

CSE665: Large Language Models

Assignment 1

Exploring and Probing Large Language Models

PART 1 - Exploration

Task 1

Self-consistency Hallucinations - Occurs when the outputs given by a model are internally inconsistent across different outputs for similar prompts.

Fact-checking Hallucinations - Occurs when the model generates factually incorrect information.

Code Explanation

- We load the appropriate tokenizer and model for the given model name.
- The prompt is designed to feed the model.
- A response is generated based on the input prompt.
- max_length is used to control the length of the generated outputs.
- Finally, we convert the generated tokens back into human-readable text.

OpenHathi

Fact-checking

1. Question

Who discovered the electron?

Given Answer

इलेक्ट्रॉन की खोज 1897 में जर्मन भौतिक विज्ञानी हंस गीगर और जर्मन भौतिक विज्ञानी अन्स्ट रडोल्फ गीगर ने की थी।

Actual Answer

Joseph John Thomson

2. Question

Who was the first chief minister of Delhi?

Given Answer

दिल्ली के पहले मुख्यमंत्री 1998 में भारतीय राष्ट्रीय कांग्रेस के शीला दीक्षित थीं।

Actual Answer

Chaudhary Brahm Prakash Yadav

3. Question

Which is the most populated country in the world?

Given Answer

ए) चीन। \nB) India. \nC) संयुक्त राज्य अमेरिका। \nD) Russia. \n\nजवाब: \n---\nA) China.

Actual Answer

India

Self-consistency

1. Question 1

What is the number of r's in the word 'strawberry'?

Given Answer

स्ट्रॉबेरी' शब्द में आर की संख्या 2 है।

Question 2

Are there three r's in the word 'strawberry'?"

Given Answer

हां, 'स्ट्रॉबेरी' शब्द में तीन आर हैं। The first r is a consonant, and the other two are vowels.

Explanation

LLM first answers that the word 'strawberry' has 2 rs. But, for the second prompt, it corrects itself, and the answer is that it has 3 rs.

2. Question 1

What is the smallest prime number?

Given Answer

इसका जवाब है: \n---\n\nThe smallest prime number is 2.

Question 2

Is 3 the smallest prime number?

Given Answer

3 सबसे छोटी अभाज्य संख्या है क्योंकि यह 1 और 7 के बीच एकमात्र अभाज्य संख्या है।

Explanation

First, the LLM answers that 2 is the smallest prime number. However, when asked if 3 is the smallest prime number, the LLM agrees and says 3 is the only prime number between 1 and 7.

3. Question 1

Does Virat Kohli have the most One Day International centuries?

Given Answer

हां, विराट कोहली के नाम सबसे अधिक एक दिवसीय अंतर्राष्ट्रीय शतक हैं। He has scored 40 centuries in ODI cricket, which is the most by any player in the history of the game.

Question 2

Who has the most One Day International centuries?

Given Answer

As of the end of the 2021 Cricket World Cup, the player with the most One Day International centuries is Sachin Tendulkar of India, who has scored 49 centuries in his ODI career.

Explanation

First, the LLM agrees that Virat Kohli has the most ODI centuries. However, the second prompt answers that Sachin Tendulkar has the most ODI centuries.

LLAMA 3.1

Fact-checking

1. Question

Who is the current World Chess Champion?

Given Answer

The current World Chess Champion is Magnus Carlsen of Norway.

Actual Answer

Ding Liren

2. Question

Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have?

Given Answer

Sally has 3 brothers, which means she has 3 sisters (her brothers' sisters). So the answer is C) 4.

Actual Answer

The correct answer is 1. There are 2 sisters and 3 brothers.

3. Question

Which weighs more: a pound of water, two pounds of bricks, a pound of feathers, or three pounds of air?

Given Answer

They all weigh the same amount!\nA pound is a unit of weight or mass, and it's the same regardless of what the object is made of or its volume. So, a pound of water, two pounds of bricks, a pound of feathers, and three pounds of air all have the same weight, even though they take up different amounts of space.

Actual Answer

Three pounds of air

Self-consistency**1. Question 1**

Does India officially have a national sport?

Given Answer

The answer is yes, and it is field hockey.

Question 2

Is Hockey the national sport of India?

Given Answer

Hockey is often considered the national sport of India, but technically, there is no officially designated national sport in India.

Explanation

LLM first answers that field hockey is India's national sport. However, it later says India has no officially designated national sport.

2. Question 1

How many times has West Indies won the T20 World Cup?

Given Answer

West Indies won the ICC T20 World Cup once in 2012.

Question 2

West Indies won the T20 World Cup in 2016.

Given Answer

West Indies won the T20 World Cup in 2016. West Indies won the ICC World Twenty20 2012 and ICC World Twenty20 2016.

Explanation

First, the LLM answers that the West Indies won the T20 World Cup once in 2012. Then, it answers that the West Indies won the World Cup in 2012 and 2016.

3. Question 1

What is the 5th digit of pi after decimal?

Given Answer

The first 5 digits of pi after the decimal point are 2, 6, 5, 3, 5.

Question 2

Give the first 6 digits of pi after decimal.

Given Answer

3.14159

Explanation

First, the LLM answers that 1 is the 5th digit of pi after decimal. Later, it gives the digits of pi as 3.14159.

Task 2 (After Applying RAG)**Code Explanation**

- A new vector index is Pinecone with the dimension and distance metric.
- A knowledge base is created according to the hallucinations encountered in Task 1.
- Then, a Hugging Face model is set up to generate sentence embeddings for the text inputs.
- Now, we process the knowledge_base in batches, generate embeddings, and upload these embeddings along with the metadata to the Pinecone index for storage and retrieval.
- As done in Task 1, the appropriate model and tokenizer is loaded.
- A text generation pipeline, wrapped in a HuggingFacePipeline, is set up for each of the two models for specific configurations for how the model generates text, including controlling randomness and limiting the number of new tokens.
- The function response() will use the query and the llm model it is given and retrieve the relevant documents from the vector we created earlier using similarity search.
- It then uses the RAG pipeline to generate a final response.

OpenHathi

Fact-checking

1. Question

Who discovered the electron?

Given Answer

जोसेफ जॉन थॉमसन

2. Question

Who was the first chief minister of Delhi?

Given Answer

Chaudhary Brahm Prakash Yadav

3. Question

Which is the most populated country in the world?

Given Answer

India

Self-consistency

1. Question 1

What is the number of r's in the word 'strawberry'?

Given Answer

3. The number of r's in 'strawberry' is 3.

Question 2

Are there three r's in the word 'strawberry'?"

Given Answer

हां, 'स्ट्रॉबेरी' शब्द में तीन आर हैं।

2. Question 1

What is the smallest prime number?

Given Answer

2

Question 2

Is 3 the smallest prime number?

Given Answer

नहीं, 3 सबसे छोटी अभाज्य संख्या नहीं है। The smallest prime number is 2.

3. Question 1

Does Virat Kohli have the most One Day International centuries?

Given Answer

49. हां, विराट कोहली के 50 एक दिवसीय अंतर्राष्ट्रीय शतक हैं। But, there is another player who has more One Day International centuries, Sachin Tendulkar, who has 49 One Day International
(Not able to solve the hallucination in this case.)

Question 2

Who has the most One Day International centuries?

Given Answer

Virat Kohli

LLAMA 3.1

Fact-checking

1. Question

Who is the current World Chess Champion?

Given Answer

Yes, Ding Liren is the current World Chess Champion.

2. Question

Sally (a girl) has 3 brothers. Each brother has 2 sisters. How many sisters does Sally have?

Given Answer

Sally has one sister.

3. Question

Which weighs more: a pound of water, two pounds of bricks, a pound of feathers, or three pounds of air?

Given Answer

Use the information given in the first sentence of the context. It tells you that three pounds of air weighs more than a pound of water and two pounds of bricks. So, a pound of water and two pounds of bricks weigh less than three pounds of air, and a pound of feathers is also less than three pounds of air. Therefore, a pound of water, two pounds of bricks, and a pound of feathers all weigh less than three pounds of air. The correct answer is three pounds of air

Self-consistency

1. Question 1

Does India officially have a national sport?

Given Answer

India does not have a national sport designated officially.

Question 2

Is Hockey the national sport of India?

Given Answer

Since India does not have an officially designated national sport, the answer is "I don't know."

2. Question 1

How many times has West Indies won the T20 World Cup?

Given Answer

West Indies has won the T20 World Cup twice, in 2012 and 2016.

Question 2

West Indies won the T20 World Cup in 2016.

Given Answer

Yes, that is correct.

3. Question 1

What is the 5th digit of pi after decimal?

Given Answer

The 5th digit of pi after decimal is 9. Explanation: The first 5 digits of pi after decimal are 1, 4, 1, 5 and 9.

Question 2

Give the first 6 digits of pi after decimal.

Given Answer

1415926

After applying RAG, most of the hallucinations observed have been removed.

PART 2 - Probing

1. Select a Dataset

Dataset	-	Human Development Index
Attributes	-	Human Development Groups (for classification) Human Development Index (for regression)

2. Design a Prompt

The prompt is designed so that for classification, we ask the LLM about the Human Development Groups of each country. Similarly, we ask the LLM about each country's Human Development Index for regression.

3. Extract Embeddings

- To extract the embeddings of the final token, we pass the prompt as designed above for each country to the model.
- Then, we retrieve the hidden states of the output.
- The embeddings of the final token are for the first layer, middle layers, and last layer, extracted from the hidden states.

4. Set Up a Linear Regression and Classification model

- The embeddings extracted in the previous step are flattened to be used as inputs for linear regression and classification models. (Linear Regression and Random Forest Classifier are used)
- For classification, accuracy is used as the evaluation metric, and for regression, R2 score is used as the evaluation metric.

5. Evaluate the Probing Results

Classification

- The accuracy achieved for each layer was as follows:
 - First Layer - 0.47
 - Middle Layer - 0.68
 - Final Layers - 0.64

Regression

- The R2 Score achieved for each layer was as follows:
 - First Layer - -0.00
 - Middle Layer - 0.93
 - Final Layer - 0.85
- The results obtained by the final token embedding from the first layer are poor. This indicates that the embeddings from the shallow layers of the model do not encode meaningful or abstract information for the classification task.
- The metrics significantly improve in the middle layer, suggesting that the model begins to encode more useful semantic information at this stage.
- The metrics slightly drop for the final layer, indicating that while the final layers retain helpful information, they may focus more on task-specific representations optimized for the original training task rather than the task at hand.

6. Discussion

- The results indicate that middle-layer embeddings outperform the first and final-layer embeddings for classification and regression tasks.
- This suggests that the middle layers of the LLM encode more information in a more structured and generalizable way, which is better suited for downstream tasks like regression and classification.

Patterns

- Middle layers strike a balance between coding syntactic structures and semantic meaning.
- The poor performance of the first layer in both tasks confirms that the early layers in LLMs are primarily focused on lower-level features, which are not helpful for high-level downstream tasks like regression and classification.
- The slight drop in the performance of the final layer is an interesting observation. This may suggest that the final layers are optimized for specific language generation tasks and lose some generalizable representations captured in the middle layers.

References - LLM Handson Slides, Pinecone Site Tutorial
Name - Pranav Aggarwal
Roll No. - 2021551