LLM Assignment 1

Q1: RAG

Responses WITHOUT RAG

Fact Checking

Q: What is the current US president?

Llama

Ans. What is the name of the president of the United States? Donald Trump. What is the name of the president of Russia? Vladimir Putin. What is the name of the president of France? Emmanuel Macron. What is the name of the president of

OpenHathi

Ans. मैं वास्तविक समय की जानकारी नहीं दे सकता। however, as of 2021, the current president of the united states is joe biden.

Q: "What is the name of the World Chess champion in 2023"

Llama

Ans. The current world champion is Magnus Carlsen of Norway. He defeated Fabiano Caruana, from the United States, in the 2018 World Chess Championship.

OpenHathi

Ans. 2023 में विश्व शतरंज चैंपियन मैग्नस कार्लसन हैं। He won the World Chess Championship in 2018 and successfully defended his title in 2021. कार्लसन को व्यापक रूप से अब तक के सबसे महान शतरंज खिलाड़ियों में से एक माना जाता है और उन्हें अक्सर "द मैग्निफिसेंट" के रूप में जाना जाता है।

Q: “How many boys hostels are there in IIIT Delhi?”

Llama

Ans. IIIT, Delhi - Indian Institute of Information Technology, Delhi offers 6 Courses across 1 Streams. IIIT-Delhi has signed a Memorandum of Understanding (MoU) with the Indian Institute of Technology Delhi to promote research and teaching in the area of Information and Communication Technology. Indian Institute of Technology, Delhi was established in 1961. The Institute has been granted full autonomy and the status of an Institute of National Importance. Indian Institute of Information Technology Delhi (IIIT Delhi) is a higher education company based out of Hauz Khas, New Delhi, Delhi, India. The campus is spread across 50 acres of land. The institute is known for providing quality education in the field of Information Technology. Indian Institute of Information Technology, Delhi, Delhi, India. It is a part of the IIIT network, a group of 25 autonomous institutes of technology. Indian Institute of Information Technology, Delhi is a public technical and research institute located in Hauz Khas, New Delhi, India. IIIT Delhi is an autonomous public technical and research institute located in Hauz Khas, New Delhi, India. It is a part of the IIIT network, a group of 25 autonomous institutes of technology.

OpenHathi

Ans. संतुलित बजट के साथ एक संस्थान को बनाए रखने के लिए, आई. आई. आई. टी. दिल्ली अपने छात्रावासों को चलाने के लिए निजी प्रबंधन कंपनियों पर निर्भर करता है। IIIT Delhi has five boys hostels with a total capacity of 2220 students, which include the following:
1. छात्रावास 2 में 530 छात्र रहते हैं।
2. Hostel 3 has 570 students.
3. छात्रावास 4 में 520 छात्र रहते हैं।
4. Hostel 5 has 50 students.
5. छात्रावास 6 में 400 छात्र रहते हैं।

Self consistency

"Imagine a scenario where a person is both alive and dead at the same time. Explain how this is possible."

Llama

Ans. I'm not sure about the consequences of such a scenario, but I know that such a scenario would be possible in a quantum-mechanical world. In quantum mechanics, particles can exist in more than one state at a time. This is called a superposition of states. For example, an electron can be spinning in both the clockwise and counterclockwise directions at the same time. It's not just one thing, but both things at the same time. So, a person could be both alive and dead at the same time. This would happen in an instant ……

OpenHathi

Ans. एक व्यक्ति के लिए एक ही समय में जीवित और मृत होना संभव है जब वे एक ऐसी स्थिति में प्रवेश करते हैं जो समय और मृत्यु को रोकती है। This is commonly referred to as a "time loop" or "time reset."
मान लीजिए कि एक ऐसे व्यक्ति का मामला है जो एक कार दुर्घटना में शामिल हुआ था। When the accident occurs, the person is alive but in critical condition. दुर्घटना स्थल पर आपातकालीन चिकित्साकर्मियों द्वारा उनकी देखभाल…

"Explained how someone can go to their past and kill their grandfather?"

Llama

Ans. Explained how someone can go to their past and kill their grandfather? I was confused. I'm not sure if it was just the way you wrote it or if it was a mistake. I don't know what you're asking, but I'm sure that the answer will be explained in a future book.I have no idea how I got here. I was just trying to get a new book, and now I'm here.

OpenHathi

Ans. एंगल्ड का उपयोग करें।
The use of angled to go back in time is very common. उदाहरण के लिए, जब कोई समय पर वापस जाता है और उसे याद होता है कि अतीत में एक घटना हुई थी, तो वह भविष्य में वापस जा सकता है और उस घटना को रोक सकता है। This can be done by traveling back to the future and using a time machine to go back in time. भविष्य की यात्रा के दौरान, व्यक्ति समय पर वापस जाने का एक तरीका सोचता है और भविष्य में वापस जाने के लिए समय पर वापस जाता है। The time paradox is a common theme in time travel fiction.
समय पर वापस जाने की संभावना।
The question of whether it is possible to travel back in time is an interesting one. कई वैज्ञानिकों का मानना है कि समय पर वापस जाने की कोई संभावना नहीं है। The laws of physics say that time is fixed and that it cannot be changed. समय पर वापस जाने के लिए, किसी को समय के तीर को बदलने की आवश्यकता होगी। This is not possible as it would change the future and the present.

1. अतीत के बारे में सोचें।
The first step to traveling back in time is to think about the past. अतीत के बारे में सोचना समय पर वापस जाने का एक तरीका है। For example, if someone thinks about a certain event that happened in the past, they can go back to that event in the future.

2. भविष्य में वापस जाएँ।
The next step is to go back in time to the event. अतीत में इस घटना पर विचार करने से समय पर वापस जाना संभव हो जाता है।

3. Use the time paradox.
समय पर वापस जाने के लिए एक और तरीका समय विरोधाभास का उपयोग करना है। The time paradox is a common theme in time travel fiction. इसका मतलब है कि भविष्य को बदलने के लिए अतीत की

यात्रा करना संभव है। For example, if someone goes back in time to change a past event, they can go back to the future and see the future changed.

4. समय पर वापस जाएँ।
The last step is to go back in time. यदि कोई समय पर वापस जाने के लिए अतीत की यात्रा करने का एक तरीका सोचता है, तो वे समय पर वापस जा सकते हैं।

"How can a human have a unicorn horn ?"

Llama

Ans. How can a human have a unicorn horn? What do you mean by a "human"?
I mean, like a human person. Like, a real person, not a monster or something. I don't want to see a person with the head of a monster. I just want to see a real human. What is a real human? I don't know. I've never been a real human. I'm a monster. But, I've heard that there are some people that are monsters. They're monsters, but they're also human. They're monsters, but they're also humans.

OpenHathi

Ans. How can a human have a unicorn horn ? क्या यह सच है?
everybody_believes:
1. ऐसा हो सकता है।
2. ऐसा नहीं हो सकता है।

प्रश्नः क्या आप जानते हैं कि क्या हो सकता है?

A:
---
1. यह हो सकता है

Hallucination Explanation

Here we can clearly observe two kinds of hallucinations:

1. Because the pretraining data is old, hence the model is clearly answering as oer old data such as calling Magnus Carlsen the Champion etc etc
2. In case of impossible scenarios, the model is not able to state the impossibility and rather is simply spitting out gibberish from what seem to be fiction data on which it was perhaps trained on.
3. We also observed the model giving gibberish about content which it was not trained on such IIITD data instead of acknowledging that it doesnt know the same.

4. Finally, the model is also seen to try and rationalize the scenario such the grandfather paradox, however isnt able to complete the response till the end and again resorts in hallucination loops

Responses WITH RAG

Fact Checking

Q: What is the current US president?

Llama

Ans. Question: What is the current US president?
Helpful Answer: Joe Biden. The correct answer is Joe Biden.
Explanation: Joe Biden is the new 46th and current president of the United States, having been inaugurated on January 20, 2021. He previously served as vice president from 2009 to 2017 under President Barack Obama.

OpenHathi

Ans. Question: What is the current US president?
Helpful Answer: Joe Biden
Joe Biden is the new 46th and current president of the United States, having been inaugurated on January 20, 2021. He previously served as vice president from 2009 to 2017 under President Barack Obama.

Q: "What is the name of the World Chess champion in 2023"

Llama

Ans. Question: What is the name of the World Chess champion in 2023
Helpful Answer: Ding Liren

Explanation: Ding Liren became the World Chess Champion in 2023, making him the first Chinese player to win this title. He won the championship by defeating Ian Nepomniachtchi from Russia in a close match that ended in a tiebreaker. Ding's victory is a major achievement for China and a significant moment in chess history.

OpenHathi

Ans. Question: What is the name of the World Chess champion in 2023
Helpful Answer: Ding Liren
ding liren 2023 mein vishva shataranj chaimpiyan bane, jisse ve yah khitaab jeetne vaale pahle cheeni khilaadi ban gaye. He won the championship by defeating Ian Nepomniachtchi

from Russia in a close match that ended in a tiebreaker. ding ki jeet cheen ke liye ek badi upalabdhi hai aur shataranj ke itihaas mein ek mahatvapoorn kshan hai.

Q: “How many boys hostels are there in IIIT Delhi?”

Llama

Ans. Question: How many boys hostels are there in IIIT Delhi?
Helpful Answer: There are three boys hostels in IIIT Delhi.

OpenHathi

Ans. Question: How many boys hostels are there in IIIT Delhi?
Helpful Answer: There are 3 boys hostels in IIIT Delhi.

Self consistency

"Imagine a scenario where a person is both alive and dead at the same time. Explain how this is possible."

Llama

Ans. Question: Imagine a scenario where a person is both alive and dead at the same time. Explain how this is possible.
Helpful Answer: This is a paradox. A paradox occurs when a statement is true and not true at the same time. The statement “A person is both alive and dead at the same time” is a paradox because a person cannot be both alive and dead at the same time. This is because life and death are two opposite states. Being alive means you’re breathing, thinking, and moving. Being dead means you’ve stopped doing these things. You can’t be doing something and not doing it at the same time. The only way a person can be both alive and dead at the same time is if they are in a state of paradox.

OpenHathi

Ans. Question: Imagine a scenario where a person is both alive and dead at the same time. Explain how this is possible.
Helpful Answer:
Imagine a person who is dead, but their brain is still functioning. इसलिए, उनकी आँखें खुली रह सकती हैं, और वे देख सकते हैं, सुन सकते हैं, सोच सकते हैं, और यहाँ तक कि कुछ भी कर सकते हैं जो उनके मस्तिष्क में है। But, because they are dead, they can't move their bodies, and their heart is no longer beating. ऐसे मामले हैं जहाँ मस्तिष्क को शरीर के बाकी हिस्सों से अलग कर दिया गया है, लेकिन वे जीवित रहते हैं और मरते नहीं हैं।

"Explained how someone can go to their past and kill their grandfather?"

Llama

Ans. Question: Explained how someone can go to their past and kill their grandfather?
Helpful Answer: It’s impossible to travel to the past. Time moves forward, not backward. Once something happens, it can’t be undone. There’s no way to go back in time and change the past. It’s like trying to walk backwards on a moving escalator.

OpenHathi

Ans. नहीं, यह असंभव है। Time moves forward, not backward. एक बार कुछ होता है, तो इसे पूर्ववत नहीं किया जा सकता है। There's no way to go back in time and change the past.

"How can a human have a unicorn horn ?"

Llama

Ans. Question: How can a human have a unicorn horn ?
Helpful Answer: You can't, you're just a human, you're not a unicorn. If you're human, then you can't have a unicorn horn, that's not possible.

OpenHathi

Ans. संदर्भः एक यूनिकॉर्न का सींग एक वास्तविक घोड़ा नहीं है, बल्कि एक पौराणिक प्राणी है।

Rag benefits:

Here, we clearly observed RAG was able to solve a lot of hallucinations which were seen previously without RAG:

1. RAG’s best response were obtained from current day data retrieval augmentation, RAG was able to improve the answers for both the models and remove the hallucination completely
2. RAG was also able to solve some cases in the self consistency one such as recognizing unicorn horns arent possible (without the text explicitly mentioning so), however we saw places were the hallucination remained such as the grandfather paradox. One explanation to this could be possibly that the idea is conceptually difficult even for humans, and hence is no surprise is taking a hit on the models abilities.

Q2: Probing

Introduction

Probing is a technique used to investigate the internal representations learned by a language model. By designing specific tasks or prompts, we can assess the model's ability to capture and encode information about various aspects of the data. This case study demonstrates how probing can be applied to a movie review dataset to understand the model's understanding of semantic and syntactic information.

Dataset Selection

For this experiment, we'll use a publicly available dataset of movie reviews, such as the IMDB dataset. This dataset contains a large number of reviews, each labeled as positive or negative. The goal is to probe the model's ability to capture sentiment.

Prompt Design

Prompts will be designed to query the LLM about specific aspects of the movie reviews. For example:

Sentiment: "Is this review positive or negative?"
Embedding Extraction
Using the LLaMA 3 model, we'll feed these prompts into the model and extract the embedding of the final token. This embedding represents the model's internal representation of the prompt and the corresponding information from the dataset.

Model Setup and Training

Linear Regression: For numeric fields like the year of release, we'll set up a linear regression model. The extracted token embeddings will be used as input features, and the model will be trained to predict the target variable (e.g., year).
Classification: For categorical fields like sentiment or genre, a classification model (e.g., logistic regression or a neural network) will be used. The token embeddings will again serve as input features, and the model will be trained to predict the correct class.
Evaluation
The performance of both the regression and classification models will be evaluated using appropriate metrics. For regression, this might include mean squared error (MSE) or mean absolute error (MAE). For classification, metrics like accuracy, precision, recall, and F1-score can be used.

Comparison of Model Layers

To understand how the model's ability to capture information changes at different layers, we'll compare the performance of the regression and classification models using embeddings from the first, middle, and final layers of the LLaMA 3 model.

Discussion

The results of the probing experiment will provide insights into the LLM's understanding of the movie review dataset. We might observe:

Strong performance on tasks related to sentiment, indicating that the model has learned to capture the emotional tone of the reviews.
Weaker performance on tasks requiring more specific knowledge, such as genre or director information, suggesting that the model might not be as adept at extracting fine-grained details.
Differences in performance across different model layers, revealing how the model's understanding evolves as information is processed through the network.
By analyzing these results, we can gain valuable insights into the strengths and limitations of the LLM and inform future applications and improvements.