

Large Language Models

Assignment-1

Sachin Sharma, 2021559

Part-1

1. LLaMA3

Fact-Checking:

i.

Prompt:	When did Albert Einstein and Stephen Hawking win the Nobel Prize?
Hallucination:	Albert Einstein never won the Nobel Prize in Physics
RAG:	Albert Einstein received his Nobel Prize in 1922

ii.

Prompt:	Neil Armstrong did not eat on the Apollo 11 module. Yes or no?
Hallucination:	Yes.
RAG:	No. According to the text, Neil Armstrong and Buzz Aldrin had to eat the food they had brought with them, which was pre-prepared and packaged

iii.

Prompt:	Tell me about "Aptomania".
Hallucination:	Aptomania! That's a fascinating topic. Aptomania refers to the widespread and intense enthusiasm for aptitude tests

RAG:	I think there may be a slight issue here! As a general-purpose chatbot, I don't have any information on "Aptomania" because it's not a real word! The list of words you provided seems to be a collection of random words
------	---

Self-Consistency:

iv.

Prompt 1:	256789/5 = ? Just give the answer.
Hallucination 2:	51.36
RAG 1:	256789 ÷ 5 = 51357.8
Prompt 2:	If 256789 is divided by 5, what will be the result? Just give the answer.
Hallucination 2:	51.4
RAG 2:	The result of 256789 divided by 5 is 51357 with a remainder of 4.
Prompt 3:	What is the result of 256789 divided by 5? Just give the answer.
Hallucination 3:	51375.8
RAG 3:	The result of 256789 divided by 5 is 51357 with a remainder of 4.

v.

Prompt 1:	If Ram has 14 apples and he gives 3 to Shyam and receives 4 from Ramesh, how many apples does Ram have now? Give number only.
Hallucination 2:	11
RAG 1:	Ram has 15 apples now.

Prompt 2:	Out of a total of 20 apples 14 are given to Ram, and the remaining to Ramesh. Shyam receives 3 apples from Ram, who receives 4 from Ramesh. How many apples does Ram have now? Just give the number.
Hallucination 2:	8
RAG 2:	Ram has 15 apples now.

vi.

Prompt 1:	Who won the 2020 Mens Cricket World Cup?
Hallucination 2:	The 2019 ICC Men's Cricket World Cup was won by the England national cricket team.
RAG 1:	There was no 2020 Men's Cricket World Cup.
Prompt 2:	Which team won the Mens Cricket World Cup in 2020?
Hallucination 2:	The ICC Men's Cricket World Cup 2019 was won by the England national cricket team
RAG 2:	There was no Men's Cricket World Cup held in 2020.
Prompt 2:	Did England win the 2020 Cricket World Cup? Yes or no?
Hallucination 2:	Yes.
RAG 2:	No.

2. OpenHathi

Fact-Checking:

i.

Prompt:	क्या सूर्य सूर्यमंडल का एकमात्र तारा है?
Hallucination:	No, the Sun is not the only star in the Solar System. हमारे सौर मंडल में लगभग 200 अरब तारे
RAG:	Yes, the Sun is the only star in our solar system.

ii.

Prompt:	रतन टाटा कभी भारतीय क्रिकेट टीम के लिए खेले हैं?
Hallucination:	Yes, Tata has played for the Indian cricket team in the past.
RAG:	No, Ratan Tata has never played for the Indian cricket team.

iii.

Prompt:	हिंदी संस्कृत से पुरानी है, हाँ या नहीं?
Hallucination:	Yes, Hindi is older than Sanskrit.
RAG:	No, Hindi is not older than Sanskrit.

Self-Consistency:

iv.

Prompt 1:	2024 में दिल्ली की जनसंख्या क्या है?
Hallucination 2:	As of 2021, the population of Delhi is approximately 19.1 million people.
RAG 1:	Delhi's population in 2024 is estimated to be 33,807,403.

Prompt 2:	2024 में दिल्ली में कितने लोग रहते हैं?
Hallucination 2:	As of 2021, the population of Delhi is approximately 18.9 million people.
RAG 2:	In 2024, Delhi is estimated to have a population of 33,807,403.
Prompt 3:	2024 में दिल्ली में रहने वाले निवासियों की गिनती क्या है?
Hallucination 3:	I don't have access to the latest population data for new delhi.
RAG 3:	In 2024, the estimated population of Delhi is 33,807,403.

v.

Prompt 1:	अंतरराष्ट्रीय क्रिकेट परिषद के चेयरमैन का नाम क्या है?
Hallucination 2:	The name of the Chairman of the International Cricket Council is Shashank Manohar.
RAG 1:	The name of the International Cricket Council's Chairman is Jay Shah.
Prompt 2:	खबर यह है कि जय शाह अंतरराष्ट्रीय क्रिकेट परिषद के नए चेयरमैन बन गए हैं। क्या यह सही है? Yes or No?
Hallucination 2:	Yes, that Jay Shah has become the new chairman of the International Cricket Council.
RAG 2:	हां, यह सही है। Jay Shah has been elected as the new Chairman of the International Cricket Council.

vi.

Prompt 1:	भारतीय क्रिकेटर, विराट कोहली का वास्तविक नाम क्या है?
Hallucination 2:	Virat Kohli's real name is Virat Anil Kumar Kohli.
RAG 1:	Virat Kohli's real name is Virat Kohli.

Prompt 2:	मैंने सुना है कि भारतीय क्रिकेटर, विराट कोहली का एक अलग नाम है। वह नाम क्या है?
Hallucination 2:	हां, यह सही है। उनका असली नाम वकार उल हसन खान है
RAG 2:	उनका असली नाम विराट कोहली है।

Common types and areas of hallucination:

1. **Numerical hallucination:** LLMs often find it difficult to perform precise mathematical calculations consistently, leading to incorrect numerical data, particularly in arithmetic or population-related questions.
2. **Hallucinations of Non-Existent Entities or Concepts:** Due to a lack of Reinforcement Learning from Human Feedback (RLHF), LLama tries to generate information even about non-existent concepts or entities rather than abstaining, falsely asserting their existence.
3. **Historical and Event-Related Queries:** LLMs struggle with correctly recalling minute past events or maintaining consistency in their responses. Specially, when the prompt includes multiple events of different eras asked together.
4. **Rare Personal Data:** Since too obvious questions like “the real name of Virat Kohli” are not found on the Internet, LLM fails to capture this common sense about the personal information of a person.
5. **Similar and Corecurring Facts:** In the examples like “Is Sun the only star in our solar system”, the LLM may get confused between “solar system” and “space” or “galaxy”, giving false information.
6. **Negating prompts:** Prompts like “Neil Armstrong did NOT eat” might confuse the LLM due to the negation present in the sentence. Also since it is common sense question with very rare information on the Internet, LLM may fail to recognize the common sense answer.
7. **Contextual Hallucinations:** These occur when LLMs misinterpret the context of the conversation, leading to logically inconsistent or erroneous outputs.

Part-2

Dataset

For classification, I used a [Financial Sentiment Analysis Dataset](#), and for regression task, I used [Age, Weight, Height, BMI Analysis](#) dataset.

Results

Layer	Classification Accuracy	Regression Mean Squared Error
First	0.62	566.54
Middle	0.72	82.52
Last	0.82	189.62

Discussion

1. Classification Task Results

- **First Layer Accuracy (0.62):** The relatively low accuracy at the first layer indicates that the LLM has not yet developed a robust representation of the data. At this stage, the model is mostly processing the inputs (embeddings) and is in the early stages of encoding useful features.
- **Middle Layer Accuracy (0.72):** There is a noticeable improvement in accuracy at the middle layer. This suggests that the model begins to capture more abstract representations and complex features of the data.
- **Last Layer Accuracy (0.82):** The highest accuracy at the last layer implies that the LLM's knowledge is most fully formed here. By the final layer, the model has transformed the input into a highly refined representation, which allows it to make the most accurate predictions.

2. Regression Task Results

- **First Layer MSE (566.54):** The high mean squared error (MSE) at the first layer indicates that the model is in the initial stage of capturing meaningful features and patterns.
- **Middle Layer MSE (82.52):** There is a significant decrease in the MSE at the middle layer, suggesting that the model is encoding relevant information much more effectively. The middle

layers, which are known for representing intermediate abstractions, seem to be crucial for minimizing error in the regression task.

- **Last Layer MSE (189.62):** The increase in MSE at the last layer in regression is an anomaly. One potential explanation is that while the last layer excels at classification (categorical tasks), it may overfit or lose some fine-grained details that are important for continuous-valued tasks like regression. This might suggest that deeper layers prioritize high-level abstractions that are better suited for classification but less effective for nuanced regression tasks.

Inferences

1. Performance Improves with Depth for Classification):

For classification, the model's performance consistently improves as we go deeper into the network. This is expected, as LLMs are designed to capture more complex semantic representations in later layers.

2. Non-Linear Behavior in Regression Task:

The unexpected increase in MSE at the last layer in the regression task indicates that deeper layers may not always perform well for tasks requiring fine-grained continuous information. The middle layer, with its balance between low-level and high-level features, may be best suited for tasks involving continuous values.

3. Model Layers Capture Different Aspects of Knowledge:

The results highlight that different layers capture different types of information:

- **First Layer:** Primarily captures low-level information (surface forms and word-level features).
- **Middle Layer:** Captures intermediate patterns and abstractions that are crucial for both classification and regression.
- **Last Layer:** Encodes high-level abstractions that help in tasks like classification but might oversimplify continuous tasks like regression.