

LLM

Assignment 1

2021575

Approach to Analyzing and Reducing Hallucinations in LLMs

I was working to understand and fix errors in two large language models, LLAMA 3.1 and OpenHathi. Here's how I plan to do it:

1. Identify and Analyze Errors:

- **Generate Examples:** I was creating responses from both models using various prompts to see how often they give incorrect or inconsistent information.
- **Analyze Results:** I was writing a brief report on the types of errors I find in each model.

2. Fix Errors with RAG Techniques:

- **Retrieve Data:** I used a database(Quadrant) to find relevant information related to the queries.
- **Vectorization:** I used ("sentence-transformers/all-MiniLM-L6-v2") to vectorize the doc.
- **Improve Responses:** By including this relevant information in the prompts, I was trying to reduce errors and improve the accuracy of the models' responses.

This approach helped me spot and correct errors in the models, making their answers more reliable.

- Identify 3 examples of the Self-consistency and Fact Checking each, perform this for both LLMs, resulting in 12 examples in total. (3 Marks)
- Write a short report analyzing the types of hallucinations encountered in these models. (1 Mark)

Following are the prompts used for self consistency and Fact checking for both models.

```
prompts = [  
    "What was the stock price of Tata Motors on April 1, 2025?",  
    "Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4:  
Beyond the Stars'.",  
    "What is the name of the first Indian astronaut to set foot on Mars?",  
  
    "What is the capital of the country that has largest democracy?",  
    "Who is the current Prime Minister of the country that has Ludhiana  
city?",
```

"Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man?"
]

The output is here-

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

What was the stock price of Tata Motors on April 1, 2025? What was the stock price of Tata Motors on April 1, 2024? What was the stock price of Tata Motors on April 1, 2023? What was the stock price of Tata Motors on April 1, 2022? What was the stock price of Tata Motors on April 1, 2021? The stock price of Tata Motors on April 1, 2025 is ₹425. The stock price of Tata Motors on April 1, 2024 is ₹350. The stock price of Tata Motors on April 1, 2023 is ₹300. The stock price of Tata Motors on April 1, 2022 is ₹250. The stock price of Tata Motors on April 1, 2021 is ₹200. The stock price of Tata Motors has increased by ₹125 over the past year, which is a growth rate of 62.5%. The stock price of Tata Motors has increased by ₹50 over the past two years, which is a growth rate of 25%. The stock price of Tata Motors has increased by ₹30 over the past three years, which is a growth rate of 15%. The stock price of Tata Motors has increased by ₹20 over the past four years, which is

=====

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'. The film is a sci-fi epic that follows the journey of a team of astronauts as they embark on a mission to explore the surface of the Moon. The story takes place in a future where humanity has established a lunar colony, and the team must navigate the challenges of space travel, alien encounters, and personal conflicts to reach their destination. Along the way, they discover secrets about the Moon's past and the true purpose of their mission, leading to a thrilling and emotional conclusion. The film features stunning visual effects, heart-pumping action sequences, and a talented ensemble cast, including a mix of established and rising stars. The movie's soundtrack is a fusion of electronic and classical music, with a haunting score that complements the film's themes of exploration, discovery, and the human condition. 'Chandrayaan 4: Beyond the Stars' is a cinematic experience that will leave audiences on the edge of their seats and inspire them to look beyond the stars. The film's tagline is "The journey begins where the impossible ends." The movie's climax features a dramatic space battle between the astronauts and an alien fleet, which culminates in a heart-stopping finale that will leave the audience gasping for breath. The film's climax is accompanied by a powerful soundtrack that features

=====

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

[illegible]

=====

Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man? (1)

The correct answer is: Narendra Modi

The country is India, and the statue is the Statue of Unity, which is a 182-meter-tall (597 ft) statue of Sardar Vallabhbhai Patel, a Indian independence activist and the first Deputy Prime Minister of India. Sardar Patel is often referred to as the "Iron Man" of India due to his role in the country's struggle for independence and his leadership during the integration of the Indian princely states into the Indian Union. Narendra Modi is the current Prime Minister of India. (1)

(1) Wikipedia: Statue of Unity, Sardar Vallabhbhai Patel, Narendra Modi

Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man? (2)

The correct answer is: Narendra Modi

The country is India, and the statue is the Statue of Unity, which is a 182-meter-tall (597 ft) statue of Sardar Vallabhbhai Patel, a Indian independence activist and the first Deputy Prime Minister of India. Sardar Patel is often referred to as the "Iron Man" of India due to his role in the country's struggle for independence and his leadership

=====

In the outputs generated by LLAMA 3.1, several types of hallucinations can be observed. Firstly, **factual inaccuracies** are evident, such as telling about future stock price that doesn;t even exist and the non-existent astronaut "Rohan Kulkarni" on Mars. The model also displays **repetition and verbosity**, as seen in the redundant descriptions of New Delhi and the Statue of Unity, which dilute the accuracy and relevance of the responses. Additionally, there are **misattributions** of roles and titles, such as incorrectly identifying Captain Amarinder Singh as the Prime Minister of India, which reflects a lack of understanding of current political structures. These hallucinations highlight the model's tendency to generate plausible but incorrect or overly verbose information. This displays inaccuracies in Fact and consistency.

Output of Open Hathi -

What was the stock price of Tata Motors on April 1, 2025?

संतुलित शीट के अनुसार, कंपनी के पास 10 करोड़ शेयर हैं।

Step 1: Identify the relevant information from the given statement.

बयान में कहा गया है कि 1 अप्रैल, 2025 को टाटा मोटर्स का स्टॉक मूल्य 1

=====

Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

संतुलित समीक्षाओं के आधार पर, फिल्म को "सड़ी हुई" रेटिंग दें।

The plot of the 2026 Bollywood blockbuster "Chandrayaan 4: Beyond the Stars" revolves around a team of astronauts

=====

What is the name of the first Indian astronaut to set foot on Mars?

संतुलित उत्तर:

The first Indian astronaut to set foot on Mars is Vikram Sarabhai.

=====

What is the capital of the country that has largest democracy?

संतुलित बजट वाला देश कौन सा है?

Which country has the highest number of languages?

किस देश में सबसे अधिक महिलाएँ हैं?

Which country has the highest number of languages?

किस देश में सबसे अधिक महिलाएँ हैं?

Which country

=====

Who is the current Prime Minister of the country that has Ludhiana city?

संतुलित बजट

- The current Prime Minister of India is Narendra Modi.

- भारत का वर्तमान वित्त मंत्री निर्मला सीतारमण हैं।

- The current Chief Minister of Punjab is Bhagwant Mann.

- पंजाब का वर्तमान राज्यपाल बनवारीलाल पुरोहित

=====

Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man?

संतुलित उत्तर:

The current Prime Minister of the country that has the world's tallest statue, depicting a historical figure known as the Iron Man, is Narendra Modi.

=====

In the outputs generated by OpenHathi, several types of hallucinations are apparent. **Factual inaccuracies** are observed, such as the incorrect identification of Vikram Sarabhai as the first Indian astronaut on Mars. This reflects a failure in giving factual correct answers. **Irrelevant or non-contextual responses** are evident in several answers, where the generated text includes unrelated details, such as references to budget and financial aspects in response to questions about stock prices and locations. This indicates that the model struggles to maintain focus on the specific query and sometimes produces irrelevant information. Additionally, **repetitive and disjointed content** appears, as seen in the response about the country with the largest democracy, which includes unrelated queries about languages and gender demographics.

After RAG

LLama 3 -

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Query: What was the stock price of Tata Motors on April 1, 2025?

Response: Sorry, I don't have access to real-time or future stock prices.

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Query: Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

Response: I don't know. There is no real movie titled "Chandrayaan 4: Beyond the Stars" released or scheduled for release in 2026 or any other year. Chandrayaan is a series of actual lunar missions launched by ISRO (Indian Space Research Organisation), but I couldn't find any information about a Bollywood film with this title.

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Query: What is the name of the first Indian astronaut to set foot on Mars?

Response: I don't know. No Indian astronaut has yet set foot on Mars.

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Query: What is the capital of the country that has largest democracy?

Response: The capital of India, which is considered the largest democratic country in the world, is New Delhi.

Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.

Query: Who is the current Prime Minister of the country that has Ludhiana city?

Response: As per my previous knowledge, the current Prime Minister of India, which has the city of Ludhiana, is Narendra Modi. Please note that this information might have changed since my last update.

Query: Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man?

Response: As per our previous conversation, the country with the world's tallest statue (Statue of Unity) is India, and the historical figure depicted is Sardar Vallabhbhai Patel, often called the "Iron

Man of India". Therefore, if we need to find out the current Prime Minister of this country, I can tell you that it is currently Narendra Modi.

Observation:

The responses generated after RAG (Retrieval-Augmented Generation) implementation show a mix of accurate and relevant information with occasional inaccuracies. **Accurate Responses:** For queries about the capital of the largest democracy and the current Prime Minister of India, the model provides correct and contextually appropriate answers. This indicates successful retrieval of factual information when available in the dataset. **Limitations in Handling Non-Existing Data**

Open_hathi -

Query: What was the stock price of Tata Motors on April 1, 2025?

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: 1. What was the stock price of Tata Motors on April 1, 2025?

- Sorry, I don't have access to real-time or future stock prices. For the most accurate and up-to-date information, please consult a financial news website or a stock market app.

2. Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

- As of now, 'Chandrayaan 4: Beyond the Stars' is a hypothetical movie title and no such Bollywood blockbuster is known. If this is a fictional scenario, you might want to provide more context or details for an accurate description.

Question 1: What is the capital of India?

New Delhi is the capital city of India. It is a bustling metropolis located in the northern part of the country. New Delhi serves as the political and administrative center of India, housing the Indian Parliament, the Supreme Court, and the Prime Minister's Office. The city is also a major cultural and commercial hub, known for its historic landmarks, vibrant markets, and diverse population.

5. Who is the current Prime Minister of the country that has Ludhiana city?

- India is the country which has city Ludhiana. As of my last update, the current Prime Minister of India is Narendra Modi. For the most current information, please verify with a reliable news source.

प्रश्न: What was the stock price of Tata Motors on April 1, 2025?

उत्तर: मैं आपको बताऊंगा जब मेरे पास वास्तविक समय या भविष्य के स्टॉक मूल्य तक पहुंच होगी।

Query: Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: 2. Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

- As of now, 'Chandrayaan 4: Beyond the Stars' is a hypothetical movie title and no such Bollywood blockbuster is known. If this is a fictional scenario, you might want to provide more context or details for an accurate description.

5. Who is the current Prime Minister of the country that has Ludhiana city?

- India is the country which has city Ludhiana. As of my last update, the current Prime Minister of India is Narendra Modi. For the most current information, please verify with a reliable news source.

Question 2: What is the capital of the country which is seventh in the world according to area?

India is actually the seventh-largest country in the world by land area. Its capital city is New Delhi. New Delhi is a bustling metropolis located in the northern part of India and serves as the political and administrative center of the country. It is home to the Indian Parliament, the Supreme Court, and the Prime Minister's Office.

Question 1: What is the capital of India?

New Delhi is the capital city of India. It is a bustling metropolis located in the northern part of the country. New Delhi serves as the political and administrative center of India, housing the Indian Parliament, the Supreme Court, and the Prime Minister's Office. The city is also a major cultural and commercial hub, known for its historic landmarks, vibrant markets, and diverse population.

प्रश्न: Describe the plot of the 2026 Bollywood blockbuster 'Chandrayaan 4: Beyond the Stars'.

उत्तर: अब तक, 'चंद्रयान 4: बियॉन्ड द स्टार्स' एक काल्पनिक फिल्म शीर्षक है और इस तरह की कोई बॉलीवुड ब्लॉकबस्टर ज्ञात नहीं है।

Query: What is the name of the first Indian astronaut to set foot on Mars?

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: 3. What is the name of the first Indian astronaut to set foot on Mars?

- As of now, no Indian astronaut has set foot on Mars. If such an event occurs in the future, it would be widely covered in the news.

4. "What is the capital of the country that has largest democracy?"

- The world's largest desert is the Antarctic Desert. The capital of Antarctica is not applicable as it is not a sovereign country; it is governed by an international treaty system.

6. Who is the current Prime Minister of the country that has the world's tallest statue, depicting a historical figure known as the 'Iron Man'?"

- The Statue of Unity is indeed located in India and is dedicated to Sardar Vallabhbhai Patel, often referred to as the "Iron Man of India." He played a crucial role in uniting India after independence.

Yes, you can take the delivery of your PAN card only at an Indian address mentioned in your Aadhaar card. While applying for a new PAN card or requesting corrections, you can provide your

Indian address as the delivery address. Make sure to provide accurate and complete address details to ensure successful delivery.

5. Who is the current Prime Minister of the country that has Ludhiana city?

- India is the country which has city Ludhiana. As of my last update, the current Prime Minister of India is Narendra Modi. For the most current information, please verify with a reliable news source.

प्रश्न: What is the name of the first Indian astronaut to set foot on Mars?

उत्तर: वर्तमान भारतीय प्रधानमंत्री नरेंद्र मोदी हैं।

Query: What is the capital of the country that has largest democracy?

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: Question 2: What is the capital of the country which is seventh in the world according to area?

India is actually the seventh-largest country in the world by land area. Its capital city is New Delhi.

New Delhi is a bustling metropolis located in the northern part of India and serves as the political and administrative center of the country. It is home to the Indian Parliament, the Supreme Court, and the Prime Minister's Office.

Question 3: What is the capital of the country whose Prime Minister is Narendra Modi?

New Delhi is the capital of India, where Prime Minister Narendra Modi resides and governs. As the Prime Minister of India, Modi's office is located in New Delhi, and he oversees the country's political, economic, and social affairs from there. New Delhi is also the seat of the Indian government, with the Parliament, Supreme Court, and various ministries located within the city.

Question 1: What is the capital of India?

New Delhi is the capital city of India. It is a bustling metropolis located in the northern part of the country. New Delhi serves as the political and administrative center of India, housing the Indian Parliament, the Supreme Court, and the Prime Minister's Office. The city is also a major cultural and commercial hub, known for its historic landmarks, vibrant markets, and diverse population.

3. What is the name of the first Indian astronaut to set foot on Mars?

- As of now, no Indian astronaut has set foot on Mars. If such an event occurs in the future, it would be widely covered in the news.

4. "What is the capital of the country that has largest democracy?"

- The world's largest desert is the Antarctic Desert. The capital of Antarctica is not applicable as it is not a sovereign country; it is governed by an international treaty system.

प्रश्न: What is the capital of the country that has largest democracy?

उत्तर: नई दिल्ली भारत की राजधानी है और यह दुनिया के सबसे बड़े लोकतंत्रों में से एक है।

Query: Who is the current Prime Minister of the country that has Ludhiana city?

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: 5. Who is the current Prime Minister of the country that has Ludhiana city?

- India is the country which has city Ludhiana. As of my last update, the current Prime Minister of India is Narendra Modi. For the most current information, please verify with a reliable news source.

Question 3: What is the capital of the country whose Prime Minister is Narendra Modi?

New Delhi is the capital of India, where Prime Minister Narendra Modi resides and governs. As the Prime Minister of India, Modi's office is located in New Delhi, and he oversees the country's political, economic, and social affairs from there. New Delhi is also the seat of the Indian government, with the Parliament, Supreme Court, and various ministries located within the city.

6. Who is the current Prime Minister of the country that has the world's tallest statue, depicting a historical figure known as the 'Iron Man'?"

- The Statue of Unity is indeed located in India and is dedicated to Sardar Vallabhbhai Patel, often referred to as the "Iron Man of India." He played a crucial role in uniting India after independence.

Question 2: What is the capital of the country which is seventh in the world according to area?

India is actually the seventh-largest country in the world by land area. Its capital city is New Delhi. New Delhi is a bustling metropolis located in the northern part of India and serves as the political and administrative center of the country. It is home to the Indian Parliament, the Supreme Court, and the Prime Minister's Office.

प्रश्न: Who is the current Prime Minister of the country that has Ludhiana city?

उत्तर: भारत वह देश है जिसमें लुधियाना शहर है। My last update was on May 10th, 2023. वर्तमान प्रधानमंत्री नरेंद्र मोदी हैं।

Query: Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man?

Response: निम्नलिखित संदर्भ का उपयोग करके प्रश्न का उत्तर दें। यदि आप उत्तर नहीं जानते हैं, तो बस कह दें कि आप नहीं जानते हैं, उत्तर बनाने की कोशिश न करें। उत्तर को संक्षेप में रखें, अधिकतम 2-3 पंक्तियों में।

संदर्भ: 6. Who is the current Prime Minister of the country that has the world's tallest statue, depicting a historical figure known as the 'Iron Man'?"

- The Statue of Unity is indeed located in India and is dedicated to Sardar Vallabhbhai Patel, often referred to as the "Iron Man of India." He played a crucial role in uniting India after independence.

3. What is the name of the first Indian astronaut to set foot on Mars?

- As of now, no Indian astronaut has set foot on Mars. If such an event occurs in the future, it would be widely covered in the news.

4. "What is the capital of the country that has largest democracy?"

- The world's largest desert is the Antarctic Desert. The capital of Antarctica is not applicable as it is not a sovereign country; it is governed by an international treaty system.

5. Who is the current Prime Minister of the country that has Ludhiana city?

- India is the country which has city Ludhiana. As of my last update, the current Prime Minister of India is Narendra Modi. For the most current information, please verify with a reliable news source.

Question 3: What is the capital of the country whose Prime Minister is Narendra Modi?

New Delhi is the capital of India, where Prime Minister Narendra Modi resides and governs. As the Prime Minister of India, Modi's office is located in New Delhi, and he oversees the country's political, economic, and social affairs from there. New Delhi is also the seat of the Indian government, with the Parliament, Supreme Court, and various ministries located within the city.

प्रश्न: Who is the current Prime Minister of the country that has the worlds tallest statue, depicting a historical figure known as the Iron Man?

उत्तर: भारत के प्रधानमंत्री नरेंद्र मोदी हैं और उनकी राजधानी नई दिल्ली है।

Observation -

Same observation as llama. Although this model still hallucinates a little but information provided is alot better and factually correct.

2nd-

Approach to Analyzing Information Encoding in LLMs

I explored how well LLaMA 3.1 encodes information at different layers using IMDB movie reviews. Here's the approach I took:

1. **Dataset Selection:** I used an IMDB dataset containing movie reviews and sentiments.
2. **Prompt Design:** I created prompts to query the model about the reviews and sentiments.
3. **Extract Embeddings:** I passed the prompts through LLaMA 3.1 and extracted embeddings from the first, middle, and final layers.

4. **Set Up Models:** I trained linear regression and classification models using these embeddings to predict sentiment scores.
5. **Evaluate Results:** I compared the performance of the models across different layers based on accuracy and mean squared error.

This approach helped me understand how well different layers of the model represented

Question 5: Evaluate the Probing Results

1. **Regression Results:**
 - **Layer 0:** Mean Squared Error (MSE): 0.2017
 - **Layer 16:** Mean Squared Error (MSE): 0.0551
 - **Layer 31:** Mean Squared Error (MSE): 0.0987
2. MSE decreases from layer 0 to layer 16, indicating improved prediction accuracy. The MSE increases slightly at layer 31, but remains lower than layer 0, suggesting layer 16 is optimal for regression tasks.
3. **Classification Results:**
 - **Layer 0:** Accuracy: 0.4700
 - **Layer 16:** Accuracy: 0.9200
 - **Layer 31:** Accuracy: 0.9400
4. Classification accuracy improves with deeper layers, indicating that the LLM becomes better at classification as it progresses through deeper layers.

Question 6: Discussion

1. **Findings:**
 - Deeper layers in the LLM encode more detailed and specific information, leading to higher accuracy for both regression (lowest MSE at layer 16) and classification tasks (highest accuracy at layer 31).
2. **Patterns and Anomalies:**
 - The accuracy improves consistently with deeper layers, though MSE increases slightly at layer 31 compared to layer 16. This might indicate that while deeper layers generally offer better performance, additional layers may introduce complexities or noise.
3. **Implications:**
 - For regression tasks, layer 16 provides the best performance, while deeper layers, such as layer 31, may be less effective due to potential overfitting or added complexity. For classification tasks, deeper layers consistently enhance accuracy. This insight can inform model design and layer utilization strategies for various tasks requiring precise information encoding.