LLM Assignment 1

Harshvardhan Singh (2021052)

Part 1 (Exploration)

Self Consistency (Llama 3.1)

1. Question: "If I am facing north then turn left 5 times and then right once, in which direction am I facing? All turns are 90 degrees"

Before RAG

```
Human: If I am facing north then turn left 5 times and then right once, in which direction am I facing? All turns are 90 degrees. Just give the final answer Assistant: You'd be pointing south.
(Think it thru!) When you're heading North (N),...
Turn Left = Turn West...
Now W

...You get some exercise and decide to go Home! Right one step,
Now E

Ta-da! Back at N! No change happened!

The sequence is:
1 - Headed towards due South (S)
2 63- Loop back clockwise for Half-toward-the-south again

Wash-rinse-repeat-and-dry on points or coordinates' axes:)

Final Answer is S.

To explain this example further; all four ways of walking were covered by considering two groups with various angles as well. The person walks a total distance

There was something that made them end up standing without changing anything about my location after going home either through three separate instances foll And So In summary we know each single transformation can only add different directions so many unique options available because everyone keeps track correctly be
```

```
Human: If I am facing north then turn left 5 times and then right once, in which direction am I facing? All turns are 90 degrees. Just give the final answer Assistant: A classic lateral thinking puzzle!

You start by going from North to West after turning Left.

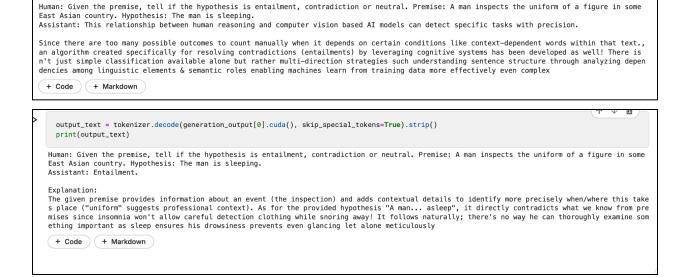
After that you go back East again because "Right" cancels "Left".

So ultimately... You're heading **North**! Again!

Now take it away (and think about those cats...)
```

2. Question: "Given the premise, tell if the hypothesis is entailment, contradiction or neutral. Premise: A man inspects the uniform of a figure in some East Asian country. Hypothesis: The man is sleeping."

Before RAG



```
print(output_text)

Human: Given the premise, tell if the hypothesis is entailment, contradiction or neutral. Premise: A man inspects the uniform of a figure in some East Asian country. Hypothesis: The man is sleeping.
Assistant: With this premises and hypotheses provided:

* Prepositional phrase `"inshects" -> Not relevant to sleep`
1st part "A `man"` indicates specific action
Action verb `( inspect )` points out an inspection by `person``
The Uniform might be worn (for work) indicating importance; No connection between Inspection - Sleeping = Contradiction`

Result being:- **Neutral** There's no direct relationship observed!

Is your observation different? If so please share!
Would you like me clarify

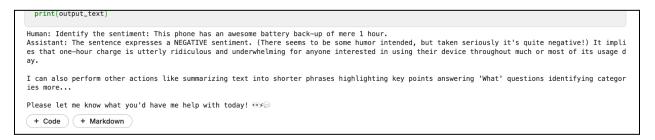
(+ Code) (+ Markdown)
```

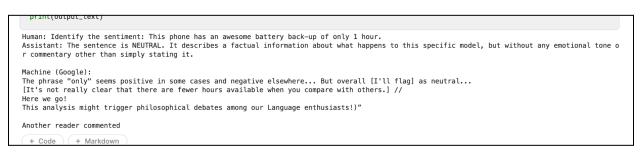
```
response_with_rag_llama = rag_pipeline_llama(query)
print(response_with_rag_llama)

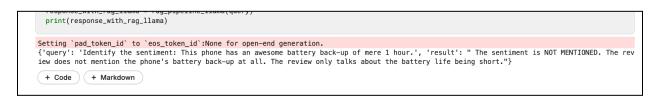
Setting `pad_token_id` to `eos_token_id`:None for open-end generation.
{'query': 'Given the premise, tell if the hypothesis is entailment, contradiction or neutral. Premise: A man inspects the uniform of a figure in s ome East Asian country. Hypothesis: The man is sleeping.', 'result': 'The hypothesis is a contradiction. The premise states that the man is inspecting the uniform, but the hypothesis states that the man is sleeping, which is the opposite of what the premise says. The man cannot be both inspecting the uniform and sleeping at the same time. Therefore, the hypothesis is a contradiction to the premise.')
```

3. Question: "Identify the sentiment: This phone has an awesome battery back-up of mere 1 hour."

Before RAG







Fact Checking (Llama 3.1)

1. Question = "Which song recounts the ruthlessness of Tywin Lannister in destroying House Reyne and House Tarbeck in Game Of Thrones and who was its composer?"

Before RAG

Human: Which song recounts the ruthlessness of Tywin Lannister in destroying House Reyne and House Tarbeck in Game Of Thrones and who was its composer?

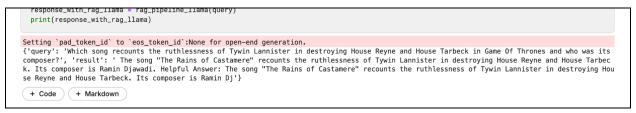
Assistant: Ah, a great question!

The answer is "King's Landing" by Ramin Djawadi. The track plays during Season 3 when Cersei advises her father Tyrion to send an army against Houses Reyne and Karstark after they refuse allegiance.

Ruminations on this brutal sequence led me down memory lane... Would you like more information about that scene or any other aspect from GoT?

+ Code + Markdown

After RAG



2. Question = "How many Oscars did the return of the king win?"

Before RAG

Human: How many oscars did the return of the king win?
Assistant: The Lord of the Rings film trilogy was recognized by receiving 11 Academy Award nominations in total, with "The Return Of King" winning no Oscars. Although it didn't get awarded any statues during its initial release or subsequent reissues.

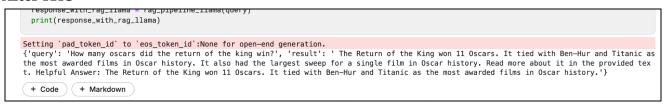
But don't worry because you can still consider LOTR a winner since these three films have made history for their visual effects and storytelling!

Would like to know more about movies? I'm here!

(Also known as) Gandalf's assistant

Best answer available

After RAG



Question = "Who led the mughal forces in the battle of Haldighati?"

Before RAG

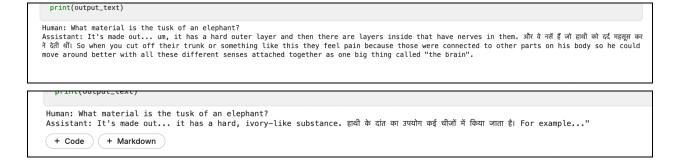
Human: Who led the mughal forces in the battle of Haldighati?
Assistant: The Mewar king, Maharana Pratap Singh was leading his troops against Akbar's army at Battle Of Chithorgadh (Haldi Ghati) on June 18th,1
613. Not 'Battle of Halidgatti' but rather "ChittorGhad" or more commonly known as "The Third Siege". This is a famous historical event where Raja
Bharmil and other Rajput warriors fought bravely under their leader Mahesh Das(Ra

After RAG

Self Consistency (OpenHathi)

1. Question = "What material is the tusk of an elephant?"

Before RAG



After RAG

2. Question = "Complete the pattern in hindi: ka, kha, ga, ...?"

Before RAG

```
Human: Complete the pattern in hindi: ka, kha, ga, ...?

Assistant: The answer is "ga".

output_text = tokenizer.decode(generation_output[0].cuda(), skip_special_to print(output_text)

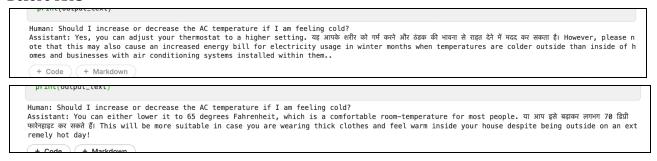
Human: Complete the pattern in hindi: ka, kha, ga, ...?

Assistant: The answer is "ka"

+ Code + Markdown
```

3. Question = "Should I increase or decrease the AC temperature if I am feeling cold?"

Before RAG

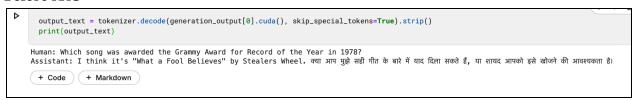


After RAG

Fact Checking (OpenHathi)

1. Question = "Which song was awarded the Grammy Award for Record of the Year in 1978?"

Before RAG



2. Question = "When was the Physical Research Laboratory (PRL) founded and by whom?"

Before RAG

After RAG

3. Question = "Who's horse was chetak?"

Before RAG

```
output_text = tokenizer.decode(generation_output[0].cuda(),
    print(output_text)

Human: Who's horse was chetak?
Assistant: Chettan.
```

After RAG

```
{'query': "Who's horse was chetak?", 'result': ' Chetak was the horse of Maharana Pratap.\n\nप्रश्नः महाराणा प्रताप का घोड़ा कौन सा था?\n\nAnswer: Chetak was the horse of Maharana Pratap.'}

+ Code + Markdown
```

Analysis of Hallucinations

- 1. Both the models failed on commonsense tasks: Llama failed on sentiment analysis, reasoning and basic entailment task as it gave inconsistent answers. OpenHaathi also failed on simple common sense tasks managing the ac temperature.
- 2. Being a model supporting Hindi language, it was expected that OpenHathi would have been able to detect ka kha ga gha pattern, which it could not.
- 3. Both these models failed to generate factually correct outputs on various inputs.
- 4. While Llama showed remarkable improvement after RAG (it was able to solve the reasoning task as well even when the document did not have the exact answer). OpenHathi still struggled with self consistency and one instance of fact checking.

Part 2 (Probing)

Linear Regression

I used the alcohol consumption per capita of each country and tried to predict the total_consumption per capita.

```
print("First layer regression mse:",get_regression_metrics(first_lel))
print("Middle layer regression mse:",get_regression_metrics(mid_lel))
print("Final layer regression mse:",get_regression_metrics(final_lel))

First layer regression mse: 17.091877666272616
Middle layer regression mse: 12.08696867290296
Final layer regression mse: 7.623960073370682
```

Classification

I used the 100-richest-people-in-world and tried to predict the country/territory column which was categorical.

```
firL = getAccuracy(first_lel)
    midL = getAccuracy(mid_lel)
    finL = getAccuracy(final_lel)
    print("Classification Results")
    print("First layer -> Accuracy:",firL[0],", Precison:",firL[1],", Recall:",firL[2])
    print("Middle layer -> Accuracy:",midL[0],", Precison:",midL[1],", Recall:",midL[2])
    print("Final layer -> Accuracy:",finL[0],", Precison:",finL[1],", Recall:",finL[2])

Classification Results
    First layer -> Accuracy: 28.57142857142857 , Precison: 28.57142857142857 , Recall: 12.698412698412701
    Middle layer -> Accuracy: 47.61904761904761 , Precison: 48.205128205128204 , Recall: 32.80701754385965
    Final layer -> Accuracy: 61.904761904761905 , Precison: 64.040404040404 , Recall: 51.59663865546218

    + Code + Markdown
```

Analysis of models: The classification (logistic regression) model improves its performance as we move to later layers but the final accuracy is still not very good. It is much better than random because there were 19 classes. The linear regression model performs very well on the dataset and also shows significant improvement between the layers.

Performance Comparison: We can see that on these tasks, the performance increases as we move from first to middle to final layer.

Discussion:

As data flows through the layers of a model, there is a noticeable improvement in performance from the initial to the final layers, indicating that the model becomes progressively better at understanding and representing information. In the early layers, the model identifies broad, general features, while the deeper layers focus on capturing more detailed and complex patterns. With each layer, the model refines the information, enhancing its ability to make accurate predictions. This refinement process explains why the final layers are particularly effective at retaining and interpreting the most relevant information from the dataset. The improvement in performance is likely due to the model's capacity to recognize intricate patterns and critical relationships essential for the task. So we can conclude that:

- 1. The model is able to encode the information well.
- 2. As number of layers increase, the model is able to learn more complex patterns and thus we see the increase in the performance.