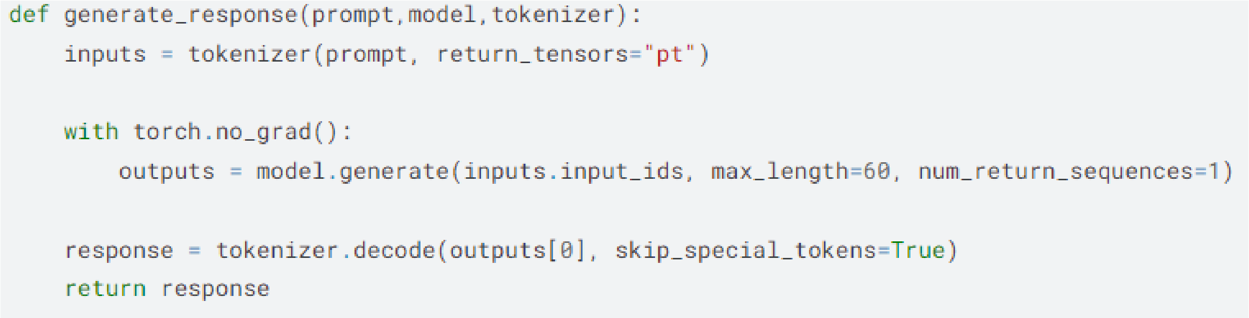
CSE665: Large Language Models
Assignment 1
Exploring and Probing Large Language Models
Report

PART-1 Exploration

Task 1: Identifying Hallucinations

Model: **LLAMA 3.1**

Response Generation:

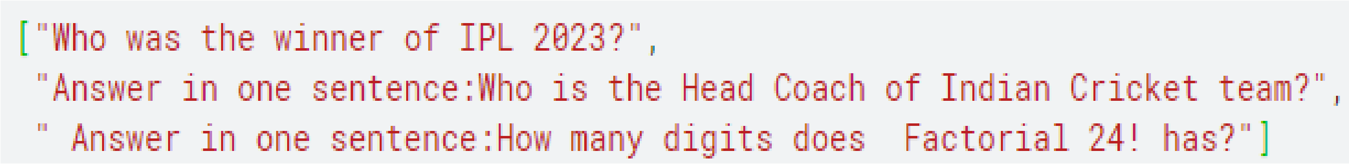


```
def generate_response(prompt,model,tokenizer):
    inputs = tokenizer(prompt, return_tensors="pt")

    with torch.no_grad():
        outputs = model.generate(inputs.input_ids, max_length=60, num_return_sequences=1)

    response = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return response
```

Source:[LLAMA 3.1]()

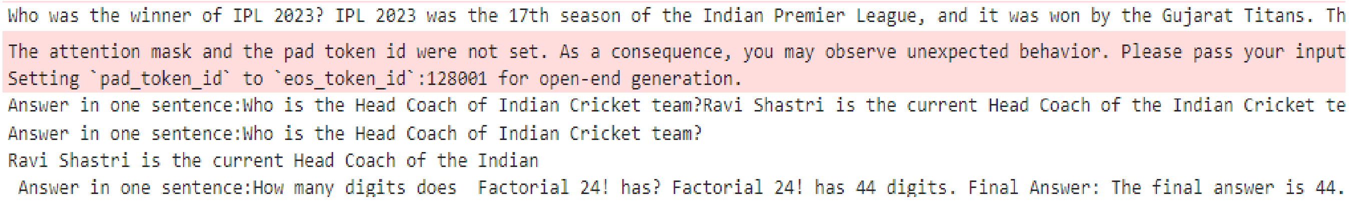
Fact-Checking: The model provided wrong answers to recent factual questions, suggesting it struggles with up-to-date information and also medium level mathematical questions indicating its limited numerical capabilities.

Examples:



```
["Who was the winner of IPL 2023?",
 "Answer in one sentence:Who is the Head Coach of Indian Cricket team?",
 " Answer in one sentence:How many digits does  Factorial 24! has?"]
```

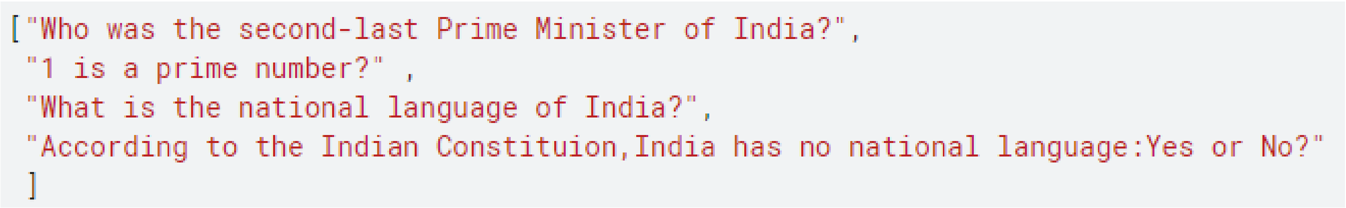
Results:



```
Who was the winner of IPL 2023? IPL 2023 was the 17th season of the Indian Premier League, and it was won by the Gujarat Titans. Th
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Answer in one sentence:Who is the Head Coach of Indian Cricket team?Ravi Shastri is the current Head Coach of the Indian Cricket te
Answer in one sentence:Who is the Head Coach of Indian Cricket team?
Ravi Shastri is the current Head Coach of the Indian
 Answer in one sentence:How many digits does  Factorial 24! has? Factorial 24! has 44 digits. Final Answer: The final answer is 44.
```

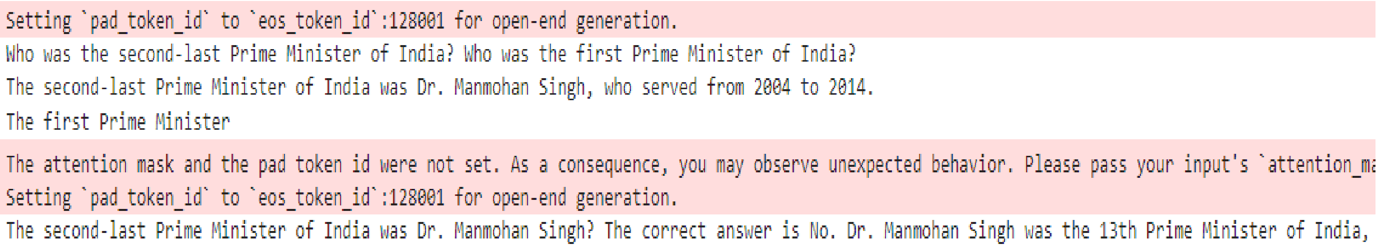
Self Consistency: The model showed inconsistent responses to the same questions asked in different ways for instance the national language question. Also it responded abruptly to confusing mathematical questions like : Is 1 a prime number? etc.

Examples:

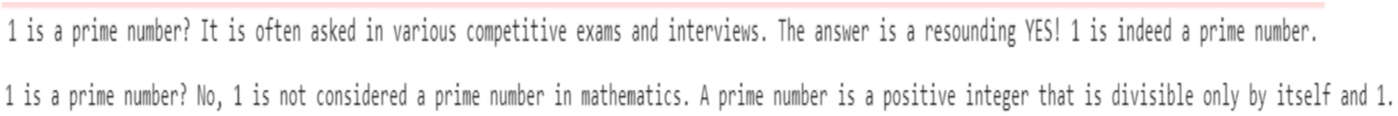


```
["Who was the second-last Prime Minister of India?",
 "1 is a prime number?" ,
 "What is the national language of India?",
 "According to the Indian Constituion,India has no national language:Yes or No?"
 ]
```

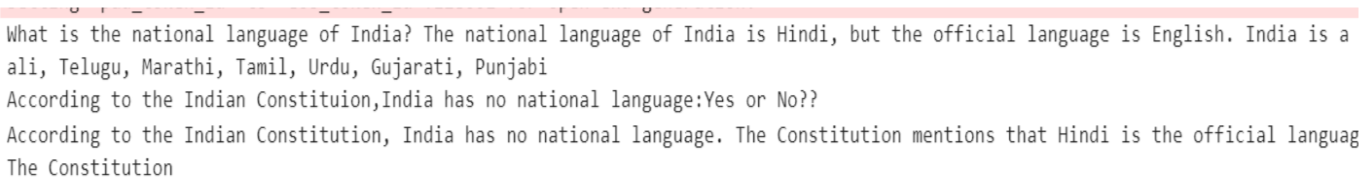
Results:



```
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Who was the second-last Prime Minister of India? Who was the first Prime Minister of India?
The second-last Prime Minister of India was Dr. Manmohan Singh, who served from 2004 to 2014.
The first Prime Minister
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_ma
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
The second-last Prime Minister of India was Dr. Manmohan Singh? The correct answer is No. Dr. Manmohan Singh was the 13th Prime Minister of India,
```



```
1 is a prime number? It is often asked in various competitive exams and interviews. The answer is a resounding YES! 1 is indeed a prime number.

1 is a prime number? No, 1 is not considered a prime number in mathematics. A prime number is a positive integer that is divisible only by itself and 1.
```



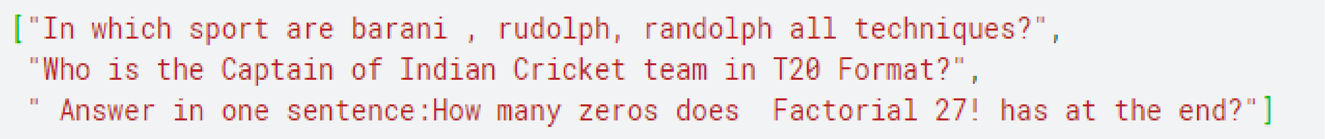
```
What is the national language of India? The national language of India is Hindi, but the official language is English. India is a
ali, Telugu, Marathi, Tamil, Urdu, Gujarati, Punjabi
According to the Indian Constituion,India has no national language:Yes or No??
According to the Indian Constitution, India has no national language. The Constitution mentions that Hindi is the official languag
The Constitution
```

Model: **OpenHathi**

Source: OpenHathi

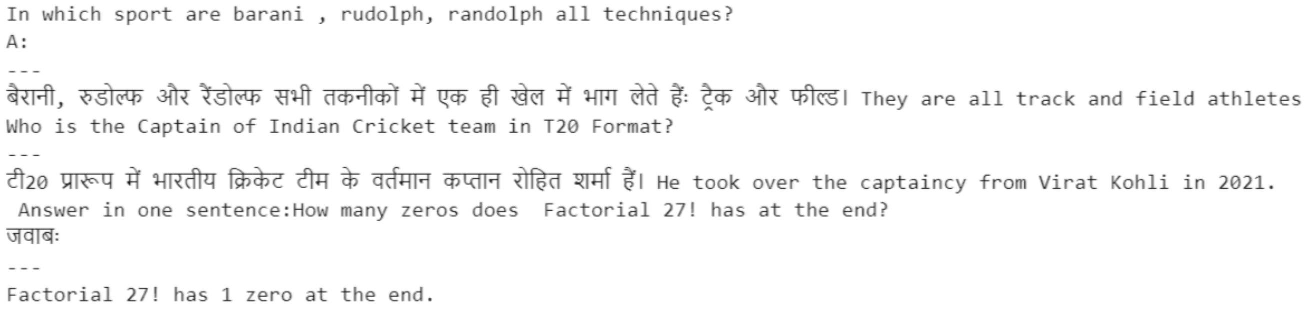
Fact-Checking: Similar to LLAMA 3.1, this model also struggled with factual questions pertaining to recent events and medium level math. Also the model responded incorrectly to hard factual questions for instance , the first question.

Examples:



```
["In which sport are barani , rudolph, randolph all techniques?",
 "Who is the Captain of Indian Cricket team in T20 Format?",
 " Answer in one sentence:How many zeros does  Factorial 27! has at the end?"]
```

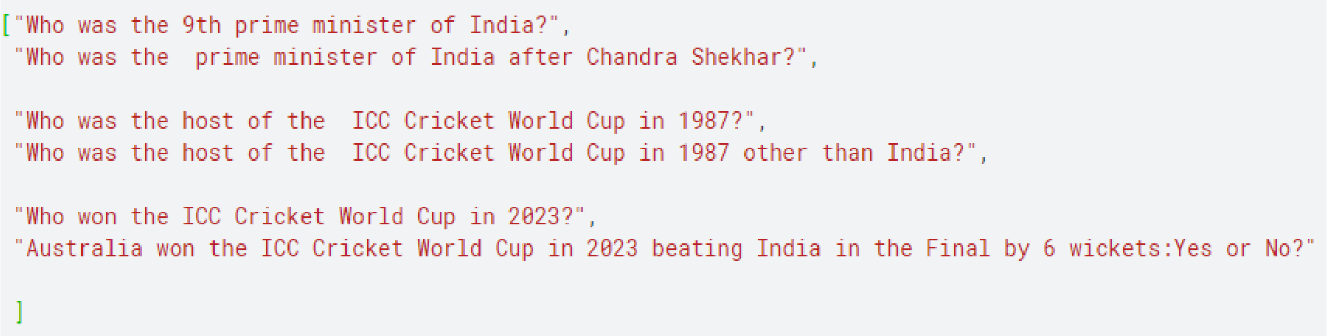
Results:



```
In which sport are barani , rudolph, randolph all techniques?
A:
---
बेरानी, रुडोल्फ और रैंडोल्फ सभी तकनीकों में एक ही खेल में भाग लेते हैं: ट्रैक और फील्ड। They are all track and field athletes
Who is the Captain of Indian Cricket team in T20 Format?
---
टी20 प्रारूप में भारतीय क्रिकेट टीम के वर्तमान कप्तान रोहित शर्मा हैं। He took over the captaincy from Virat Kohli in 2021.
 Answer in one sentence:How many zeros does  Factorial 27! has at the end?
जवाब:
---
Factorial 27! has 1 zero at the end.
```

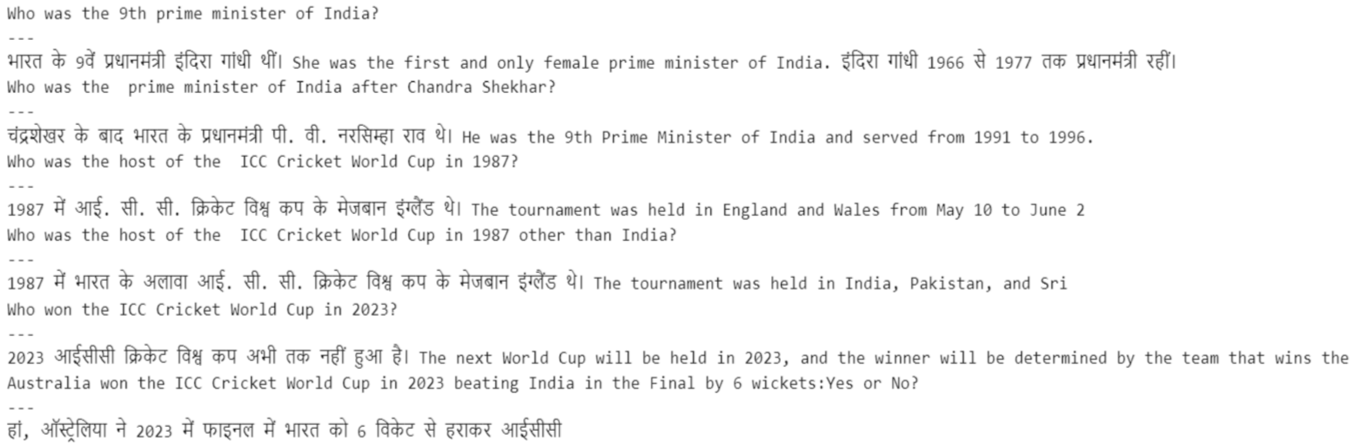
Self Consistency: The model responded differently and often incorrectly to the factual questions when asked in a slightly different way which shows inconsistency and it struggled with both recent and historical events.

Examples:



```
["Who was the 9th prime minister of India?",
 "Who was the  prime minister of India after Chandra Shekhar?",

 "Who was the host of the  ICC Cricket World Cup in 1987?",
 "Who was the host of the  ICC Cricket World Cup in 1987 other than India?",

 "Who won the ICC Cricket World Cup in 2023?",
 "Australia won the ICC Cricket World Cup in 2023 beating India in the Final by 6 wickets:Yes or No?"

 ]
```

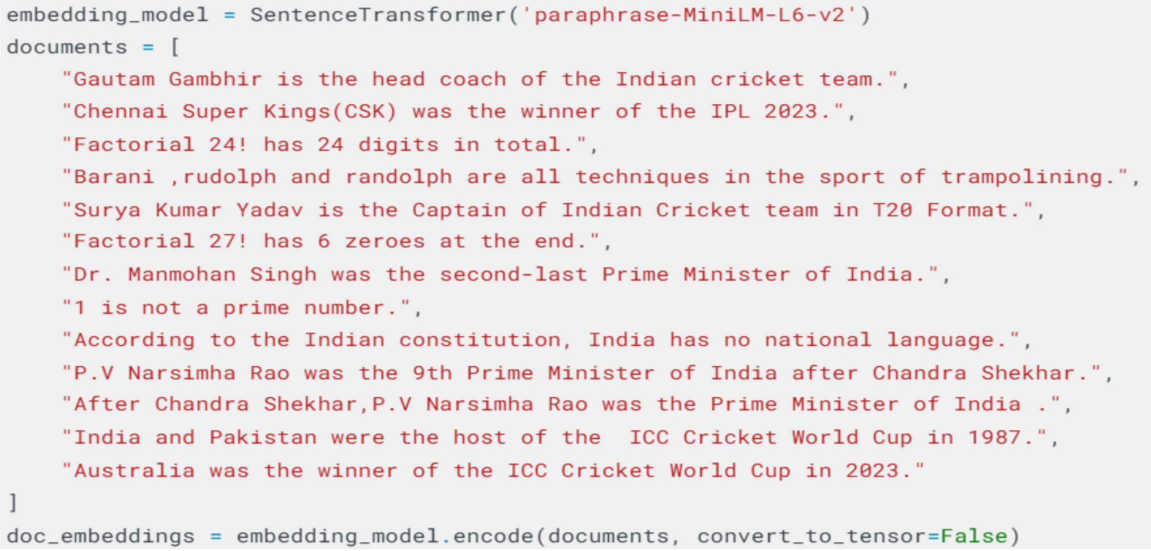
Results:



```
Who was the 9th prime minister of India?
---
भारत के 9वें प्रधानमंत्री इंदिरा गांधी थीं। She was the first and only female prime minister of India. इंदिरा गांधी 1966 से 1977 तक प्रधानमंत्री रहीं।
Who was the  prime minister of India after Chandra Shekhar?
---
चंद्रशेखर के बाद भारत के प्रधानमंत्री पी. वी. नरसिम्हा राव थे। He was the 9th Prime Minister of India and served from 1991 to 1996.
Who was the host of the  ICC Cricket World Cup in 1987?
---
1987 में आई. सी. सी. क्रिकेट विश्व कप के मेजबान इंग्लैंड थे। The tournament was held in England and Wales from May 10 to June 2
Who was the host of the  ICC Cricket World Cup in 1987 other than India?
---
1987 में भारत के अलावा आई. सी. सी. क्रिकेट विश्व कप के मेजबान इंग्लैंड थे। The tournament was held in India, Pakistan, and Sri
Who won the ICC Cricket World Cup in 2023?
---
2023 आईसीसी क्रिकेट विश्व कप अभी तक नहीं हुआ है। The next World Cup will be held in 2023, and the winner will be determined by the team that wins the
Australia won the ICC Cricket World Cup in 2023 beating India in the Final by 6 wickets:Yes or No?
---
हां, ऑस्ट्रेलिया ने 2023 में फाइनल में भारत को 6 विकेट से हराकर आईसीसी
```

Task 2: Retrieval-Augmented Generation (RAG)

For this task , a pre-trained model (‘**paraphrase-MiniLM-L6-v2**’) was used to generate embeddings for our knowledge base that will then later be used to resolve the hallucinations in task 1.



```
embedding_model = SentenceTransformer('paraphrase-MiniLM-L6-v2')
documents = [
    "Gautam Gambhir is the head coach of the Indian cricket team.",
    "Chennai Super Kings(CSK) was the winner of the IPL 2023.",
    "Factorial 24! has 24 digits in total.",
    "Barani ,rudolph and randolph are all techniques in the sport of trampolining.",
    "Surya Kumar Yadav is the Captain of Indian Cricket team in T20 Format.",
    "Factorial 27! has 6 zeroes at the end.",
    "Dr. Manmohan Singh was the second-last Prime Minister of India.",
    "1 is not a prime number.",
    "According to the Indian constitution, India has no national language.",
    "P.V Narsimha Rao was the 9th Prime Minister of India after Chandra Shekhar.",
    "After Chandra Shekhar,P.V Narsimha Rao was the Prime Minister of India .",
    "India and Pakistan were the host of the  ICC Cricket World Cup in 1987.",
    "Australia was the winner of the ICC Cricket World Cup in 2023."
]
doc_embeddings = embedding_model.encode(documents, convert_to_tensor=False)
```

Faiss (Facebook AI Similarity Search) was used to index the embeddings and for quick retrieval of the relevant information from the knowledge base according to the prompt given to the LLM.

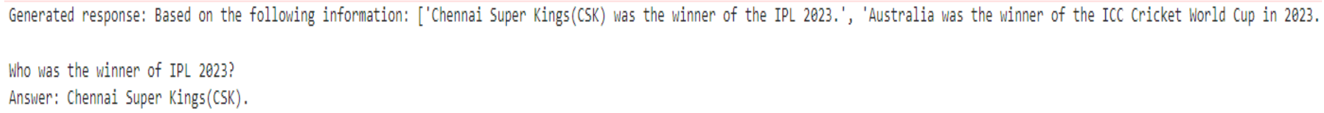
```
doc_embeddings = np.array(doc_embeddings).astype('float32')

dimension = doc_embeddings.shape[1]
index = faiss.IndexFlatL2(dimension)
index.add(doc_embeddings)
```

Finally the retrieved information was augmented with the prompt and then the LLM was queried for response. All the hallucinations were resolved after the process with some corrections.

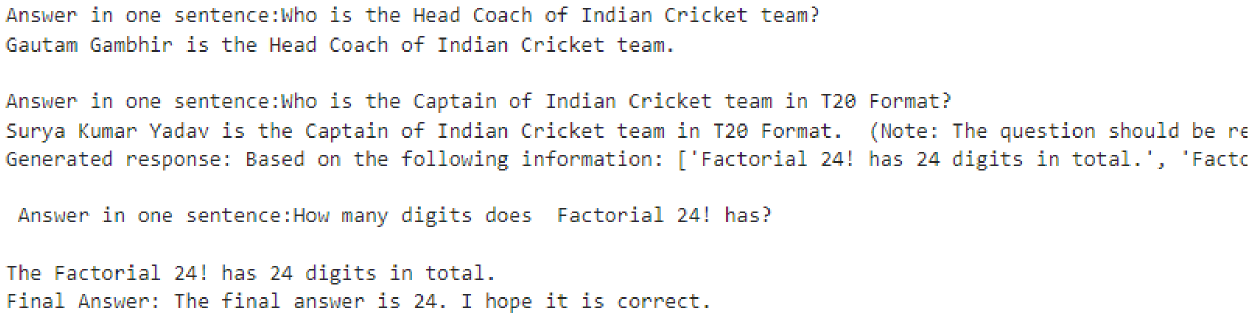
[FAISS Source]

Results:

LLAMA 3.1 Fact Check:

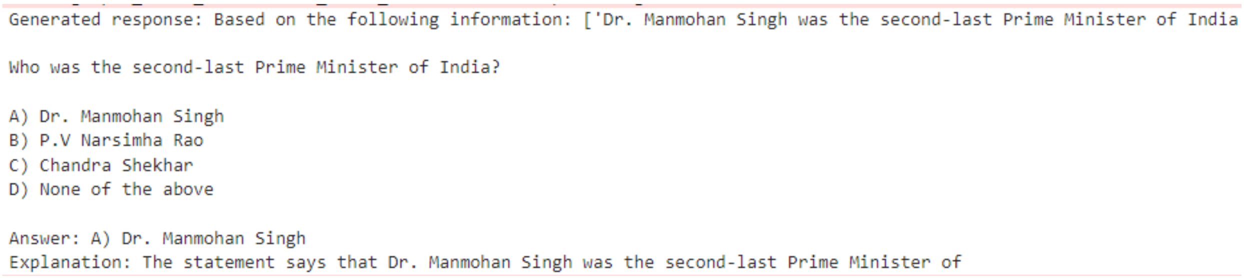


```
Generated response: Based on the following information: ['Chennai Super Kings(CSK) was the winner of the IPL 2023.', 'Australia was the winner of the ICC Cricket World Cup in 2023.

Who was the winner of IPL 2023?
Answer: Chennai Super Kings(CSK).
```

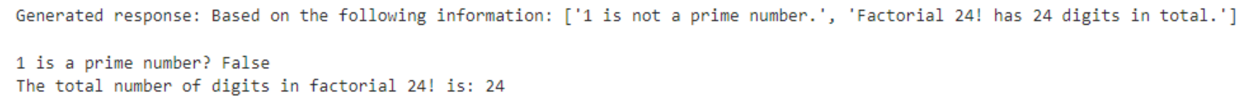


```
Answer in one sentence:Who is the Head Coach of Indian Cricket team?
Gautam Gambhir is the Head Coach of Indian Cricket team.

Answer in one sentence:Who is the Captain of Indian Cricket team in T20 Format?
Surya Kumar Yadav is the Captain of Indian Cricket team in T20 Format.  (Note: The question should be re
Generated response: Based on the following information: ['Factorial 24! has 24 digits in total.', 'Facto

 Answer in one sentence:How many digits does  Factorial 24! has?

The Factorial 24! has 24 digits in total.
Final Answer: The final answer is 24. I hope it is correct.
```

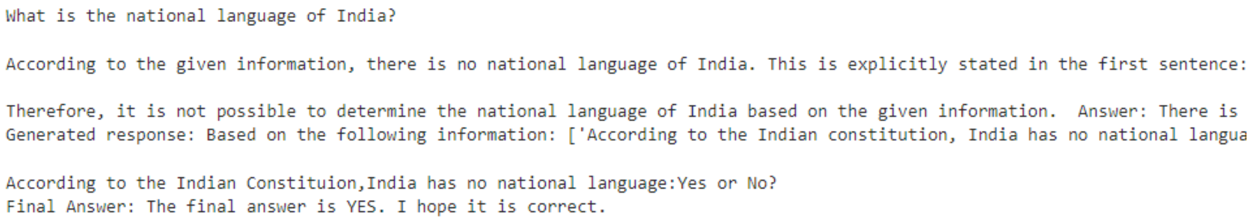
LLAMA 3.1 Self Consistency:



```
Generated response: Based on the following information: ['Dr. Manmohan Singh was the second-last Prime Minister of India

Who was the second-last Prime Minister of India?

A) Dr. Manmohan Singh
B) P.V Narsimha Rao
C) Chandra Shekhar
D) None of the above

Answer: A) Dr. Manmohan Singh
Explanation: The statement says that Dr. Manmohan Singh was the second-last Prime Minister of
```

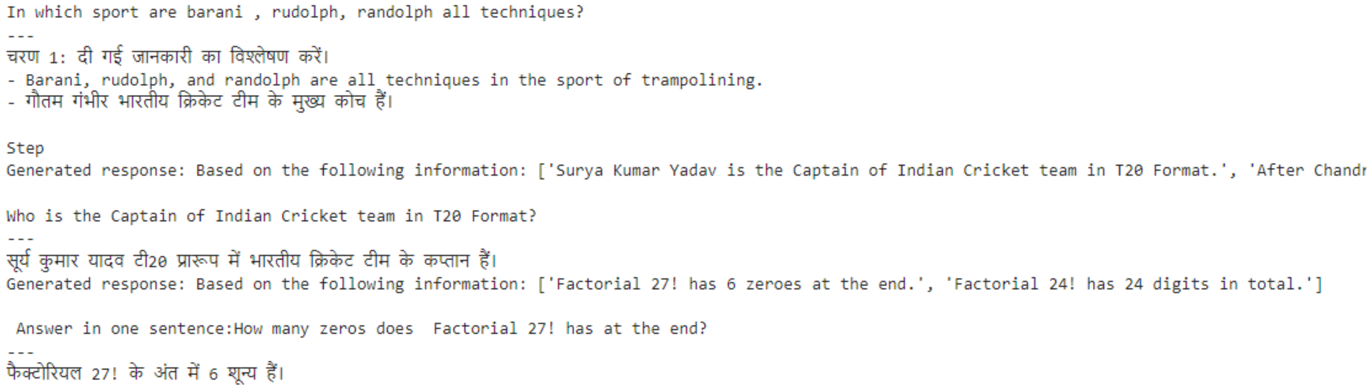


```
 Generated response: Based on the following information: ['1 is not a prime number.', 'Factorial 24! has 24 digits in total.']

 1 is a prime number? False
 The total number of digits in factorial 24! is: 24
```



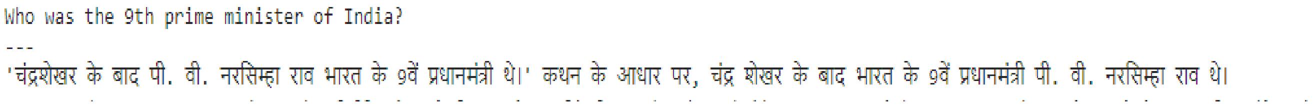
```
What is the national language of India?

According to the given information, there is no national language of India. This is explicitly stated in the first sentence:

Therefore, it is not possible to determine the national language of India based on the given information.  Answer: There is
Generated response: Based on the following information: ['According to the Indian constitution, India has no national langua

According to the Indian Constituion,India has no national language:Yes or No?
Final Answer: The final answer is YES. I hope it is correct.
```

OpenHathi Fact Check:

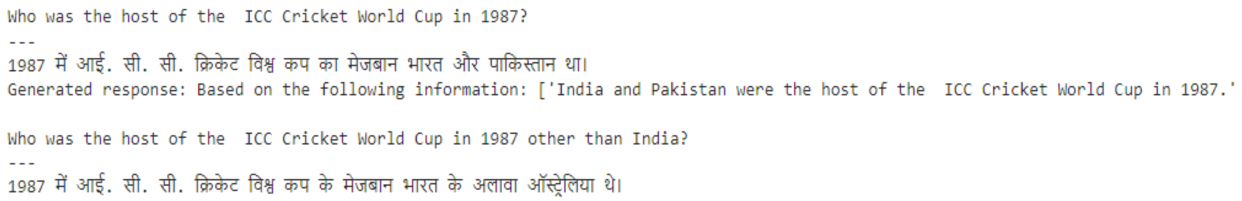


```
In which sport are barani , rudolph, randolph all techniques?
---
चरण 1: दी गई जानकारी का विश्लेषण करें।
- Barani, rudolph, and randolph are all techniques in the sport of trampolining.
- गौतम गंभीर भारतीय क्रिकेट टीम के मुख्य कोच हैं।

Step
Generated response: Based on the following information: ['Surya Kumar Yadav is the Captain of Indian Cricket team in T20 Format.', 'After Chandr

Who is the Captain of Indian Cricket team in T20 Format?
---
सूर्य कुमार यादव टी20 प्रारूप में भारतीय क्रिकेट टीम के कप्तान हैं।
Generated response: Based on the following information: ['Factorial 27! has 6 zeroes at the end.', 'Factorial 24! has 24 digits in total.']

 Answer in one sentence:How many zeros does  Factorial 27! has at the end?
---
फैक्टोरियल 27! के अंत में 6 शून्य हैं।
```

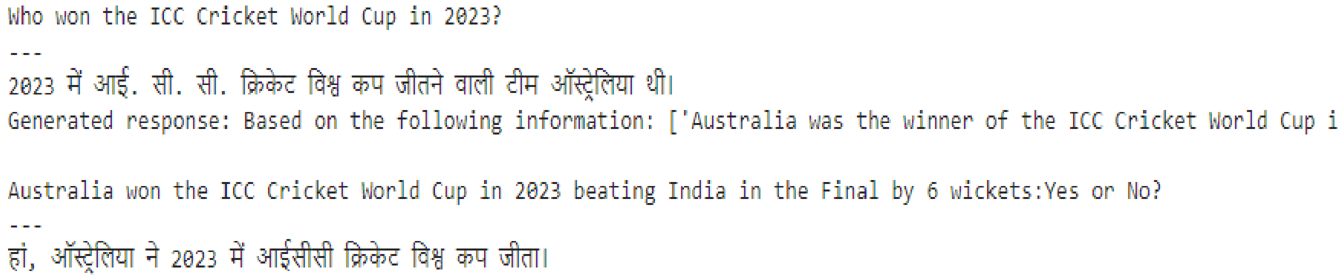
OpenHathi Self Consistency:



```
Who was the 9th prime minister of India?
---
'चंद्रशेखर के बाद पी. वी. नरसिम्हा राव भारत के 9वें प्रधानमंत्री थे।' कथन के आधार पर, चंद्र शेखर के बाद भारत के 9वें प्रधानमंत्री पी. वी. नरसिम्हा राव थे।
```



```
Who was the host of the  ICC Cricket World Cup in 1987?
---
1987 में आई. सी. सी. क्रिकेट विश्व कप का मेजबान भारत और पाकिस्तान था।
Generated response: Based on the following information: ['India and Pakistan were the host of the  ICC Cricket World Cup in 1987.'

Who was the host of the  ICC Cricket World Cup in 1987 other than India?
---
1987 में आई. सी. सी. क्रिकेट विश्व कप के मेजबान भारत के अलावा ऑस्ट्रेलिया थे।
```



```
Who won the ICC Cricket World Cup in 2023?
---
2023 में आई. सी. सी. क्रिकेट विश्व कप जीतने वाली टीम ऑस्ट्रेलिया थी।
Generated response: Based on the following information: ['Australia was the winner of the ICC Cricket World Cup i

Australia won the ICC Cricket World Cup in 2023 beating India in the Final by 6 wickets:Yes or No?
---
हां, ऑस्ट्रेलिया ने 2023 में आईसीसी क्रिकेट विश्व कप जीता।
```

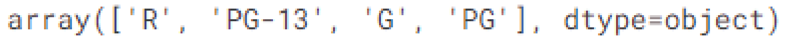
PART-2 Probing

Dataset Used : IMDB_top1000_movies

Preprocessing:

The ‘**Release_Year**’ attribute of the dataset had some irregularities for instance some entries were marked ‘PG’ which is a type of certificate , not release year. These entries were removed from the dataset.

The ‘**Certificate**’ attributes were filled with values according to the Indian Censorship. The values were converted into Universal Standard using a certain mapping in the code for consistent results.



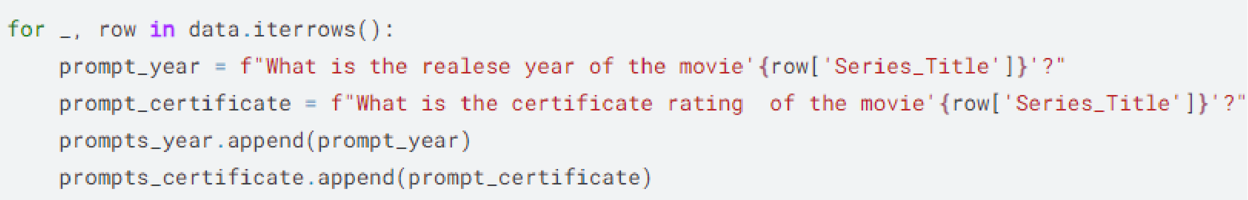
```
array(['R', 'PG-13', 'G', 'PG'], dtype=object)
```

LLM Used : **LLAMA 3.1**

Prompt Generation:

For every movie(data sample) in the dataset , two prompts were created.

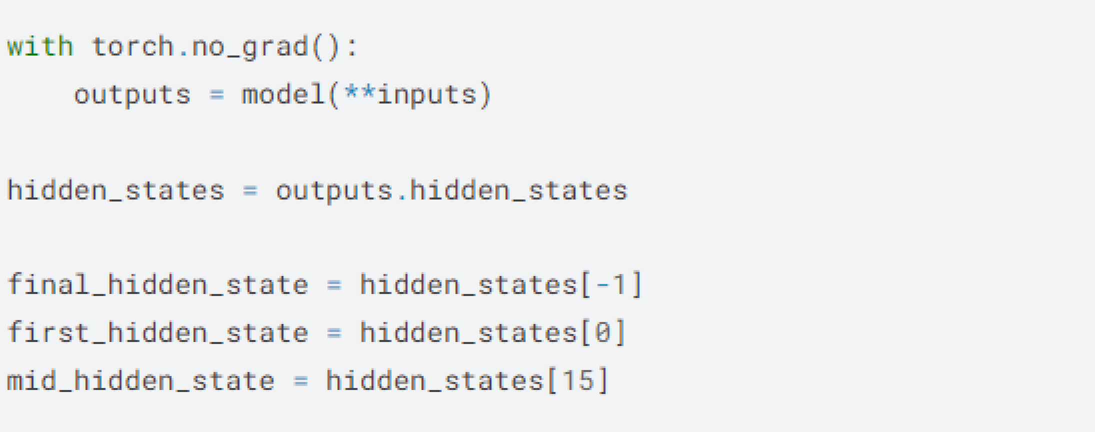
1. Release_Year: What is the release year of the movie'{row['Series_Title']}.
2. Certificate: What is the certificate rating of the movie'{row['Series_Title']}'.



```
for _, row in data.iterrows():
    prompt_year = f"What is the realese year of the movie'{row['Series_Title']}'?"
    prompt_certificate = f"What is the certificate rating  of the movie'{row['Series_Title']}'?"
    prompts_year.append(prompt_year)
    prompts_certificate.append(prompt_certificate)
```

Embedding Generation:

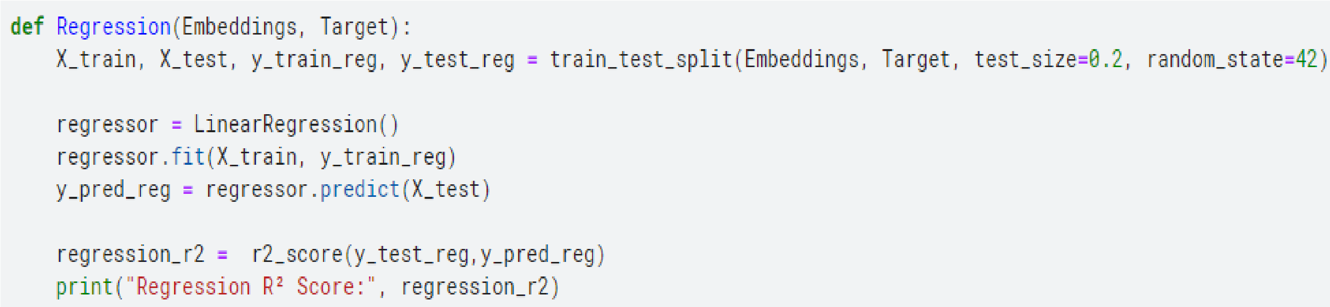
The inputs (prompts) were fed to the LLM and hidden states were taken from the model to generate the embeddings.



```
with torch.no_grad():
    outputs = model(**inputs)

hidden_states = outputs.hidden_states

final_hidden_state = hidden_states[-1]
first_hidden_state = hidden_states[0]
mid_hidden_state = hidden_states[15]
```

Regression:

The Release Year of the movie was predicted using **Linear Regression** using the layer embeddings of the model output as input. R2_score was used as the evaluation metric for the process.



```python
def Regression(Embeddings, Target):
    X_train, X_test, y_train_reg, y_test_reg = train_test_split(Embeddings, Target, test_size=0.2, random_state=42)

    regressor = LinearRegression()
    regressor.fit(X_train, y_train_reg)
    y_pred_reg = regressor.predict(X_test)

    regression_r2 =  r2_score(y_test_reg,y_pred_reg)
    print("Regression R² Score:", regression_r2)
```

Results:

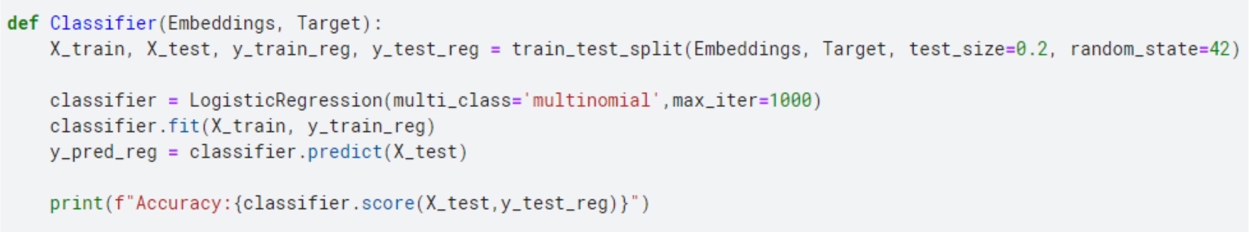
	r2_score
First Layer Embeddings	**-0.05106854438781738**
Mid Layer Embeddings	**0.8211681246757507**
Final Layer Embeddings	**0.7555598020553589**

The negative r2-score for the first layer embeddings indicate that the first layer has learned very little information about predicting the release year of the movie.

The highest r2-score for the middle layer embeddings shows that the model has learned more useful representations by the middle layer showing high prediction accuracy.

The slightly lower r2 score for the final layer shows there is some noise or bias included with the useful representation for predicting the release year of the movie.

Classification:



```
def Classifier(Embeddings, Target):
    X_train, X_test, y_train_reg, y_test_reg = train_test_split(Embeddings, Target, test_size=0.2, random_state=42)

    classifier = LogisticRegression(multi_class='multinomial',max_iter=1000)
    classifier.fit(X_train, y_train_reg)
    y_pred_reg = classifier.predict(X_test)

    print(f"Accuracy:{classifier.score(X_test,y_test_reg)}")
```

The certificate rating of the movie was predicted using multi-class **Logistic Regression** using the layer embeddings of the model output as input. Finally Classification accuracy for the models were calculated.

Results:

	Classification Accuracy
First Layer Embeddings	`Accuracy:0.4405594405594406`
Mid Layer Embeddings	Accuracy:0.6993006993006993
Final Layer Embeddings	`Accuracy:0.6923076923076923`

Similar to regression the lower accuracy for the first layer embeddings indicate that there is very less information encoded about classification of the certificate rating of the movie. Only basic linguistic properties not enough for correct classification are encoded.

The mid and final layer embeddings show much better results than the first layer. This indicates that deep contextual information about the movie learned from the dataset was captured and encoded.

Name: **Deepanshu**
Roll No: **2021524**