objective of enhancing the accuracy of the model in generating question-answer (QA) pairs. The robustness of the model's generation is directly proportional to the number of examples included in the prompt. In our methodology, we incorporated up to three movie examples along with their corresponding responses. For the exact prompt, please refer to the script "palm_process_3.py".

At present, Palm2 is only available in a public preview stage, which results in inconsistent output. Following the extraction process, a significant amount of data cleaning and augmentation were necessitated to generate a sufficiently reliable output for training the Language Model (LLM).

Several challenges were encountered during this process, including:

- The presence of duplicate questions with varying answers
- Malformed JSON data
- Absence of the movie name in the question
- Generation of out-of-context QA pairs

Upon addressing these issues, we processed approximately 7,000 movies, resulting in the generation of around 10,000 movie QA pairs post data cleaning.

4.2 Recommendation Dataset

To generate movie recommendations, we used the following prompt

```
Given that I like the movie 'Imprisoned (2018)', recommend me 2 movies.
The output should be in a list. For eg. {[movie1, movie2,...]}
```

```
Each movie should have the release year attached to it. For eg. Imprisoned (2018)
Each movie should be in order of their relevance
```

The quantity of recommendations was adjusted for each new movie. It was observed that Palm2 infrequently generated movies that were aligned with the reference. However, upon examination, it was found that the relevance list often contained films that were highly relevant to the first generated movie. As a result, the recommendation list was constructed based on the initial recommended movie. This process was executed for approximately 5,000 movies. The results generated from the Question-Answer and Recommendation dataset were subsequently merged and utilized for the fine-tuning of the Language Model (LLM).

4.3 Question to Movie Dataset

For a more realistic inference process it was necessary to procure information pertaining to movies that were not included in our training dataset. To achieve this we needed the IMDB URL corresponding to each question. To accomplish this task, we trained a retriever model designed to map a question to the respective movie in the IMDB dataset.

4.4 Movie to Recommendation Dataset

Upon extraction of the movie, it is merged with the corresponding information present in the IMDB dataset. Subsequently, a search string is formulated based on this integrated data, which is then employed to retrieve 'k' number of relevant movies. The following are a few representative rows utilized for this task. The rows were categorized into positive or negative samples based on the output from the Recommendation dataset.

5 Experiments

5.1 Tools & Hardware

For all the models, we trained them on V100 GPUs that we had access to through a Google Colab Pro subscription. We reduced the training computation required by using Q-LoRA, one of the Peft(Parameter Efficient Fine-Tuning) methods from HuggingFace.

5.2 Retriever Fine-Tuning

We trained a Sentence Transformer model using a dataset comprising 163,098 instances for training and 4,032 instances for validation. The model was initialized with the pretrained weights from the "sentence-transformers/all-MiniLM-L6-v2" model. The training process utilized the ContrastiveLoss function, which is particularly effective for learning sentence embeddings. The training parameters were set as follows: a batch size of 32 was used to ensure manageable memory usage and efficient gradient updates. The warm-up steps, which are crucial for stabilizing the learning rate in the initial phase of training, were calculated as 10

5.3 LLM Fine-Tuning

To build our chat-bot, we fine-tune Llama originally pre-trained as "meta-llama/Llama-2-70b-hf". We used a dataset comprising 18k instances for training, 2k instances for validation and 2k instances for the test set. The training parameters were meticulously chosen to optimize the learning process. We set the maximum token length to 1024 and used a batch size of 8. The learning rate was set to 0.0003, and we employed a "cosine" learning rate scheduler. The model was trained for a maximum of 1 epoch using the Adam optimizer. Additionally, a warmup ratio of 0.05 was used to gradually increase the learning rate at the beginning of training. We used the LoRA method to greatly reduce the number of trainable parameters, using only 0.06% of the total number of parameters of the 7 billion parameters present in the LLM. For the LoRA parameters, we set LORA_R to 32, LORA_ALPHA to 16 and LORA_DROUPOUT to 0.05 We experimented with different combinations of LORA Target modules and found the best responses to be when we used the combination of q_proj, v_proj, k_proj, o_proj, gate_proj, down_proj and up_proj.

6 Results

We evaluated our retriever models based on the Cosine Similarity between the original output and the output of the respective retriever. The output of our LLM was evaluated manually by randomly choosing 100 samples. Sample outputs are available in Figure 2 and Figure 3.

Model	Metric	Result
IMDB Movie Retreiver	Cosine Similarity	0.8356
Top k Retreiver	Cosine Similarity	0.7531
LLM Output	Manual Validations	0.82

Table 1: Results

7 Conclusion

In this paper, we have presented a movie chatbot that can engage users in conversations about movies and their descriptions. Our chatbot is based on the Llama 2 model, which we fine-tune on a domain-specific dataset of question-answer pairs generated from movie details. We also enhance our chatbot's abil-

```
Enter your query: Can you explain what happens in the movie 'Harry Potter and the Goblet of Fire 2005' in detail?
Below is a question regarding movies and shows paired with an input that provides further context. Write a response that appropriately completes the request.
###Instruction: Can you explain what happens in the movie 'Harry Potter and the Goblet of Fire 2005' in detail?
###Input:
Description: Harry's (Daniel Radcliffe's) fourth year at Hogwarts is about to start and he is enjoying the summer vacation with his friends. They get the tickets
 to The Quidditch World Cup Final, but after the match is over, people dressed like Lord Voldemort's (Ralph Fiennes') "Death Eaters" set a fire to all of the vis
itors' tents, coupled with the appearance of Voldemort's symbol, the "Dark Mark" in the sky, which causes a frenzy across the magical community. That same year,
Hogwarts is hosting "The Triwizard Tournament", a magical tournament between three well-known schools of magic : Hogwarts, Beauxbatons, and Durmstrang. The conte
stants have to be above the age of seventeen, and are chosen by a magical object called "The Goblet of Fire". On the night of selection, however, the Goblet spew
s out four names instead of the usual three, with Harry unwittingly being selected as the Fourth Champion. Since the magic cannot be reversed, Harry is forced to
 go with it and brave three exceedingly difficult tasks.,
Director(s): Mike Newell
Release Year: 2005
Rating: 7.7
Runtime(in minutes): 157

###Response:
Harry's fourth year at Hogwarts is about to start and he is enjoying the summer vacation with his friends. They get the tickets to The Quidditch World Cup Final,
 but after the match is over, people dressed like Lord Voldemort's "Death Eaters" set a fire to all of the visitors' tents, coupled with the appearance of Voldem
ort's symbol, the "Dark Mark" in the sky, which causes a frenzy across the magical community. That same year, Hogwarts is hosting "The Triwizard Tournament", a m
agical tournament between three well-known schools of magic : Hogwarts, Beauxbatons, and Durmstrang. The contestants have to be above the age of seventeen, and a
re chosen by a magical object called "The Goblet of Fire". On the night of selection, however, the Goblet spews out four names instead of the usual three, with H
arry unwittingly being selected as the Fourth Champion. Since the magic cannot be reversed, Harry is forced to go with it and brave three exceedingly difficult t
asks.
```

Figure 2: Q&A Sample Result

```
Enter your query: Recommend movies similar to The Conjuring

Below is a question regarding movies and shows paired with an input that provides further context. Write a response that appropriately completes the request.
###Instruction: Given the movie in The Conjuring (2013), recommend 3 similar movies for the input
###Input: The Conjuring (2013)
Poltergeist (2013)
The Haunting of Pearson Place (2013)
The Conjuring 2 (2013)
Sinister Visions (2013)
Insidious (2013)
The Autopsy of Jane Doe (2013)
Poltergeist Activity (2013)
Eerie (2013)
The Amityville Asylum (2013)
The Blair Witch Project (2013)
Poltergeist Encounters (2013)
Exorcismo Documentado (2013)
A Haunting in Ravenwood (2013)
Exorcism: Haunted Child (2013)
An Irish Exorcism (2013)
The Haunting in Connecticut (2013)
Dark Exorcism (2013)
Haunting of the Innocent (2013)
Sinister (2013)
###Response:

The Conjuring 2 (2013)
The Haunting of Pearson Place (2013)
The Autopsy of Jane Doe (2013)
```

Figure 3: Recommendation Sample Result

ity to provide accurate responses by using Retrieval Augmented Generation, which leverages external knowledge sources. Our chatbot demonstrates promising results in terms of fluency, relevance, and diversity of responses. We believe that our chatbot can be a useful and entertaining tool for movie enthusiasts and researchers alike.