Problem

In our bot, we aim to identify the genre of a movie based on its summary. To achieve this, we use an architecture that takes a paragraph summarizing a movie as input and outputs the movie genre, selecting from one of the following categories: "action," "family," "romance," "crime," or "fantasy."

Architecture

For this problem, we need LLM models and a task-relevant dataset. We fine-tune the LLM models using our task-oriented dataset.

Models

For this purpose, we chose two types of models:

1. Gpt-3.5-turbo (openai api)
2. Llama-2-7b-chat-hf

Dataset

For this purpose, we chose two types of datasets:

1. A generated dataset using the GPT-4 API.
2. A real-world dataset from Kaggle.

Type1- Dataset generation Using GPT4 API (movie_genre_API.csv)

First, create an account on OpenAI at https://platform.openai.com. After that, obtain API access, which requires a minimum billing amount of $5. From the API we can generate data set.

Input parameters are

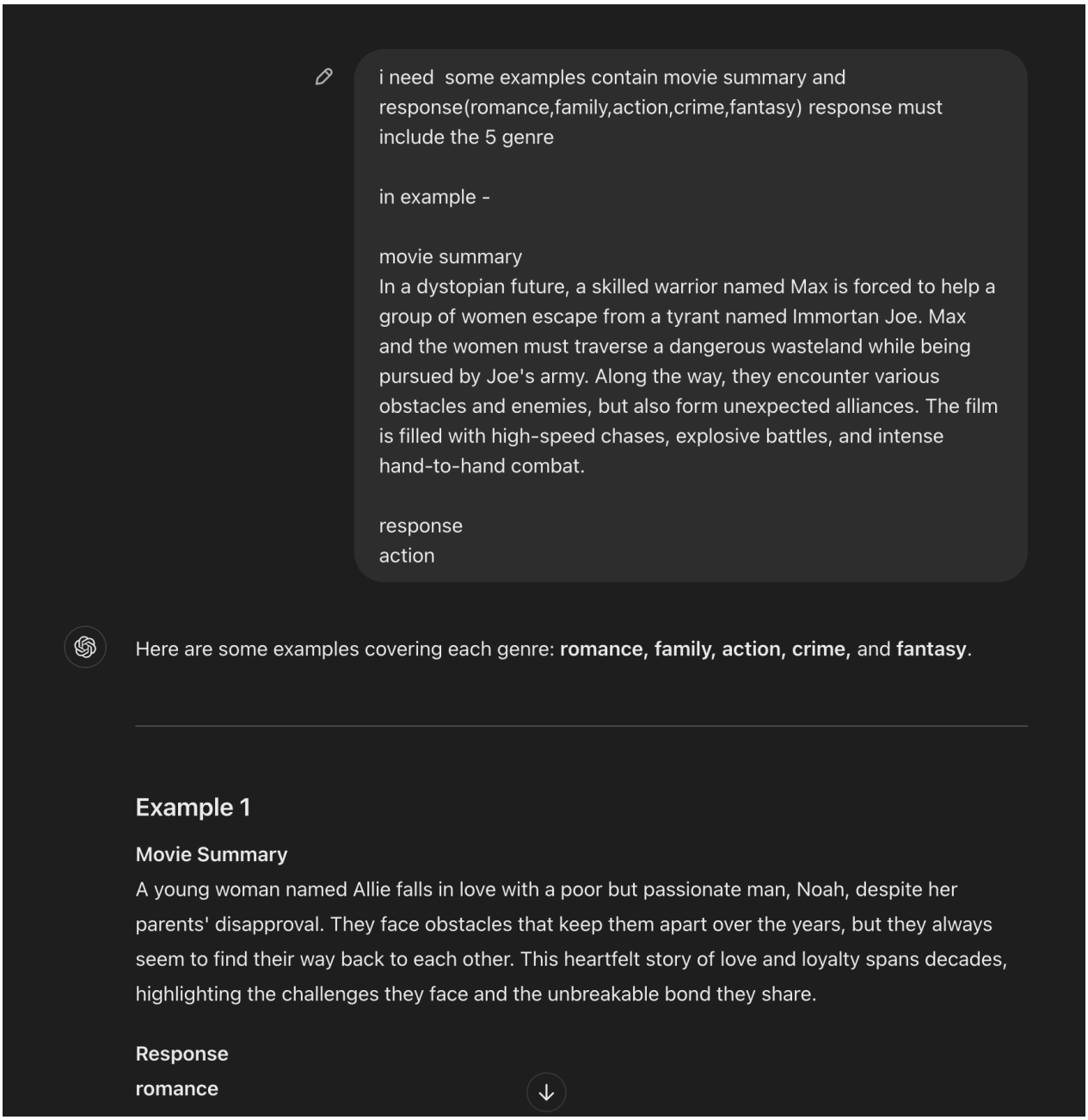
1. prompt = 'A model that takes a paragraph summarizing a movie as input and, based on the summary, outputs the movie genre as one of the following: "action," "family," "romance," "crime," or "fantasy."'
2. temperature = 0.4
3. number_of_examples = 68

So we generate 68 summary and response pair.

Note: Generating a dataset using the API can be somewhat costly.

Type2- Dataset generation Using chatgpt - prompt engineering | one shot

The dataset can also be generated using ChatGPT (open source) by applying prompt engineering; a one-shot approach has proven effective.



Example 1

Movie Summary

A young woman named Allie falls in love with a poor but passionate man, Noah, despite her parents' disapproval. They face obstacles that keep them apart over the years, but they always seem to find their way back to each other. This heartfelt story of love and loyalty spans decades, highlighting the challenges they face and the unbreakable bond they share.

Response - romance

Example 2

Movie Summary

An animated adventure follows the story of a young lion cub, Simba, who is destined to be king. After a tragic event orchestrated by his uncle, Simba runs away, only to later return as an adult to reclaim his throne. This family-friendly story is filled with songs, colorful characters, and lessons about courage, responsibility, and the importance of family.

Response - family

Type 3- Dataset collection from kaggle (movie_genre_kaggle.csv)

Models comparison

	GPT3.5-turbo	LLAMA2
parameters	~175 billion	7 billion
Performance	high-quality generation, excels in NLP tasks.	Good for NLP, but may be less fluent
Fine-tuning	relies on few-shot prompting	Fully supports fine-tuning for customization
Cost	API-dependent	Deployable locally
Hardware Requirements	Requires OpenAI's API	Can run on local GPUs(PEFT)

Result comparison

The total dataset contains 68 entries. Approximately 80% of the data will be used for training, and 20% will be used for testing.

	movie_summary	Ground Truth	Predicted
0	The movie follows the story of a former Green ...	action	action
1	In this film, a young wizard named Harry Potte...	fantasy	fantasy
2	The movie is about a group of astronauts who e...	action	action
3	The film is set in a world where magic is real...	fantasy	fantasy
4	The film tells the story of a young woman name...	romance	romance
5	The film follows the journey of a hobbit named...	fantasy	fantasy
6	The movie follows the life of a notorious mobs...	crime	crime
7	The film revolves around a young woman named A...	romance	romance
8	The film follows the life of a young boy named...	family	family
9	The film revolves around a young woman named B...	romance	romance
10	The film tells the story of a young woman name...	romance	family
11	The film tells the story of a young woman name...	romance	romance
12	The film is about a young woman who is swept o...	romance	romance
13	The film follows the life of a notorious mobst...	crime	crime

	content	ground truth	predicted
0	The movie follows the story of a former Green ...	action	action
1	In this film, a young wizard named Harry Potte...	fantasy	fantasy
2	The movie is about a group of astronauts who e...	action	Fantasy
3	The film is set in a world where magic is real...	fantasy	fantasy
4	The film tells the story of a young woman name...	romance	fantasy
5	The film follows the journey of a hobbit named...	fantasy	fantasy
6	The movie follows the life of a notorious mobs...	crime	crime
7	The film revolves around a young woman named A...	romance	romance
8	The film follows the life of a young boy named...	family	family
9	The film revolves around a young woman named B...	romance	fantasy
10	The film tells the story of a young woman name...	romance	fantasy
11	The film tells the story of a young woman name...	romance	fantasy
12	The film is about a young woman who is swept o...	romance	fantasy
13	The film follows the life of a notorious mobst...	crime	crime

- For a sample of 14 data points, GPT-3.5 Turbo made 13 correct predictions and 1 incorrect prediction. In comparison, the fine tuned LLaMA-2 model (using PEFT) made 8 correct predictions and 6 incorrect predictions.
- With a total of 68 data points divided into 5 categories, each category has approximately 13 data points. In the test case, the "romance" category has 6 data points (ground truth), while the training set contains only 7 data points for "romance." This limited representation of "romance" features may have affected the model's ability to learn these features effectively. Consequently, both GPT-3.5 Turbo and LLaMA-2 misclassified examples from the "romance" category. A review of the category distribution in the training set is recommended.
- In the LLaMA-2 inference, many "romance" instances were classified as "fantasy." This misclassification may be due to the low distance between "romance" and "fantasy" vectors in the 5-dimensional vector database, causing the model to confuse the two categories during prediction.
- The prediction accuracy of GPT-3.5 Turbo is comparatively higher than that of LLaMA-2. To enhance LLaMA-2's accuracy, it requires additional data—approximately 100 to 150 more samples. A similar approach applied to the Kaggle dataset has previously resulted in improved accuracy for LLaMA-2.