Movie Recommender Chat - Bot

Ashwin Dhanasamy
Shaun Mendes

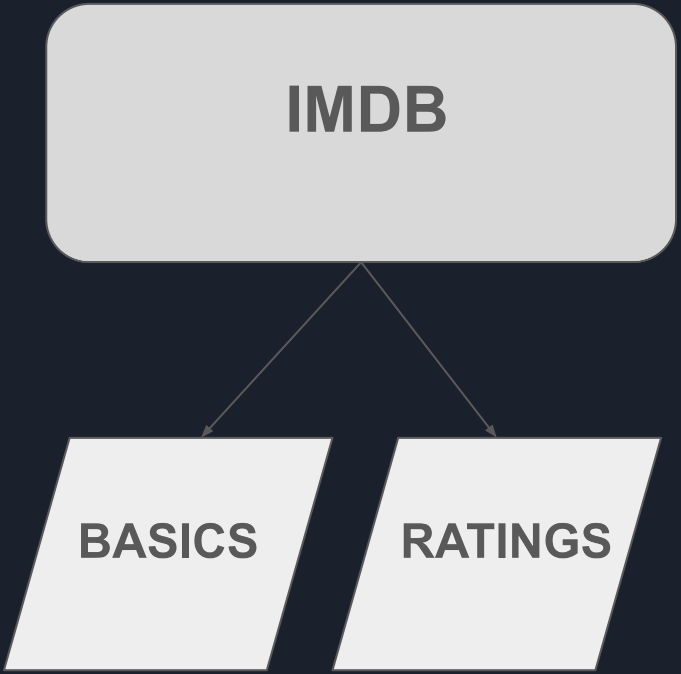
Movie Recommender Chatbot

- A chatbot that can engage in natural language conversations about movies.
- It can handle various types of queries related to films, such as genre, plot, cast, ratings, reviews, and trivia.
- It can also suggest movies that match the user’s preferences or interests

Tools & Technologies

- Google Colab - T4, V100
- Nvidia RTX 3050 Laptop GPU
- Llama 2/OpenLlama2
- PaLM: https://developers.generativeai.google/products/palm
- Cohere: Command
- ChatGPT-3.5-Turbo-Instruct

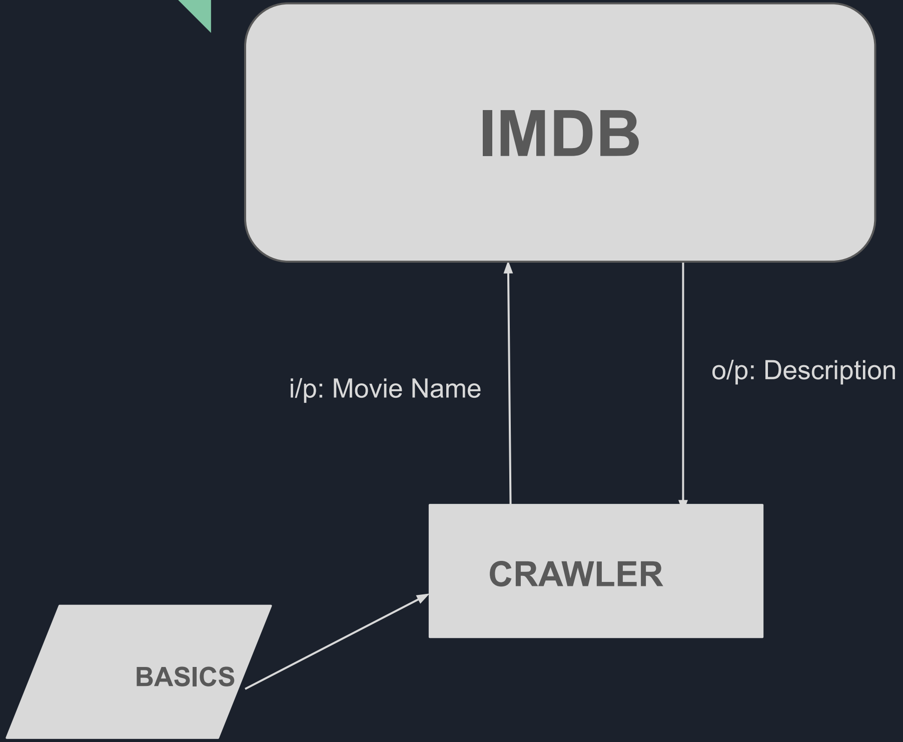
Dataset (1/3)



IMDB Non-Commercial Dataset (2019-2022)

- 35K Movies
- Basics
  - id,originalTitle,isAdult,startYear,
  - endYear,runtimeMinutes,genres
- Ratings
  - id,averageRating,numVotes

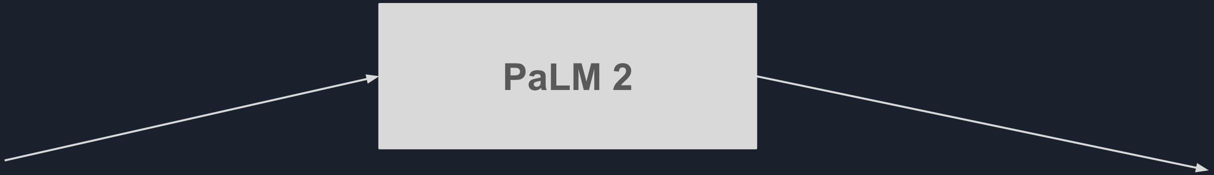
Dataset (2/3)



- Name: Photo de famille
- Year: 2018
- Genres: Comedy, Drama
- Runtime: 98
- Average Rating: 5.8
- numVotes: 826
- Description:

“Gabrielle is a "statue" for tourists, much to the chagrin of her teenage son. Elsa is in angry at the world and desperate to become pregnant. Mao is a chronically depressed video game designer who drowns his melancholy in alcohol and psychoanalysis. They are brother and sisters but do not hang out. Ever. Their parents Pierre and Claudine, separated for a long time, have really done nothing to strengthen the bonds of the family - yet, at their grandfather's funeral, they are going to have to meet, and together answer the question: "What to do with grandma?"—Hugo Van Herpe”

Dataset (3/3)

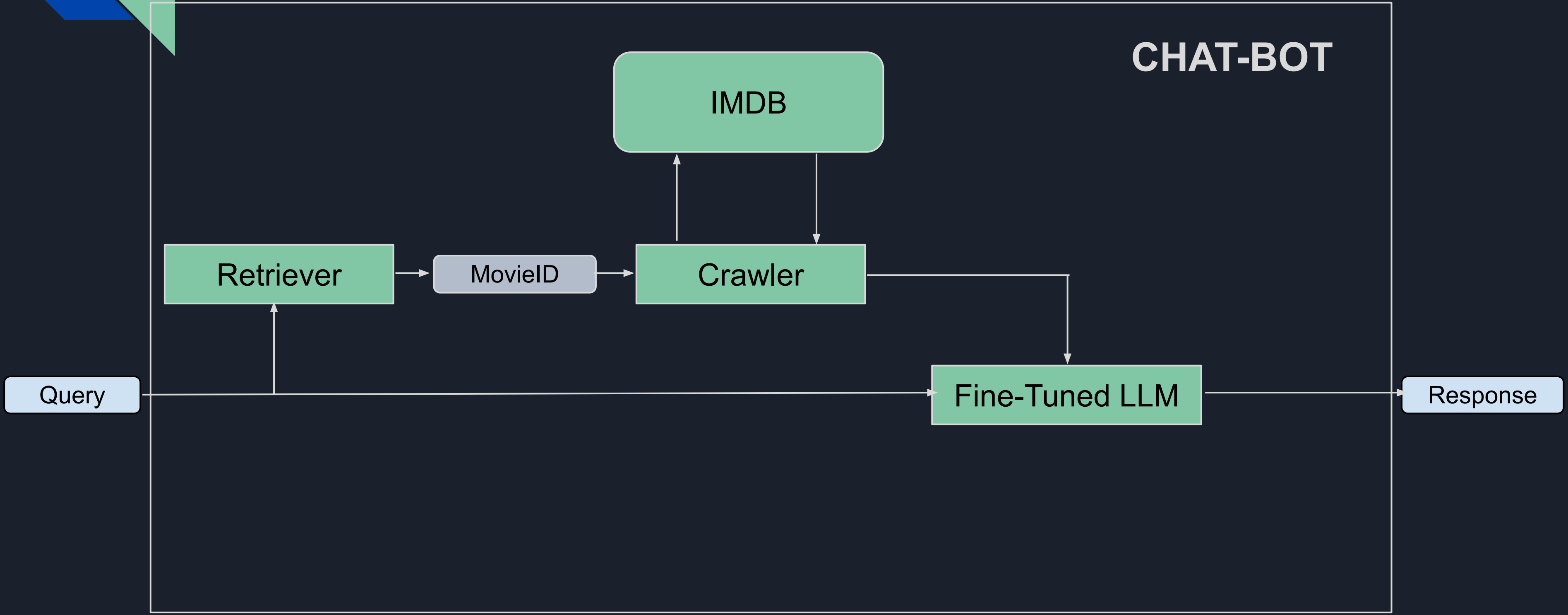


- Name: Photo de famille
- Year: 2018
- Genres: Comedy, Drama
- Runtime: 98
- Average Rating: 5.8
- numVotes: 826
- Description:

“Gabrielle is a "statue" for tourists, much to the chagrin of her teenage son. Elsa is in angry at the world and desperate to become pregnant. Mao is a chronically depressed video game designer who drowns his melancholy in alcohol and psychoanalysis. They are brother and sisters but do not hang out. Ever. Their parents Pierre and Claudine, separated for a long time, have really done nothing to strengthen the bonds of the family - yet, at their grandfather's funeral, they are going to have to meet, and together answer the question: "What to do with grandma?"”

```
{
  "question": "What is the movie Photo de famille about?",
  "answer":"The movie Photo de famille is about a family who reunites after the death of their grandfather to decide what to do with their grandmother."
},
{
  "question": "Which location(s) are featured in the movie Photo de famille?",
  "answer":"The movie is set in Brussels, Belgium."
},
{
  "question": "What is the tone, main themes and genres of the movie Photo de famille?",
  "answer":"The movie is a comedy-drama that explores the themes of family relationships, loss, and grief."
},
{
  "question": "What is the ending of the movie Photo de famille?",
  "answer":"The family is able to come to a consensus on what to do with their grandmother."
},
```

Inference Workflow



Retriever (1/2)

Task:

“What is the only refuge the characters have in Diablo Rojo PTY?“

“Diablo Rojo”

Retriever (2/2)

sentence-transformers/all-MiniLM-L6-v2: Maps sentences & paragraphs to a 384 dimensional dense vector space.

- Individual word embeddings for every word -> Embeddings for individual movies (multiple words)
- Given a question -> Search vector space -> Retrieve movie name

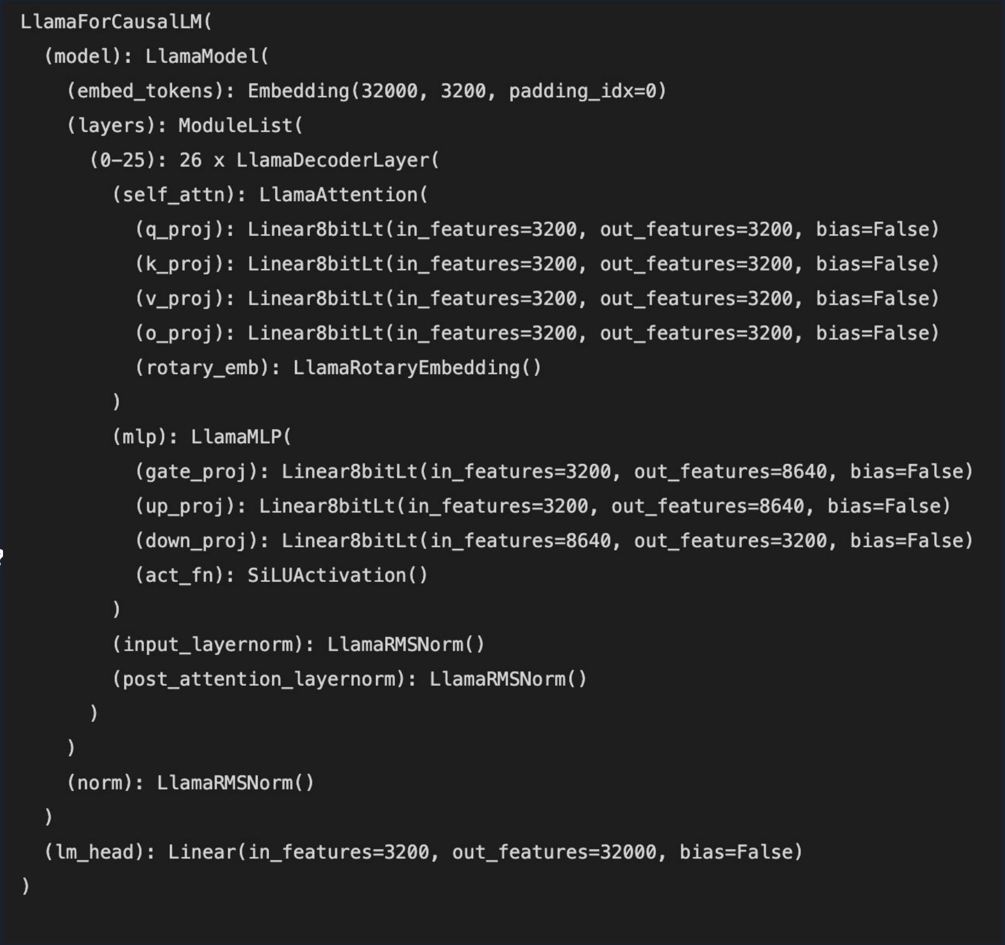
Chatbot (1/4)

openlm-research/open_llama_3b_v2

Task:

“What is the only refuge the characters have in Diablo Rojo PTY?

<Correct Answer>

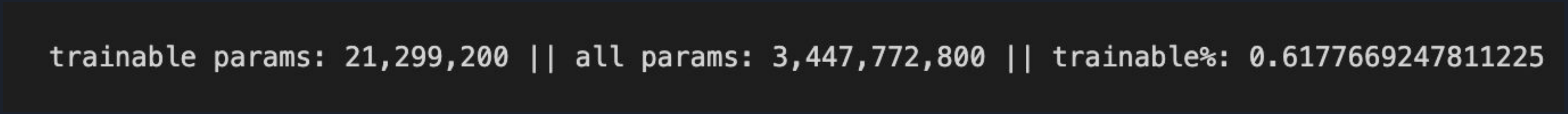


```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(32000, 3200, padding_idx=0)
    (layers): ModuleList(
      (0-25): 26 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (k_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (v_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (o_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear8bitLt(in_features=3200, out_features=8640, bias=False)
          (up_proj): Linear8bitLt(in_features=3200, out_features=8640, bias=False)
          (down_proj): Linear8bitLt(in_features=8640, out_features=3200, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): LlamaRMSNorm()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
  )
  (lm_head): Linear(in_features=3200, out_features=32000, bias=False)
)
```

Chatbot (2/4)

PEFT (Parameter-Efficient Fine-Tuning) Approach : To reduce number of parameters to train

- **LoRA: Lo**w-**R**ank **A**daptation of Large Language Models



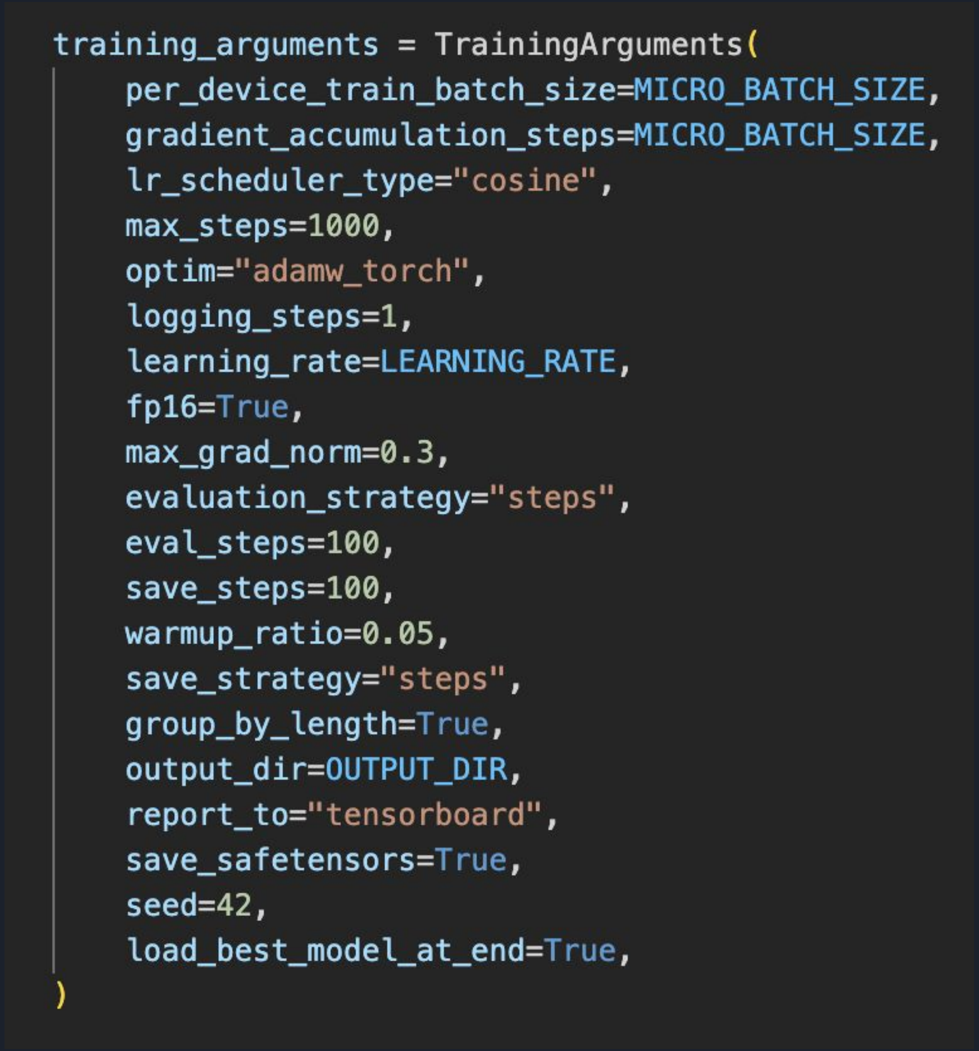
```
trainable params: 21,299,200 || all params: 3,447,772,800 || trainable%: 0.6177669247811225
```

Chatbot (3/4)

Training Arguments:

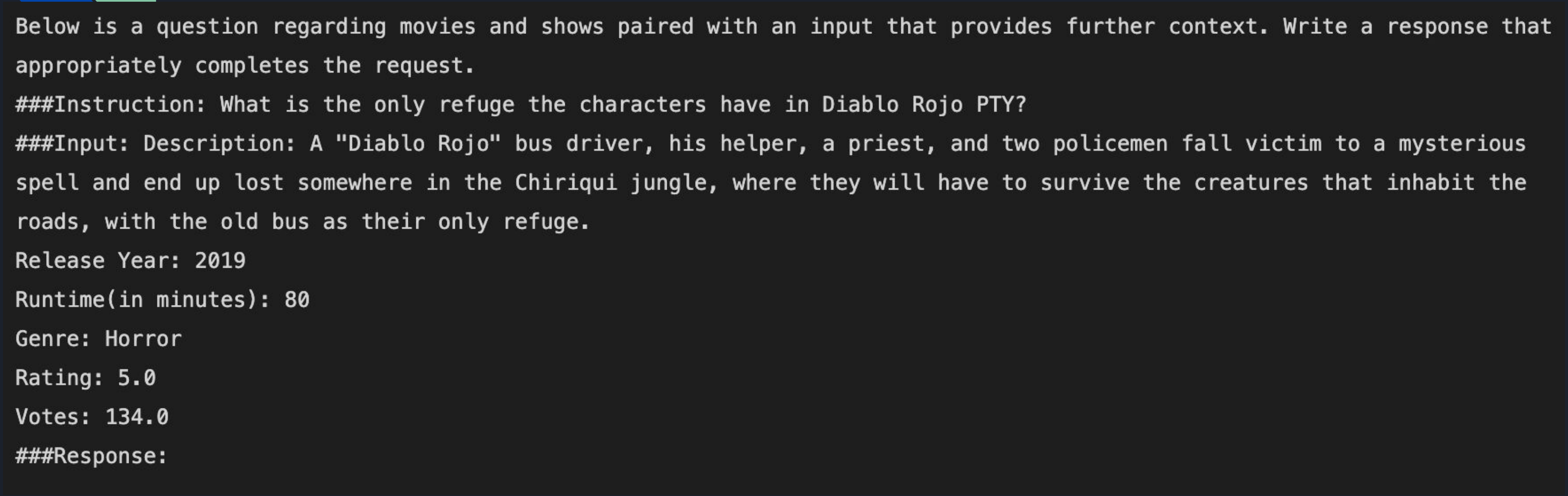
MICRO_BATCH_SIZE = 4

LEARNING_RATE = 3e-4

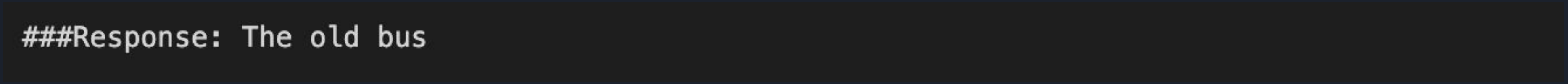


```
training_arguments = TrainingArguments(
    per_device_train_batch_size=MICRO_BATCH_SIZE,
    gradient_accumulation_steps=MICRO_BATCH_SIZE,
    lr_scheduler_type="cosine",
    max_steps=1000,
    optim="adamw_torch",
    logging_steps=1,
    learning_rate=LEARNING_RATE,
    fp16=True,
    max_grad_norm=0.3,
    evaluation_strategy="steps",
    eval_steps=100,
    save_steps=100,
    warmup_ratio=0.05,
    save_strategy="steps",
    group_by_length=True,
    output_dir=OUTPUT_DIR,
    report_to="tensorboard",
    save_safetensors=True,
    seed=42,
    load_best_model_at_end=True,
)
```

Chatbot (4/4)



```
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: What is the only refuge the characters have in Diablo Rojo PTY?
###Input: Description: A "Diablo Rojo" bus driver, his helper, a priest, and two policemen fall victim to a mysterious
spell and end up lost somewhere in the Chiriqui jungle, where they will have to survive the creatures that inhabit the
roads, with the old bus as their only refuge.
Release Year: 2019
Runtime(in minutes): 80
Genre: Horror
Rating: 5.0
Votes: 134.0
###Response:
```



```
###Response: The old bus
```

Future Scope

- Generate better training data
- Use a more powerful LLM
- Integrate Chain-of-Thought(CoT)

Conclusion

- We have successfully implemented a movie chatbot that can engage users in conversations about movies. 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 ability 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.