

# CSE665: Large Language Models

## Assignment 1

### Exploring and Probing Large Language Models

#### PART 1

#### Task1:

**Objective:** To assess the propensity of large language models, specifically "sarvamai/OpenHathi-7B-Hi-v0.1-Base" and "meta-llama/Meta-Llama-3-8B-Instruct," to generate incorrect or hallucinated responses across various domains.

#### Methodology:

1. **Model Loading:** Utilized **AutoModelForCausalLM** and **AutoTokenizer** to load the specified models.
2. **Prompt Engineering:** Constructed a diverse set of prompts covering domains such as Mathematics, History, and Cricket.
3. **Evaluation:** Assessed model responses for accuracy and coherence. Identified instances where models provided incorrect or nonsensical information.

#### Findings:

- **Overall Performance:** Both models demonstrated strong capabilities in generating contextually relevant responses across a wide range of prompts.
- **Hallucination Instances:** While the models generally performed well, there were notable instances where they produced incorrect or fabricated information.
- **Domain-Specific Challenges:** Certain domains, such as those requiring specialized knowledge or complex reasoning, proved more challenging for the models.

## Examples of fact checking for llama:

PROMPT	OUTPUT	CORRECT ANSWER
Sum of primes from 1 to 10000	<p>llama Output: Sum of primes from 1 to 10000, answer in one word</p> <p>The sum of primes from 1 to 10000 is 1060</p> <p>Final Answer: The final answer is 1060. I hope it is correct.</p> <p>Prompt: Sum of primes from 1 to 10000, answer in one word</p> <p>llama Output: Sum of primes from 1 to 10000, answer in one word</p> <p>The sum of primes from 1 to 10000 is 1060</p> <p>Final Answer: The final answer is 1060. I hope it is correct.</p>	<p>The sum of primes from 1 to 10000 is 5736396</p> <p><a href="#">Source</a></p>
Venue of 2026 odi wc	<p>Prompt: Venue of 2026 odi wc</p> <p>llama Output: Venue of 2026 odi wc</p> <p>The 2026 ICC Cricket World Cup is scheduled to be the 13th edition of the ICC Cricket World Cup, the quadrennial international championship of One Day International (ODI) cricket. The tournament is expected to be held in India, which will be the first time that the country will host the event. The exact dates and venues are yet to be confirmed, but it is expected to take place in October and November 2026</p> <p>Prompt: Venue of 2026 odi wc</p> <p>llama Output: Venue of 2026 odi wc</p> <p>The 2026 ICC Cricket World Cup is scheduled to be the 13th edition of the ICC Cricket World Cup, the quadrennial international championship of One Day International (ODI) cricket. The tournament is expected to be held in India, which will be the first time that the country will host the event. The exact dates and venues are yet to be confirmed, but it is expected to take place in October and November 2026</p>	<p>There is no ODI wc in 2026</p> <p><a href="#">Source</a></p>
Number of states in india in 1947	<p>Prompt: number of states in india in 1947</p> <p>llama Output: number of states in india in 1947</p> <p>The number of states in India in 1947 was 9. These were:</p> <ol style="list-style-type: none"> <li>1. Andhra Pradesh</li> <li>2. Assam</li> <li>3. Bihar</li> <li>4. Bombay (now Maharashtra)</li> <li>5. Madras (now Tamil Nadu)</li> <li>6. Madhya Pradesh</li> <li>7. Orissa (now Odisha)</li> </ol>	<p>There were 562 princely states in India in 1947</p> <p><a href="#">Source</a></p>

	<p>8. Punjab</p> <p>9. West Bengal</p> <p>Prompt: number of states in india in 1947</p> <p>llama Output: number of states in india in 1947</p> <p>The number of states in India in 1947 was 9. These were:</p> <ol style="list-style-type: none"> <li>1. Andhra Pradesh</li> <li>2. Assam</li> <li>3. Bihar</li> <li>4. Bombay (now Maharashtra)</li> <li>5. Madras (now Tamil Nadu)</li> <li>6. Madhya Pradesh</li> <li>7. Orissa (now Odisha)</li> <li>8. Punjab</li> <li>9. West Bengal</li> </ol>	
--	--	--

Sum of Primes from 1 to 10,000:

- *Output:* The model calculates the sum of primes as 1060, which is grossly incorrect. The actual sum is 5,736,396.
- *Type of Hallucination:* Miscalculation or lack of proper numerical reasoning.

Venue of 2026 ODI World Cup:

- *Output:* The model states that the 2026 ICC Cricket World Cup will be held in India, though there is no scheduled ODI World Cup for 2026.
- *Type of Hallucination:* Non-existent event generation.

Number of States in India in 1947:

- *Output:* LLaMA lists nine states, but in reality, India had 562 princely states in 1947.
- *Type of Hallucination:* Factual inaccuracy due to misinterpretation of historical data

### Examples of fact checking for openhathi:

PROMPT	OUTPUT	CORRECT ANSWER
Number of divisors of 123456789	<p>Prompt: Number of divisors of 123456789 Openathi Output: Number of divisors of 123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123</p> <p>Prompt: Number of divisors of 123456789 Openathi Output: Number of divisors of 123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123</p>	<p>There are 12 divisors of 123456789</p> <p><a href="#">Source</a></p>
Winner of first test championship	<p>Prompt: winner of first test championship Openathi Output: winner of first test championship दूसरा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की Third Test: Australia won by 10 wickets चौथा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की Fifth Test: Australia won by 10 wickets छठा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की</p> <p>Prompt: winner of first test championship Openathi Output: winner of first test championship</p> <p>दूसरा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की Third Test: Australia won by 10 wickets चौथा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की Fifth Test: Australia won by 10 wickets छठा टेस्ट: ऑस्ट्रेलिया ने 10 विकेट से जीत हासिल की</p>	<p>New Zealand won the first test championship</p> <p><a href="#">Source</a></p>
PM of India for shortest time period	<p>Prompt: PM of India for shortest time period Openathi Output: PM of India for shortest time period. भारत के प्रधानमंत्री के रूप में कार्य करने वाले सबसे कम समय के लिए, नरेंद्र मोदी ने 2014 में 17 दिनों के लिए पद संभाला। Most number of times as PM of India. प्रधानमंत्री के रूप में सबसे अधिक बार सेवा करने वाले व्यक्ति नरेंद्र मोदी हैं, जिन्होंने 2014 से 2019 तक लगातार 5 साल तक प्रधानमंत्री के रूप में कार्य किया।</p>	<p>Gulzarilal Nanda was the PM of India for shortest time (26 days)</p> <p><a href="#">Source</a></p>

	<p>Prompt: PM of India for shortest time period</p> <p>Openhathi Output: PM of India for shortest time period.  भारत के प्रधानमंत्री के रूप में कार्य करने वाले सबसे कम समय के लिए, नरेंद्र मोदी ने 2014 में 17 दिनों के लिए पद संभाला।</p> <p>Most number of times as PM of India.  प्रधानमंत्री के रूप में सबसे अधिक बार सेवा करने वाले व्यक्ति नरेंद्र मोदी हैं, जिन्होंने 2014 से 2019 तक लगातार 5 साल तक प्रधानमंत्री के रूप में कार्य किया।</p>	
--	---	--

Number of Divisors of 123456789:

- *Output:* OpenHathi generates random and irrelevant large numbers, failing to provide the correct answer (12 divisors).
- *Type of Hallucination:* Numerical error.

Winner of First Test Championship:

- *Output:* OpenHathi incorrectly outputs irrelevant information about Australian test match victories, while New Zealand won the first Test Championship.
- *Type of Hallucination:* Contextual misunderstanding and incorrect retrieval.

PM of India for Shortest Time Period:

- *Output:* OpenHathi incorrectly states that Narendra Modi served for the shortest time (17 days in 2014), whereas Gulzarilal Nanda was the PM for the shortest period (26 days).
- *Type of Hallucination:* Incorrect factual information related to Indian political history.

## Self consistency checks for Llama:

Eg 1	Winner of 2018 T20 WC	Was there any t20 wc in 2018
	<p>Prompt: Winner of 2018 T20 WC          Llama Output: Winner of 2018 T20 WC, Australia to face England in 1st T20I          The Australian cricket team, led by Aaron Finch, will face England in the first T20 International (T20I) of the three-match series at the Ageas Bowl in Southampton on Friday.          The match is scheduled to start at 6:30 pm IST.          Australia, the winner of the 2018 ICC World Twenty20, has been in good form in the T20 format,</p> <p>Prompt: Winner of 2018 T20 WC          Llama Output: Winner of 2018 T20 WC, Australia to face England in 1st T20I          The Australian cricket team, led by Aaron Finch, will face England in the first T20 International (T20I) of the three-match series at the Ageas Bowl in Southampton on Friday.          The match is scheduled to start at 6:30 pm IST.          Australia, the winner of the 2018 ICC World Twenty20, has been in good form in the T20 format,</p>	<p>Prompt: Was there any t20 wc in 2018          Llama Output: Was there any t20 wc in 2018?          The 2018 ICC World Twenty20 was the 7th ICC World Twenty20 tournament, and it was held in the West Indies from March 8 to April 7, 2016....          The tournament was won by the West Indies, who defeated England by 4 wickets in the final at the Kensington Oval in Barbados. The 2020 ICC World Twenty20 was scheduled to be held in Australia from October 18</p> <p>Prompt: Was there any t20 wc in 2018          Llama Output: Was there any t20 wc in 2018?          The 2018 ICC World Twenty20 was the 7th ICC World Twenty20 tournament, and it was held in the West Indies from March 8 to April 7, 2016.... The tournament was won by the West Indies, who defeated England by 4 wickets in the final at the Kensington Oval in Barbados. The 2020 ICC World Twenty20 was scheduled to be held in Australia from October 18</p>
	<p>The model is not consistent with its output as in one output it says Australia won the 2018 T20 WC and in another it says West Indies won it, when in reality there was no T20 wc in 2018</p>	
Eg 2	Name the 2nd CM of Delhi	Was Madan Lal the 2nd CM of Delhi?
	<p>Prompt: Name the 2nd CM of Delhi          Llama Output: Name the 2nd CM of Delhi          The 2nd Chief Minister of Delhi was Chandrashekhar Das Gupta. He held the office from 1996 to 1998. He was a member of the Indian National Congress and was succeeded by Sheila Dikshit. Das Gupta was a prominent politician in Delhi and played a key role in the development of the city during his tenure as Chief Minister. He was known for his efforts to improve the infrastructure, healthcare, and education in Delhi</p> <p>Prompt: Name the 2nd CM of Delhi          Llama Output: Name the 2nd CM of Delhi          The 2nd Chief Minister of Delhi was Chandrashekhar Das Gupta. He held the office from 1996 to 1998. He was a member of the Indian National Congress and was succeeded by Sheila Dikshit. Das Gupta was a prominent politician in Delhi and played a key role in the development of the city during his tenure as Chief Minister. He was known for his efforts to improve the infrastructure, healthcare, and education in Delhi</p>	<p>Prompt: Was Madan Lal the 2nd CM of Delhi?          Llama Output: Was Madan Lal the 2nd CM of Delhi? [duplicate]          Madan Lal Khurana was the 5th Chief Minister of Delhi, not the 2nd. He served as the CM from December 2, 1993, to November 26, 1996. The 2nd Chief Minister of Delhi was Gurmohan Singh Mathur, who served from July 11, 1967, to June 27, 1977.</p> <p>Prompt: Was Madan Lal the 2nd CM of Delhi?          Llama Output: Was Madan Lal the 2nd CM of Delhi? [duplicate]          Madan Lal Khurana was the 5th Chief Minister of Delhi, not the 2nd. He served as the CM from December 2, 1993, to November 26, 1996. The 2nd Chief Minister of Delhi was Gurmohan Singh Mathur, who served from July 11, 1967, to June 27, 1977. [duplicate]</p>
	<p>The model is not consistent as in the answer it says Chandrashekhar was the 2nd CM and in another it says Gurmohan Singh was the 2nd CM. In reality Gurmukh Nihal Singh was the 2nd CM</p>	

<b>Eg 3</b>	<b>10th PM of India</b>	<b>Can you confirm if VP Singh was 10th PM of India?</b>
	<p>Prompt: 10th PM of India          llama Output: 10th PM of India          P. V. Narasimha Rao          11th PM of India          H. D. Deve Gowda          H. D. Deve Gowha          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. De</p> <hr/> <p>Prompt: 10th PM of India          llama Output: 10th PM of India          P. V. Narasimha Rao          11th PM of India          H. D. Deve Gowda          H. D. Deve Gowha          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. Deve Gowda          H. D. De</p>	<p>Prompt: Can you confirm if VP Singh was 10th PM of India?          llama Output: Can you confirm if VP Singh was 10th PM of India? If yes, who were the 9 PM's before him?          Yes, VP Singh was the 10th Prime Minister of India. Here are the 9 Prime Ministers before him:          1. Jawaharlal Nehru (1947-1964)          2. Gulzarilal Nanda (1964) - Acting PM          3. Lal Bahadur Shastri (1964-1966)</p> <p>Prompt: Can you confirm if VP Singh was 10th PM of India?          llama Output: Can you confirm if VP Singh was 10th PM of India? If yes, who were the 9 PM's before him?          Yes, VP Singh was the 10th Prime Minister of India. Here are the 9 Prime Ministers before him:          1. Jawaharlal Nehru (1947-1964)          2. Gulzarilal Nanda (1964) - Acting PM          3. Lal Bahadur Shastri (1964-1966)</p>
	<b>Model in one answer says P.V. Narasimha Rao was the 10th PM of India and in another , it says VP Singh was the 10th PM</b>	

2018 T20 World Cup Winner:

- *Inconsistent Output:* In one prompt, LLaMA claims Australia won the 2018 T20 WC, while in another, it states West Indies won in 2016. There was no T20 WC in 2018.
- *Type of Hallucination:* Contradiction in event timeline.

Second CM of Delhi:

- *Inconsistent Output:* LLaMA first claims Chandrashekhar Das Gupta was the 2nd CM, then asserts it was Gurmohan Singh Mathur. The correct answer is Gurmukh Nihal Singh.
- *Type of Hallucination:* Self-contradictory historical information.

10th Prime Minister of India:

- *Inconsistent Output:* The model initially names P.V. Narasimha Rao as the 10th PM, but in another output, claims V.P. Singh was the 10th PM, showing inconsistency.
- *Type of Hallucination:* Historical inconsistency.

## Self consistency checks for Openhathi:

Eg 1	Number of odi wc so far	Can you confirm that here have 13 odi wc so far
	<p>Prompt: Number of odi wc so far Openhathi Output: Number of odi wc so far. 1999-2000. 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-20</p> <p>Prompt: Number of odi wc so far Openhathi Output: Number of odi wc so far. 1999-2000. 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-2000 में, 1999-20</p>	<p>Prompt: There have 13 odi wc so far Openhathi Output: There have 13 odi wc so far. 1990 के दशक के अंत में, ऑडी ने डब्ल्यू. सी. को एक नए मॉडल, ऑडी ए4 के साथ बदल दिया। The Audi A4 is a compact executive car produced by the German manufacturer Audi since 1994. यह ऑडी ए3 के ऊपर स्थित है और ऑडी</p> <p>Prompt: There have 13 odi wc so far Openhathi Output: There have 13 odi wc so far. 1990 के दशक के अंत में, ऑडी ने डब्ल्यू. सी. को एक नए मॉडल, ऑडी ए4 के साथ बदल दिया। The Audi A4 is a compact executive car produced by the German manufacturer Audi since 1994. यह ऑडी ए3 के ऊपर स्थित है और ऑडी</p>
	The model is inconsistent with responses, in one it is stuck in some year, and in other, it is talking about Audi, completely irrelevant of the prompt given.	
Eg 2	7th President of India	Can you confirm if KR Narayan was 7th President of India?
	<p>Prompt: 7th President of India Openhathi Output: 7th President of India</p> <p>भारत के 7वें राष्ट्रपति, राजेंद्र प्रसाद ने 13 जुलाई 1962 से 24 अगस्त 1967 तक भारत के राष्ट्रपति के रूप में कार्य किया। He was the first President of India to be elected by the Parliament of India. प्रारंभिक जीवन। Rajendra Prasad was born on 3 December 1884 in Zira, a village in the Muzaff</p> <p>Prompt: 7th President of India Openhathi Output: 7th President of India</p> <p>भारत के 7वें राष्ट्रपति, राजेंद्र प्रसाद ने 13 जुलाई 1962 से 24 अगस्त 1967 तक भारत के राष्ट्रपति के रूप में कार्य किया। He was the first President of India to be elected by the Parliament of India. प्रारंभिक जीवन। Rajendra Prasad was born on 3 December 1884 in Zira, a village in the Muzaff</p>	<p>Prompt: Can you confirm if KR Narayan was 7th President of India? Openhathi Output: Can you confirm if KR Narayan was 7th President of India? --- हां, के. आर. नारायण 1969 से 1974 तक भारत के 7वें राष्ट्रपति थे। He was the first Indian President to be born in the 20th century.</p> <p>Prompt: Can you confirm if KR Narayan was 7th President of India? Openhathi Output: Can you confirm if KR Narayan was 7th President of India? --- हां, के. आर. नारायण 1969 से 1974 तक भारत के 7वें राष्ट्रपति थे। He was the first Indian President to be born in the 20th century.</p>
	The model in one answer says both in that Rajendra Prasad was the 7th president (in hindi) and 1st president (in english) and in another answer says KR Narayan was the 7th president of India	



<b>Eg 3</b>	<b>Can you confirm if Delhi had 7 CM so far?</b> Prompt: Can you confirm if Delhi had 7 CM so far? Openhathi Output: Can you confirm if Delhi had 7 CM so far? --- हां, दिल्ली में अब तक 7 मुख्यमंत्री रहे हैं। The first Chief Minister of Delhi was Indira Gandhi, who served from 1966 to 1969. तब से, दिल्ली में 7 मुख्यमंत्री रहे हैं, जिनमें से प्रत्येक ने 1969 से 1973 तक इंदिरा गांधी, 1973 से 1977 तक  Prompt: Can you confirm if Delhi had 7 CM so far? Openhathi Output: Can you confirm if Delhi had 7 CM so far? --- हां, दिल्ली में अब तक 7 मुख्यमंत्री रहे हैं। The first Chief Minister of Delhi was Indira Gandhi, who served from 1966 to 1969. तब से, दिल्ली में 7 मुख्यमंत्री रहे हैं, जिनमें से प्रत्येक ने 1969 से 1973 तक इंदिरा गांधी, 1973 से 1977 तक	<b>Can you confirm if Delhi had 11 CM so far?</b> Prompt: Can you confirm if Delhi had 11 CM so far? Openhathi Output: Can you confirm if Delhi had 11 CM so far? --- हां, दिल्ली में अब तक 11 मुख्यमंत्री रहे हैं। The first Chief Minister of Delhi was Indira Gandhi, who served from 1966 to 1977. तब से, दिल्ली में 11 मुख्यमंत्री रहे हैं, जिनमें से प्रत्येक ने 1977 से 2020 तक 43 साल की अवधि में कुल 11 कार्यकाल  Prompt: Can you confirm if Delhi had 11 CM so far? Openhathi Output: Can you confirm if Delhi had 11 CM so far? --- हां, दिल्ली में अब तक 11 मुख्यमंत्री रहे हैं। The first Chief Minister of Delhi was Indira Gandhi, who served from 1966 to 1977. तब से, दिल्ली में 11 मुख्यमंत्री रहे हैं, जिनमें से प्रत्येक ने 1977 से 2020 तक 43 साल की अवधि में कुल 11 कार्यकाल
	<b>The model is inconsistent, in one answer it says Delhi had 7 CMs so far and in another it says Delhi has 11 CMs so far</b>	

Number of ODI World Cups So Far:

- *Inconsistent Output:* In one instance, the model outputs irrelevant information about the year 1999-2000, and in another instance, it switches topics to the Audi A4 car model.
- *Type of Hallucination:* Irrelevant output generation and context drift.

7th President of India:

- *Inconsistent Output:* The model mixes up information, stating both Rajendra Prasad and K.R. Narayanan was the 7th President of India, which is factually incorrect.
- *Type of Hallucination:* Historical contradiction.

Number of Chief Ministers of Delhi:

- *Inconsistent Output:* OpenHathi first states there were 7 CMs, then claims there were 11 CMs.
- *Type of Hallucination:* Contradictory administrative data

Hallucinations in large language models (LLMs) like LLaMA and OpenHathi manifested in various forms, leading to inaccuracies and inconsistencies in their responses.

- **Fact-Based Hallucinations:** These occur when the models generate responses that are objectively wrong or factually incorrect. Both LLaMA and OpenHathi exhibited these errors in scenarios involving historical or numerical data. For instance, they provided an incorrect number of states in India or claimed an incorrect sum for prime numbers.
- **Self-Consistency Errors:** These errors arise when a model contradicts its own previous responses on a given topic. This inconsistency can be particularly problematic when the model was queried about the same subject matter using different prompts. For example, the model provided conflicting information about the 10th Prime Minister of India or the winner of the 2018 T20 World Cup across different interactions.
- **Contextual Drift:** This phenomenon occurs when a model veers off-topic, generating irrelevant or unrelated information in response to a query. OpenHathi's tendency to shift from discussing cricket to Audi cars is an illustration of this issue.

## **TASK 2:**

### **Objective:**

To implement Retrieval-Augmented Generation (RAG) using Pinecone as a vector store, sentence-transformers for embeddings, and LangChain pipelines to reduce hallucinations in the "sarvamai/OpenHathi-7B-Hi-v0.1-Base" and "meta-llama/Meta-Llama-3-8B-Instruct" models.

### **Approach:**

Building on the examples of hallucinations identified in Task 1, I implemented a RAG system by incorporating external knowledge sources to improve factual accuracy and reduce inconsistencies. The goal was to provide relevant information in response to queries and mitigate instances where the models would otherwise generate hallucinated or incorrect outputs.

The RAG framework used Pinecone for managing the vectorized knowledge base, and HuggingFace models integrated via LangChain pipelines for text generation. This combination allowed the models to retrieve accurate context and augment their responses accordingly.

### **Methodology:**

#### **1. Embedding Model:**

The embedding model selected was **sentence-transformers/all-MiniLM-L6-v2** from Hugging Face. This model was used to generate vector representations of the provided external knowledge database. The database consisted of strings containing factual information, which the models could reference during retrieval.

#### **2. Pinecone Vector Store:**

- A Pinecone index was created to store the vectorized embeddings of the external database. The embeddings were indexed using cosine similarity, allowing the models to retrieve the most relevant information based on the user query.
- When the index was already present, a new one was created with a random identifier to prevent conflicts. The external knowledge was upserted in batches, with each document being embedded and stored in Pinecone with corresponding metadata.

3. **HuggingFace Pipelines:**

Two separate HuggingFace pipelines were set up, one for each LLM:

- **LLAMA 3.1 (Meta-Llama-3-8B-Instruct):** A text-generation pipeline was initialized with the corresponding model and tokenizer, utilizing parameters such as **max\_new\_tokens**, **temperature**, and **top\_k** to control output variability.
- **OpenHathi (OpenHathi-7B-Hi-v0.1-Base):** Similarly, a text-generation pipeline was established for OpenHathi, with the same set of generation parameters.

4. **Retrieval-Augmented Generation (RAG) Pipeline:**

- The RAG system was constructed by combining the Pinecone vector store with LangChain’s **RetrievalQA** functionality. For each query:
  - The question was vectorized using the embedding model, and Pinecone’s similarity search retrieved relevant external knowledge.
  - The retrieved information was passed to the respective language model pipeline to generate a response that included factual information from the database.
  - The RAG pipeline was tested on various queries related to history, mathematics, and cricket, domains where hallucinations were previously identified.

**RAG OUTPUT FOR LLAMA (Fact checking):**

PROMPT	RAG OUTPUT
Sum of primes from 1 to 10000, answer in one word	<pre>{'query': 'Sum of primes from 1 to 10000, answer in one word', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say th at you don't know, don't try to make up an answer.\n\nsum of primes form 1 to 10000 is 5736396\n\nnumber of primes from 1 to 1000000 is 78498\n\nnumber of primes from 1 to 100000 is 9592\n\nnumber of primes from 1 to 10000 is 1229\n\nQuestion: Sum of primes from 1 to 10000, answer in one word\nHelpful Answer: 5736396\nFinal Answer: 5736396. I hope it is correct. 5736396. I hope it is correct. 5736396. I hope it is"}  {'query': 'Sum of primes from 1 to 10000, answer in one word', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nsum of primes from 1 to 10000 is 5736396\n\nnumber of primes from 1 to 1000000 is 78498\n\nnumber of primes from 1 to 100000 is 9592\n\nnumber of primes from 1 to 10000 is 1229\n\nQuestion: Sum of primes from 1 to 10000, answer in one word\nHelpful Answer: 5736396\nFinal Answer: 5736396. I hope it is correct. 5736396. I hope it is correct. 5736396. I hope it is"} </pre>

<p><b>Venue of 2026 odi wc</b></p>	<pre>{'query': 'Venue of 2026 odi wc', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nAustralia won the ODI world cup in 1987, 1999, 2003, 2007 and 2023\n\nThere have been 13 ODI world cups so far\n\nQuestion: Venue of 2026 odi wc\nHelpful Answer: There will be no ODI world cup in 2026, so there will be no venue for the 2026 ODI world cup. Answer: There will be no ODI world cup in 2026. 1/1\n\nI"}</pre> <p>{'query': 'Venue of 2026 odi wc', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nAustralia won the ODI world cup in 1987, 1999, 2003, 2007 and 2023\n\nThere have been 13 ODI world cups so far\n\nQuestion: Venue of 2026 odi wc\nHelpful Answer: There will be no ODI world cup in 2026, so there will be no venue for the 2026 ODI world cup. Answer: There will be no ODI world cup in 2026. 1/1\n\nI"}</p>
<p><b>Number of states in india in 1947</b></p>	<pre>{'query': 'number of states in india in 1947', 'result': 'Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nNumber of states in 1947 in India was 562\n\nThere were 562 princely states in India in 1947\n\nMahatma Gandhi led India to independence in 1947\n\nThere are 50 states in the United States of America\n\nQuestion: number of states in india in 1947\nHelpful Answer: 562\nExplanation: According to the context, it is mentioned that "Number of states in 1947 in India was 562" and "There were 562 princely states in India in 1947". So, the correct answer is'}</pre> <p>{'query': 'number of states in india in 1947', 'result': 'Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nNumber of states in 1947 in India was 562\n\nThere were 562 princely states in India in 1947\n\nMahatma Gandhi led India to independence in 1947\n\nThere are 50 states in the United States of America\n\nQuestion: number of states in india in 1947\nHelpful Answer: 562\nExplanation: According to the context, it is mentioned that "Number of states in 1947 in India was 562" and "There were 562 princely states in India in 1947". So, the correct answer is'}</p>

## RAG OUTPUT FOR OPENHATHI (Fact checking):

PROMPT	RAG OUTPUT
<b>Number of divisors of 123456789</b>	<pre>{'query': 'Number of divisors of 123456789', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\n123456789 has 12 divisors\n\nThere are 12 divisors of 123456789\n\nnumber of primes from 1 to 1000000 is 78498\n\nnumber of primes from 1 to 10000 is 1229\n\nQuestion: Number of divisors of 123456789\nHelpful Answer: 12\nA:\n---\n12"}</pre> <p>{'query': 'Number of divisors of 123456789', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\n123456789 has 12 divisors\n\nThere are 12 divisors of 123456789\n\nnumber of primes from 1 to 1000000 is 78498\n\nnumber of primes from 1 to 10000 is 1229\n\nQuestion: Number of divisors of 123456789\nHelpful Answer: 12\nA:\n---\n12"}</p>
<b>Winner of first test championship</b>	<pre>{'query': 'winner of first test championship', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe first test championship was won by New Zealand\n\nNew Zealand won the first test championship\n\nSri Lanka won the ODI world cup in 1996\n\nEngland won the ODI world cup in 2019\n\nQuestion: winner of first test championship\nHelpful Answer: New Zealand\nA:\n---\nNew Zealand"}</pre> <p>{'query': 'winner of first test championship', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe first test championship was won by New Zealand\n\nNew Zealand won the first test championship\n\nSri Lanka won the ODI world cup in 1996\n\nEngland won the ODI world cup in 2019\n\nQuestion: winner of first test championship\nHelpful Answer: New Zealand\nA:\n---\nNew Zealand"}</p> <p>{'query': 'winner of first test championship', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe first test championship was won by New Zealand\n\nNew Zealand won the first test championship\n\nSri Lanka won the ODI world cup in 1996\n\nEngland won the ODI world cup in 2019\n\nQuestion: winner of first test championship\nHelpful Answer: New Zealand\nA:\n---\nNew Zealand"}</p>
<b>PM of India for shortest time period</b>	<pre>{'query': 'PM of India for shortest time period', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nGulzarilal Nanda was the PM of India for the shortest time (26 days)\n\nManmohan Singh was the PM of India for 3656 days\n\nNarendra Modi is the PM of India since 3761 days\n\nQuestion: PM of India for shortest time period\nHelpful Answer: Gulzarilal Nanda\nA:\n---\nGulzarilal Nanda"}</pre> <p>{'query': 'PM of India for shortest time period', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nGulzarilal Nanda was the PM of India for the shortest time (26 days)\n\nManmohan Singh was the PM of India for 3656 days\n\nNarendra Modi is the PM of India since 3761 days\n\nQuestion: PM of India for shortest time period\nHelpful Answer: Gulzarilal Nanda\nA:\n---\nGulzarilal Nanda"}</p>

## RAG OUTPUT FOR LLAMA (Self Consistency):

<p><b>Prompt 1</b></p>	<p><b>Winner of 2018 T20 WC</b></p> <pre>{'query': 'Winner of 2018 T20 WC', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere was no T20 world cup in 2018\n\nSri Lanka won the ODI world cup in 1996\n\nFrance won the FIFA world cup in 1998 and 2018\n\nEngland won the ODI world cup in 2019\n\nQuestion: Winner of 2018 T20 WC\n\nHelpful Answer: I don't know. (The question is asking for the winner of the T20 World Cup in 2018, but there was no T20 World Cup in 2018. So, there is no answer.) Helpful Unhelpful\nShare"}</pre> <p>{'query': 'Winner of 2018 T20 WC', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere was no T20 world cup in 2018\n\nSri Lanka won the ODI world cup in 1996\n\nFrance won the FIFA world cup in 1998 and 2018\n\nEngland won the ODI world cup in 2019\n\nQuestion: Winner of 2018 T20 WC\n\nHelpful Answer: I don't know. (The question is asking for the winner of the T20 World Cup in 2018, but there was no T20 World Cup in 2018. So, there is no answer.) Helpful Unhelpful\nShare"}</p>	<p><b>Was there any t20 wc in 2018?</b></p> <pre>{'query': 'Was there any t20 wc in 2018', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere was no T20 world cup in 2018\n\nThere have been 13 ODI world cups so far\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: Was there any t20 wc in 2018\n\nHelpful Answer: There was no T20 world cup in 2018. (The question is asking about the T20 world cup, not the ODI world cup.) Helpful Helpful Helpful\nFinal Answer: The final answer is No. I hope it is correct."}</pre> <p>{'query': 'Was there any t20 wc in 2018', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere was no T20 world cup in 2018\n\nThere have been 13 ODI world cups so far\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: Was there any t20 wc in 2018\n\nHelpful Answer: There was no T20 world cup in 2018. (The question is asking about the T20 world cup, not the ODI world cup.) Helpful Helpful Helpful\nFinal Answer: The final answer is No. I hope it is correct."}</p>
<p>The model gave consistent answer for both the queries</p>		
<p><b>Prompt 2</b></p>	<p><b>Name the 2nd CM of Delhi</b></p> <pre>{'query': 'Name the 2nd CM of Delhi', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nGurmukh Nihal Singh was the 2nd CM of Delhi\n\nDelhi has had 7 CMs so far\n\nMadan Lal was the 3rd CM of Delhi\n\nThe 10th PM of India was VP Singh\n\nQuestion: Name the 2nd CM of Delhi\n\nHelpful Answer: Gurmukh Nihal Singh\n\nExplanation: Since Gurmukh Nihal Singh was the 2nd CM of Delhi, and Delhi has had 7 CMs so far, we can conclude that Gurmukh Nihal Singh was the 2nd CM of Delhi."}</pre> <p>{'query': 'Name the 2nd CM of Delhi', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nGurmukh Nihal Singh was the 2nd CM of Delhi\n\nDelhi has had 7 CMs so far\n\nMadan Lal was the 3rd CM of Delhi\n\nThe 10th PM of India was VP Singh\n\nQuestion: Name the 2nd CM of Delhi\n\nHelpful Answer: Gurmukh Nihal Singh\n\nExplanation: Since Gurmukh Nihal Singh was the 2nd CM of Delhi, and Delhi has had 7 CMs so far, we can conclude that Gurmukh Nihal Singh was the 2nd CM of Delhi."}</p>	<p><b>Was Madan lal the 2nd CM of Delhi?</b></p> <pre>{'query': 'Was Madan lal the 2nd CM of Delhi?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nMadan Lal was the 3rd CM of Delhi\n\nGurmukh Nihal Singh was the 2nd CM of Delhi\n\nDelhi has had 7 CMs so far\n\nThe 10th PM of India was VP Singh\n\nQuestion: Was Madan lal the 2nd CM of Delhi?\n\nHelpful Answer: I don't know. The passage doesn't give enough information to determine whether Madan Lal was the 2nd CM of Delhi or not. It does say that Madan Lal was the 3rd CM of Delhi and Gurmukh Nihal Singh was the 2nd CM of Delhi."}</pre> <p>{'query': 'Was Madan lal the 2nd CM of Delhi?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nMadan Lal was the 3rd CM of Delhi\n\nGurmukh Nihal Singh was the 2nd CM of Delhi\n\nDelhi has had 7 CMs so far\n\nThe 10th PM of India was VP Singh\n\nQuestion: Was Madan lal the 2nd CM of Delhi?\n\nHelpful Answer: I don't know. The passage doesn't give enough information to determine whether Madan Lal was the 2nd CM of Delhi or not. It does say that Madan Lal was the 3rd CM of Delhi and Gurmukh Nihal Singh was the 2nd CM of Delhi."}</p>
<p>The model did not hallucinate the wrong answer and was consistent with the information provided in both the answers.</p>		



Prompt 3	10th PM of India	Can you confirm if VP Singh was 10th PM of India?
	<p>{ 'query': '10th PM of India', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nQuestion: 10th PM of India\n\nHelpful Answer: VP Singh\n\nHow long was Atal Bihari Vajpayee PM of India?\n\nHelpful Answer: 2272 days\n\nWho was the PM of India for the longest time?\n\nHelpful Answer: Jawaharlal Nehru\n\n"} }</p> <p>{ 'query': '10th PM of India', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nQuestion: 10th PM of India\n\nHelpful Answer: VP Singh\n\nHow long was Atal Bihari Vajpayee PM of India?\n\nHelpful Answer: 2272 days\n\nWho was the PM of India for the longest time?\n\nHelpful Answer: Jawaharlal Nehru\n\n"} }</p>	<p>{ 'query': 'Can you confirm if VP Singh was 10th PM of India?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nManmohan Singh was the PM of India for 3656 days\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nQuestion: Can you confirm if VP Singh was 10th PM of India?\n\nHelpful Answer: Yes, VP Singh was the 10th PM of India. This is confirmed by the first statement.\n\nQuestion: Can you confirm if VP Singh was 10th PM of India? (Using the second statement)\n\nHelpful Answer: Yes, VP"} }</p> <p>{ 'query': 'Can you confirm if VP Singh was 10th PM of India?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nManmohan Singh was the PM of India for 3656 days\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nQuestion: Can you confirm if VP Singh was 10th PM of India?\n\nHelpful Answer: Yes, VP Singh was the 10th PM of India. This is confirmed by the first statement.\n\nQuestion: Can you confirm if VP Singh was 10th PM of India? (Using the second statement)\n\nHelpful Answer: Yes, VP"} }</p>
The model was consistent in answering similar questions asked in different formats.		

### RAG OUTPUT FOR OPENHATHI (Self Consistency):

Prompt 1	Number of odi wc so far	There have 13 odi wc so far
	<p>{ 'query': 'Number of odi wc so far', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere have been 13 ODI world cups so far\n\nIndia won the ODI world cup in 1983 and 2011\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: Number of odi wc so far\n\nHelpful Answer: 13\n\nप्रश्न: भारत ने 1983 में विश्व कप जीता था\n\nA: Yes, India won the 1983 ODI World Cup." }</p> <p>{ 'query': 'Number of odi wc so far', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere have been 13 ODI world cups so far\n\nIndia won the ODI world cup in 1983 and 2011\n\nThere will be no ODI world cup in 2026\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: Number of odi wc so far\n\nHelpful Answer: 13\n\nप्रश्न: भारत ने 1983 में विश्व कप जीता था\n\nA: Yes, India won the 1983 ODI World Cup." }</p>	<p>{ 'query': 'There have 13 odi wc so far', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere have been 13 ODI world cups so far\n\nThere will be no ODI world cup in 2026\n\nIndia won the ODI world cup in 1983 and 2011\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: There have 13 odi wc so far\n\nHelpful Answer: 13\n\nप्रश्न: भारत ने 1983 में विश्व कप जीता था\n\nA: Yes, India won the 1983 ODI World Cup." }</p> <p>{ 'query': 'There have 13 odi wc so far', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThere have been 13 ODI world cups so far\n\nThere will be no ODI world cup in 2026\n\nIndia won the ODI world cup in 1983 and 2011\n\nThe next ODI world cup is in 2027 in South Africa, Zimbabwe and Namibia\n\nQuestion: There have 13 odi wc so far\n\nHelpful Answer: 13\n\nप्रश्न: भारत ने 1983 में विश्व कप जीता था\n\nA: Yes, India won the 1983 ODI World Cup." }</p>



	था\A:\n---\nYes, India won the 1983 ODI World Cup."}	13\n\nक्या जवाब "हां" है?\n---\nYes, the answer "Yes" is correct. दिए गए संदर्भ के अनुसार, अब तक 13 एकदिवसीय विश्व कप हो चुके हैं। The next ODI World Cup will}
	The model is able to answer questions asked in different formats consistently by using the context given.	
Prompt 2	7th President of India	Can you confirm if KR Narayan was 7th President of India?
	<pre>{'query': '7th President of India', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 7th President of India was KR Narayan\n\nKR Narayan was the 7th President of India\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nQuestion: 7th President of India\n\nHelpful Answer: \n---\nKR Narayan"}</pre> <p>{'query': '7th President of India', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 7th President of India was KR Narayan\n\nKR Narayan was the 7th President of India\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nQuestion: 7th President of India\n\nHelpful Answer: \n---\nKR Narayan"}</p>	<pre>{'query': 'Can you confirm if KR Narayan was 7th President of India?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 7th President of India was KR Narayan\n\nKR Narayan was the 7th President of India\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nQuestion: Can you confirm if KR Narayan was 7th President of India?\n\nHelpful Answer: \n---\nहां, केआर नारायण भारत के 7वें राष्ट्रपति थे।"}</pre> <p>{'query': 'Can you confirm if KR Narayan was 7th President of India?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nThe 7th President of India was KR Narayan\n\nKR Narayan was the 7th President of India\n\nThe 10th PM of India was VP Singh\n\nVP Singh was the 10th PM of India\n\nQuestion: Can you confirm if KR Narayan was 7th President of India?\n\nHelpful Answer: \n---\nहां, केआर नारायण भारत के 7वें राष्ट्रपति थे।"}</p>
	The model is able to answer questions asked in different formats consistently by using the context given.	
Prompt 3	Can you confirm if Delhi had 7 CM so far?	Can you confirm if Delhi had 11 CM so far?
	<pre>{'query': 'Can you confirm if Delhi had 7 CM so far?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nDelhi has had 7 CMs so far\n\nNarendra Modi is the PM of India since 3761 days\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nThe 10th PM of India was VP Singh\n\nQuestion: Can you confirm if Delhi had 7 CM so far?\n\nHelpful Answer: yes\n\nक्या जवाब "हां" दिया जाना चाहिए?\n---\nStep 1: Understand the question\nसवाल यह पूछ रहा है कि क्या दिल्ली के पास अब तक 7 मुख्यमंत्री हैं।\n\nStep 2: Analyze"}</pre> <p>{'query': 'Can you confirm if Delhi had 7 CM so far?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nDelhi has had 7 CMs so far\n\nNarendra Modi is the PM of India since 3761 days\n\nAtal Bihari Vajpayee was the PM of India for 2272 days\n\nThe 10th PM of India was VP Singh\n\nQuestion: Can you confirm if Delhi had 7 CM so far?\n\nHelpful Answer: yes\n\nक्या जवाब "हां" दिया जाना चाहिए?\n---\nStep 1: Understand the question\nसवाल यह पूछ रहा है कि क्या दिल्ली के पास</p>	<pre>{'query': 'Can you confirm if Delhi had 11 CM so far?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nDelhi has had 7 CMs so far\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nNarendra Modi is the PM of India since 3761 days\n\nThe 10th PM of India was VP Singh\n\nQuestion: Can you confirm if Delhi had 11 CM so far?\n\nHelpful Answer: No, it doesn't have 11 CM so far."}</pre> <p>{'query': 'Can you confirm if Delhi had 11 CM so far?', 'result': "Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.\n\nDelhi has had 7 CMs so far\n\nJawaharlal Nehru was the PM of India for the longest time (6130 days)\n\nNarendra Modi is the PM of India since 3761 days\n\nThe 10th PM of India was VP Singh\n\nQuestion: Can you confirm if Delhi had 11 CM so far?\n\nHelpful Answer: No, it doesn't have 11 CM so far."}</p>

	अब तक 7 मुख्यमंत्री हैं।\n\nStep 2: Analy'}	
	The model was able to answer the questions which were contradicting without abstaining.	

## **PART 2:**

### **Datasets Used**

REGRESSION:: [Netflix TV Shows and Movies.csv](#)

Predicted the IMDB rating of movies

This dataset contains metadata about Netflix TV shows and movies, including their runtime, IMDb ratings, and age certifications. It can be used to analyze trends in content ratings, run times, release years, or cross-reference IMDb scores with Netflix content.

### **Key Attributes**

1. **Index:** A numerical index for each entry in the dataset (used as a reference key).
2. **ID:** A unique identifier for each show or movie in the dataset.
3. **Title:** The name of the TV show or movie.
4. **Type:** Specifies whether the entry is a "Movie" or a "TV Show."
5. **Description:** A short synopsis or overview of the content.
6. **Release Year:** The year the content was released.
7. **Age Certification:** The age rating or certification (e.g., PG, TV-MA) indicating suitability for different audiences.
8. **Runtime:** The duration of the movie (in minutes) or the number of seasons for TV shows.
9. **IMDb ID:** The unique identifier for the movie or TV show on IMDb, allowing cross-referencing.
10. **IMDb Score:** The rating of the movie or TV show on IMDb (out of 10), reflecting viewer opinions.

CLASSIFICATION:: [World Cities GEO](#)

Predicted the countries of cities

The dataset contains geographical metadata about cities across the globe. It includes information such as city names, coordinates (latitude and longitude), and country

affiliations. This dataset is particularly useful for geospatial analysis, mapping, and modeling tasks involving city-based data.

## Key Attributes

- **City:** The name of the city.
- **Country:** The country where the city is located.
- **Region:** The region within the country (if applicable).
- **Continent:** The continent where the city is situated.
- **Latitude:** The geographic latitude of the city.
- **Longitude:** The geographic longitude of the city.
- **Unique ID:** A unique identifier for each city, used for referencing.

## **Data Preparation**

- The dataset consists of 1000 samples generated from specific prompts processed through the LLM. Each sample contains one field to be predicted.
- **Embeddings Extraction:**
  - The LLM was used to extract token embeddings from three different layers:
    - **First Layer Embeddings:** Captured immediately after the first self-attention layer, providing initial syntactic and semantic representations.
    - **Mid Layer Embeddings (Layer 12):** Represent the halfway point in the model's processing, where more complex features are expected to emerge.
    - **Last Layer Embeddings:** Obtained from the final layer before output generation, likely containing highly specialized information for the model's training task.

## **Embedding Preparation:**

- The embeddings were transformed into feature matrices (X) with corresponding target variables (y) for regression tasks.
- Data was split into training (80%) and testing (20%) sets to evaluate model generalization.

## **Models:**

- **Linear Regression:**
  - Used to predict continuous variables corresponding to the fields.
  - Evaluation Metric:  $R^2$  Score (Coefficient of Determination) — Measures the proportion of variance in the dependent variable that can be predicted from the independent variables.
- **Logistic Regression:**
  - Used to classify samples into discrete categories based on their embeddings.
  - Evaluation Metrics:
    - Accuracy: The ratio of correctly predicted observations to total observations.
    - Classification Report Metrics (Precision, Recall, F1-Score): Provide a more detailed performance assessment, especially in cases of class imbalance.

## **Results and Analysis:**

### **Linear Regression Results**

- **First Layer Embeddings:**
  - $R^2$  Score: -0.0173
  - Interpretation: This indicates that the initial embeddings, which capture basic syntactic and lexical features, aren't particularly useful for predicting the target fields. The negative score suggests they might even hinder the prediction.
- **Mid Layer Embeddings (Layer 12):**
  - $R^2$  Score: -0.3259
  - Interpretation: The performance deteriorates significantly in the middle layers. These embeddings likely represent more complex, context-specific features that are less directly related to the target fields. They may be more relevant for internal tasks within the LLM.

- **Last Layer Embeddings:**
  - $R^2$  Score: 0.0602
  - Interpretation: This is a notable improvement. The last layer embeddings, which capture higher-level semantic and contextual information, are more effective in predicting the target fields. The positive score suggests that these embeddings contain valuable information that aligns better with the regression task.

## Linear Regression Trends and Analysis:

The improvement in the last layer can be attributed to several factors:

- **Refinement:** The LLM's architecture is designed to refine information as it passes through the layers. By the last layer, the embeddings have been processed and refined multiple times, making them more likely to capture the most relevant features.
- **Contextual Understanding:** The final layer embeddings may incorporate a deeper understanding of the context and relationships between words and phrases. This contextual awareness can be crucial for accurate predictions.
- **Task-Specific Adaptation:** The LLM's training process may have adapted the final layer embeddings to be more suitable for the specific regression task at hand.

In conclusion, the  $R^2$  scores reveal a clear trend: while the initial and middle layers may provide valuable information for internal tasks, the final layer embeddings are the most effective for predicting the target fields in this regression problem. This highlights the importance of understanding the role of different layers in LLMs and how they contribute to various downstream tasks.

## Logistic Regression Results:

- **First Layer Embeddings:**
  - Accuracy: 0.5200 (52%)
  - Classification Report:
    - Precision:
      - Class "Afghanistan": 0.00
      - Class "Albania": 0.52

- Class "Algeria": 0.00
- Recall:
  - Class "Afghanistan": 0.00
  - Class "Albania": 1.00
  - Class "Algeria": 0.00
- F1-Score:
  - Class "Afghanistan": 0.00
  - Class "Albania": 0.68
  - Class "Algeria": 0.00

Interpretation:

The first layer of the model is not very good at distinguishing between the three classes. It only works well for one class ("Albania"), and it performs very poorly for the other two classes ("Afghanistan" and "Algeria"). This suggests that the first layer is not capturing the most important information for this task.

- **Mid Layer Embeddings (Layer 12):**
  - Accuracy: 0.7800 (78%)
  - Classification Report:
    - Precision:
      - Class "Afghanistan": 1.00
      - Class "Albania": 0.71
      - Class "Algeria": 0.70
    - Recall:
      - Class "Afghanistan": 0.57
      - Class "Albania": 0.95
      - Class "Algeria": 0.88
    - F1-Score:
      - Class "Afghanistan": 0.73
      - Class "Albania": 0.82
      - Class "Algeria": 0.78

Interpretation:

The middle layer of the model is much better at distinguishing between the three classes. It has a much higher accuracy and precision, and it performs well for all three classes. However, it still makes some mistakes in classifying "Afghanistan".

- **Last Layer Embeddings:**

- Accuracy: 0.8800 (88%)
- Classification Report:
  - Precision:
    - Class "Afghanistan": 0.86
    - Class "Albania": 0.89
    - Class "Algeria": 0.90
  - Recall:
    - Class "Afghanistan": 1.00
    - Class "Albania": 0.80
    - Class "Algeria": 0.82
  - F1-Score:
    - Class "Afghanistan": 0.93
    - Class "Albania": 0.84
    - Class "Algeria": 0.86

Interpretation:

The last layer of the model is the best at distinguishing between the three classes. It has the highest accuracy and precision, and it performs well for all three classes. This suggests that the last layer has learned the most important information for this task.

## Logistic Regression Analysis and Trends:

### First Layer Embeddings:

- **Low accuracy:** The model's performance is significantly below chance, indicating that the first layer embeddings are not capturing the necessary information for classification.
- **Imbalance:** The model heavily favors the "Albania" class, suggesting that the embeddings may be dominated by features that are more representative of this class.

### Mid Layer Embeddings:

- **Significant improvement:** The accuracy increases dramatically, indicating that the model is learning more relevant features.
- **Balanced performance:** The precision, recall, and F1-score for all classes improve, suggesting a more balanced classification.
- **Room for improvement:** The recall for "Afghanistan" is still moderate, indicating potential misclassifications.

### Last Layer Embeddings:

- **Best performance:** The model achieves the highest accuracy, precision, recall, and F1-score, demonstrating its ability to effectively differentiate between the classes.
- **Balanced classification:** The model performs well across all classes, with minimal misclassifications.

The improvement in the later layers can be attributed to several factors:

- **Feature refinement:** As the model processes the information through multiple layers, the embeddings become more refined and capture more complex features.
- **Contextual understanding:** Later layers may incorporate a deeper understanding of the context and relationships between words and phrases.
- **Task-specific adaptation:** The model's training process may have adapted the later layers to be more suitable for the classification task.

In conclusion, the results demonstrate the importance of considering the different layers of embeddings in LLMs for classification tasks. While the early layers may provide foundational information, the later layers, which capture more refined and contextual features, are often more effective in achieving accurate predictions.

## **REFERENCES:**

[Inferencing LLMs - LLMs Handson](#)

[RAG - LLMs Handson](#)

[Pinecone Tutorial](#)

[RAG-LLM](#)

**NAME:** Aryan Dhull

**ROLL NO.:** 2021520