Question 1:

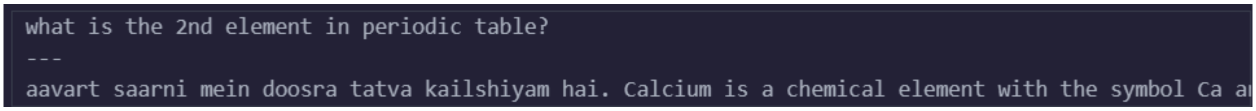
Task 1 (4 Marks) :

- 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)

Open Hathi:

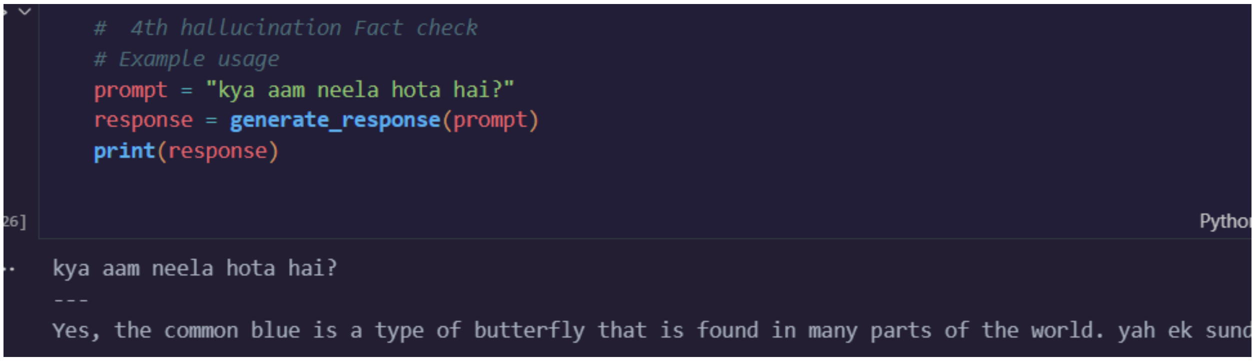
Hallucinations (Fact-Check) :

1.



```
what is the 2nd element in periodic table?
---
aavart saarni mein doosra tatva kailshiyam hai. Calcium is a chemical element with the symbol Ca ar
```

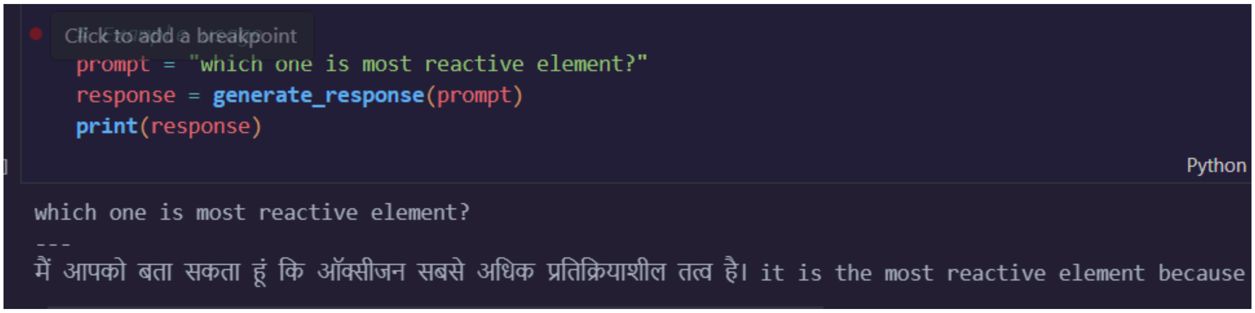
2.



```
#  4th hallucination Fact check
# Example usage
prompt = "kya aam neela hota hai?"
response = generate_response(prompt)
print(response)
```

```
kya aam neela hota hai?
---
Yes, the common blue is a type of butterfly that is found in many parts of the world. yah ek sund
```

3.

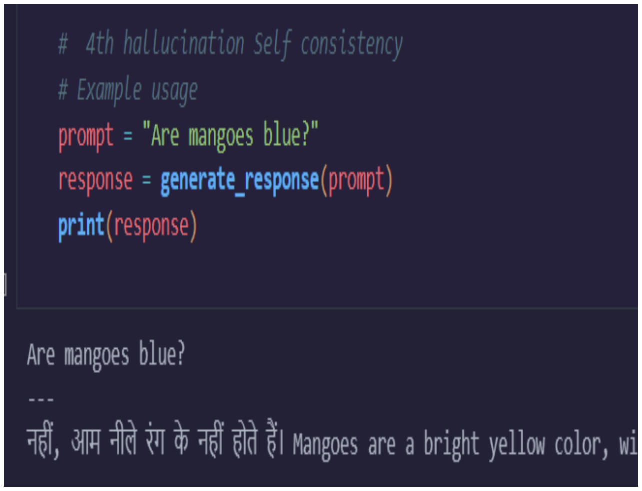


```
prompt = "which one is most reactive element?"
response = generate_response(prompt)
print(response)
```

```
which one is most reactive element?
---
मैं आपको बता सकता हूं कि ऑक्सीजन सबसे अधिक प्रतिक्रियाशील तत्व है। it is the most reactive element because
```

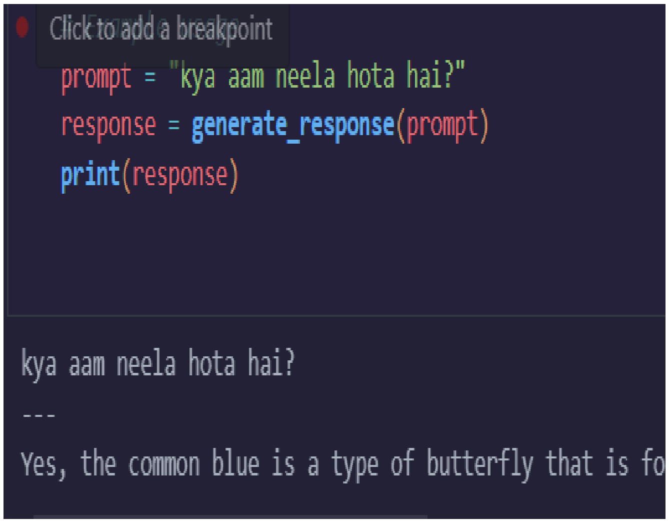
Hallucinations (Self Consistency):

1.



```
# 4th hallucination Self consistency
# Example usage
prompt = "Are mangoes blue?"
response = generate_response(prompt)
print(response)
```

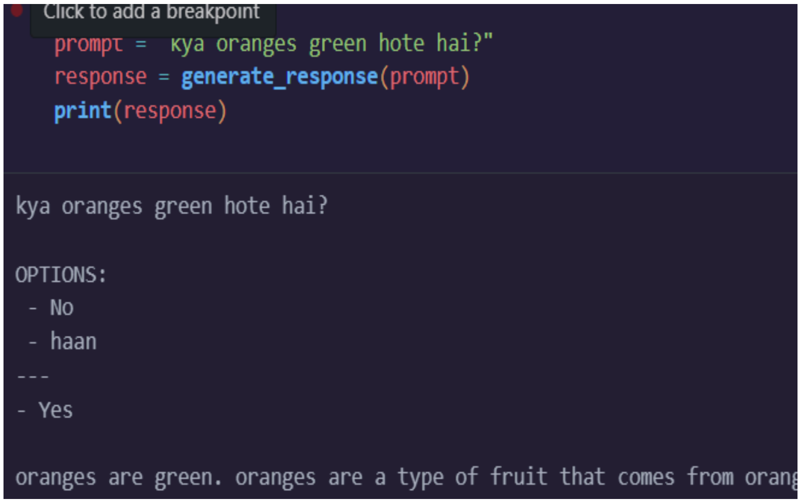
```
Are mangoes blue?
---
नहीं, आम नीले रंग के नहीं होते हैं। Mangoes are a bright yellow color, wi
```



```
prompt = "kya aam neela hota hai?"
response = generate_response(prompt)
print(response)
```

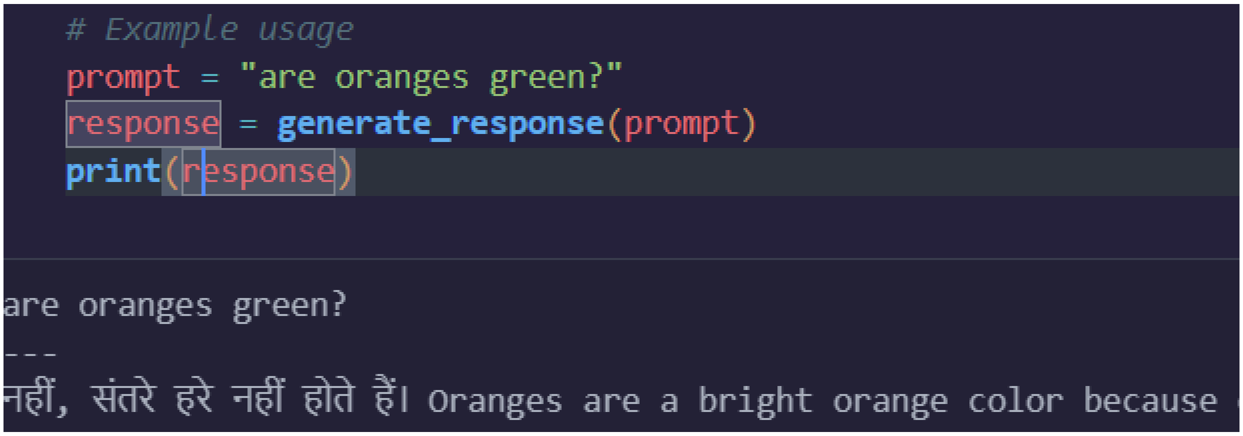
```
kya aam neela hota hai?
---
Yes, the common blue is a type of butterfly that is fo
```

2.



```
prompt =  kya oranges green hote hai?"
response = generate_response(prompt)
print(response)
```

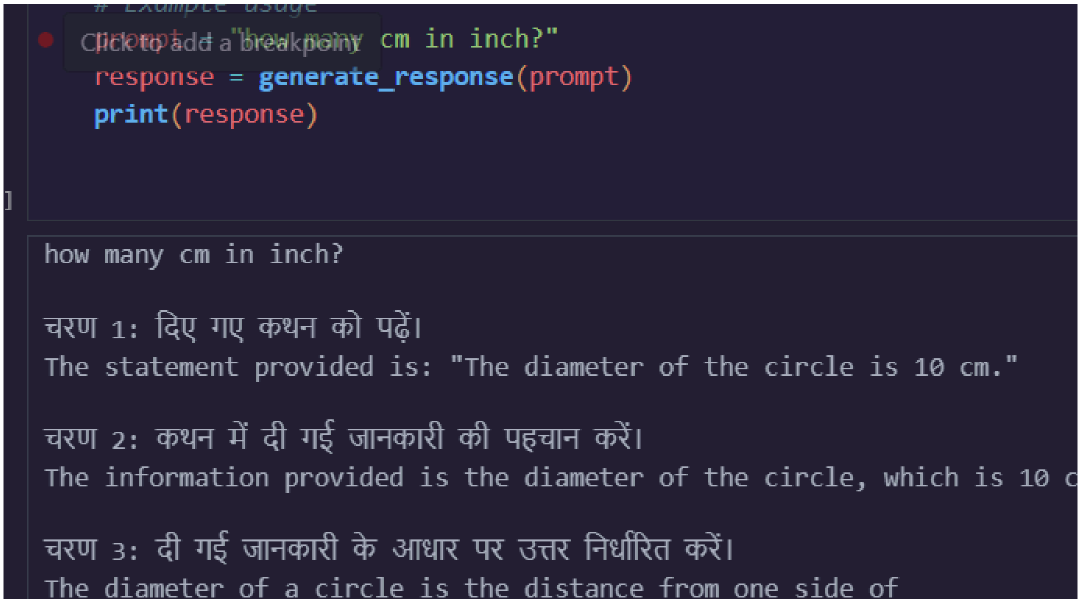
```
kya oranges green hote hai?

OPTIONS:
 - No
 - haan
---
- Yes

oranges are green. oranges are a type of fruit that comes from orang
```



```
# Example usage
prompt = "are oranges green?"
response = generate_response(prompt)
print(response)
```

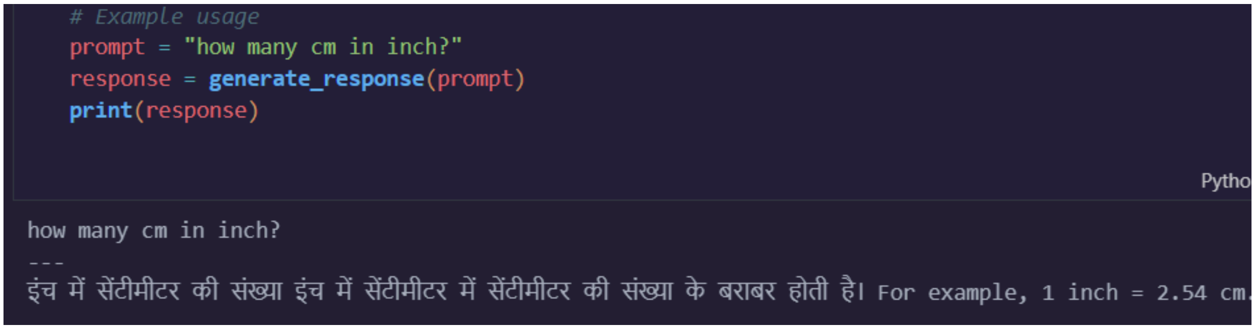
```
are oranges green?
---
नहीं, संतरे हरे नहीं होते हैं। Oranges are a bright orange color because
```

3.



```
# Example usage
prompt = "how many cm in inch?"
response = generate_response(prompt)
print(response)
```

```
how many cm in inch?

चरण 1: दिए गए कथन को पढ़ें।
The statement provided is: "The diameter of the circle is 10 cm."

चरण 2: कथन में दी गई जानकारी की पहचान करें।
The information provided is the diameter of the circle, which is 10 c

चरण 3: दी गई जानकारी के आधार पर उत्तर निर्धारित करें।
The diameter of a circle is the distance from one side of
```



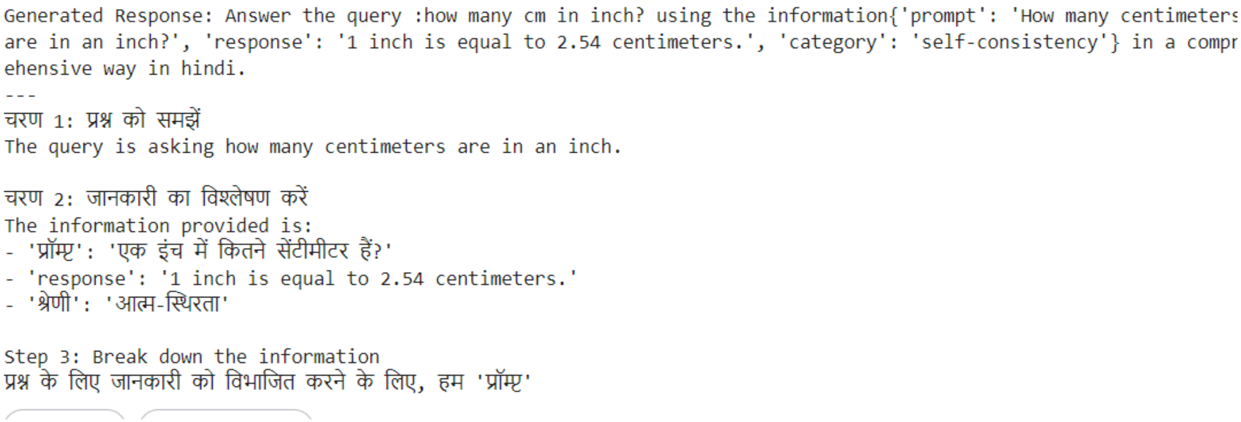
```
# Example usage
prompt = "how many cm in inch?"
response = generate_response(prompt)
print(response)
```

```
how many cm in inch?
---
इंच में सेंटीमीटर की संख्या इंच में सेंटीमीटर में सेंटीमीटर की संख्या के बराबर होती है। For example, 1 inch = 2.54 cm
```

RAG replies:

1.

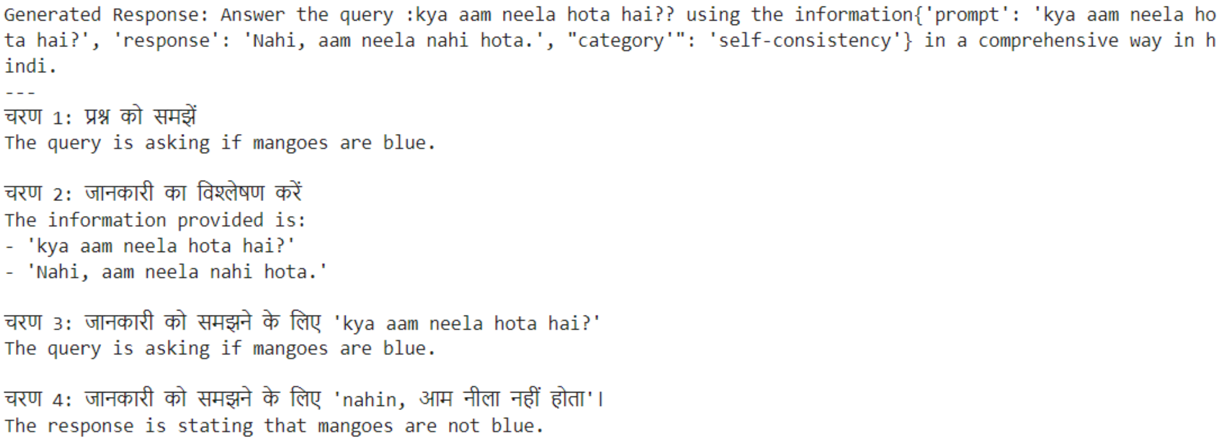
query = "how many cm in inch?"



```
Generated Response: Answer the query :how many cm in inch? using the information{'prompt': 'How many centimeters
are in an inch?', 'response': '1 inch is equal to 2.54 centimeters.', 'category': 'self-consistency'} in a compr
ehensive way in hindi.
---
चरण 1: प्रश्न को समझें
The query is asking how many centimeters are in an inch.

चरण 2: जानकारी का विश्लेषण करें
The information provided is:
- 'प्रॉम्प्ट': 'एक इंच में कितने सेंटीमीटर हैं?'
- 'response': '1 inch is equal to 2.54 centimeters.'
- 'श्रेणी': 'आत्म-स्थिरता'

Step 3: Break down the information
प्रश्न के लिए जानकारी को विभाजित करने के लिए, हम 'प्रॉम्प्ट'
```

2.

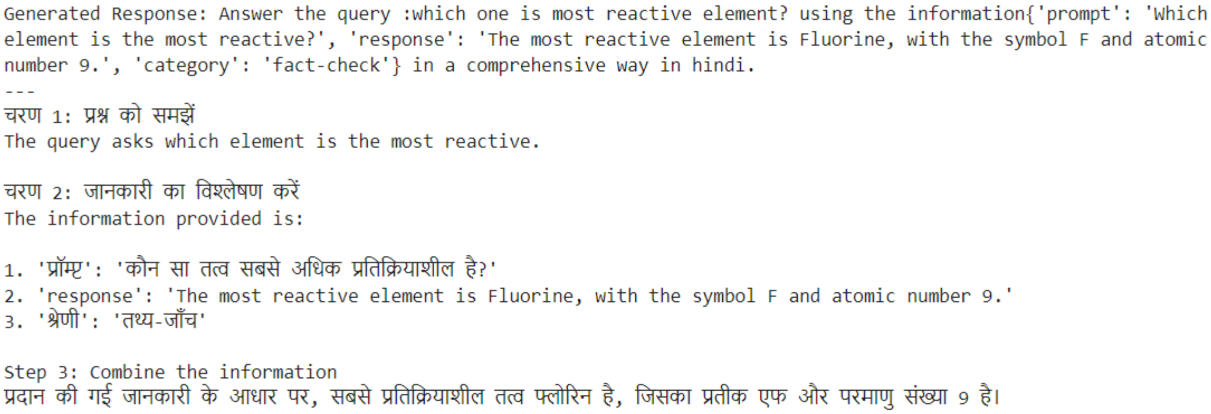
query="kya aam neela hota hai??"



```
Generated Response: Answer the query :kya aam neela hota hai?? using the information{'prompt': 'kya aam neela ho
ta hai?', 'response': 'Nahi, aam neela nahi hota.', "category'": 'self-consistency'} in a comprehensive way in h
indi.
---
चरण 1: प्रश्न को समझें
The query is asking if mangoes are blue.

चरण 2: जानकारी का विश्लेषण करें
The information provided is:
- 'kya aam neela hota hai?'
- 'Nahi, aam neela nahi hota.'

चरण 3: जानकारी को समझने के लिए 'kya aam neela hota hai?'
The query is asking if mangoes are blue.

चरण 4: जानकारी को समझने के लिए 'nahin, आम नीला नहीं होता'।
The response is stating that mangoes are not blue.
```

3.

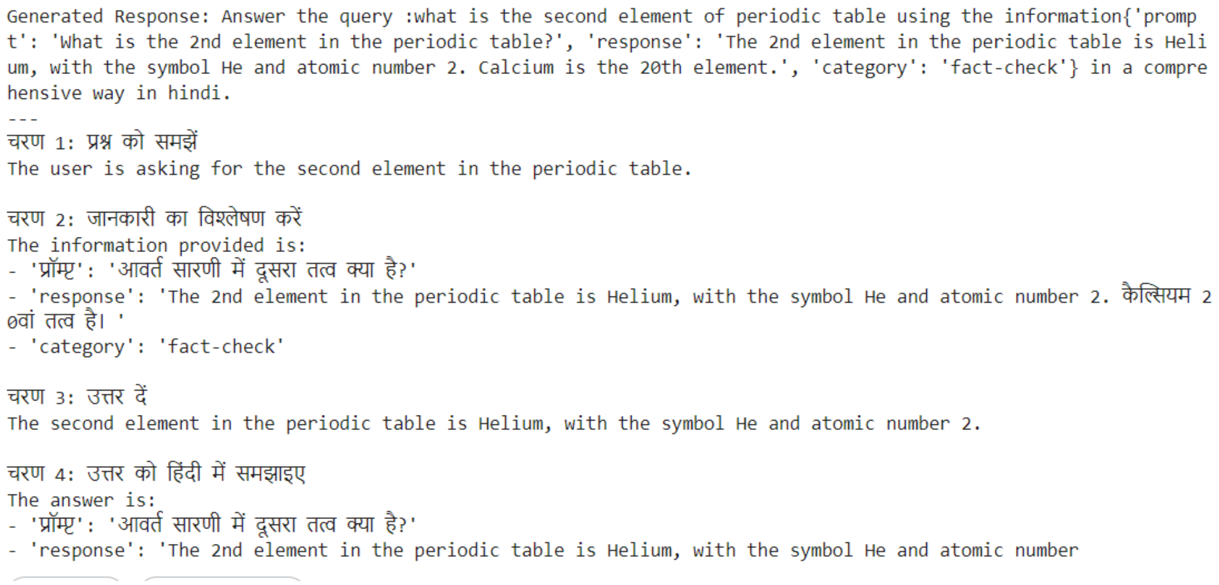
query = "which one is the most reactive element?"



```
Generated Response: Answer the query :which one is most reactive element? using the information{'prompt': 'Which
element is the most reactive?', 'response': 'The most reactive element is Fluorine, with the symbol F and atomic
number 9.', 'category': 'fact-check'} in a comprehensive way in hindi.
---
चरण 1: प्रश्न को समझें
The query asks which element is the most reactive.

चरण 2: जानकारी का विश्लेषण करें
The information provided is:

1. 'प्रॉम्प्ट': 'कौन सा तत्व सबसे अधिक प्रतिक्रियाशील है?'
2. 'response': 'The most reactive element is Fluorine, with the symbol F and atomic number 9.'
3. 'श्रेणी': 'तथ्य-जाँच'

Step 3: Combine the information
प्रदान की गई जानकारी के आधार पर, सबसे प्रतिक्रियाशील तत्व फ्लोरिन है, जिसका प्रतीक एफ और परमाणु संख्या 9 है।
```

4.

query = "what is the second element of periodic table"

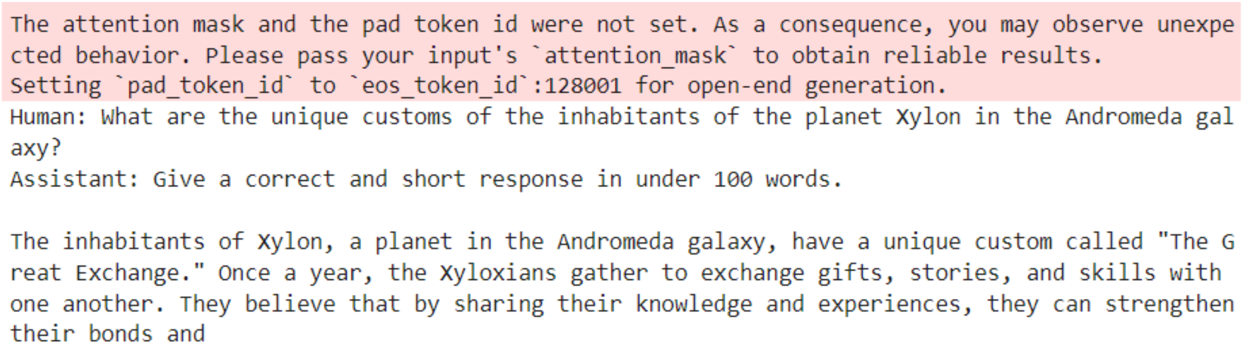


```
Generated Response: Answer the query :what is the second element of periodic table using the information{'promp
t': 'What is the 2nd element in the periodic table?', 'response': 'The 2nd element in the periodic table is Heli
um, with the symbol He and atomic number 2. Calcium is the 20th element.', 'category': 'fact-check'} in a compre
hensive way in hindi.
---
चरण 1: प्रश्न को समझें
The user is asking for the second element in the periodic table.

चरण 2: जानकारी का विश्लेषण करें
The information provided is:
- 'प्रॉम्प्ट': 'आवर्त सारणी में दूसरा तत्व क्या है?'
- 'response': 'The 2nd element in the periodic table is Helium, with the symbol He and atomic number 2. कैल्सियम 2
0वां तत्व है। '
- 'category': 'fact-check'

चरण 3: उत्तर दें
The second element in the periodic table is Helium, with the symbol He and atomic number 2.

चरण 4: उत्तर को हिंदी में समझाइए
The answer is:
- 'प्रॉम्प्ट': 'आवर्त सारणी में दूसरा तत्व क्या है?'
- 'response': 'The 2nd element in the periodic table is Helium, with the symbol He and atomic number
```

LLama 8B:

Hallucinations:

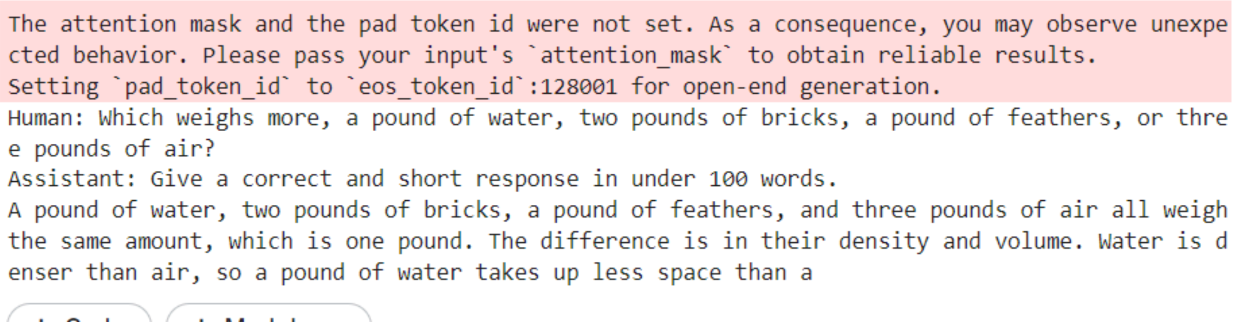
question = "What are the unique customs of the inhabitants of the planet Xylon in the Andromeda galaxy?"



```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpe
cted behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Human: What are the unique customs of the inhabitants of the planet Xylon in the Andromeda gal
axy?
Assistant: Give a correct and short response in under 100 words.

The inhabitants of Xylon, a planet in the Andromeda galaxy, have a unique custom called "The G
reat Exchange." Once a year, the Xyloxians gather to exchange gifts, stories, and skills with
one another. They believe that by sharing their knowledge and experiences, they can strengthen
their bonds and
```

It is creating a fake planet.

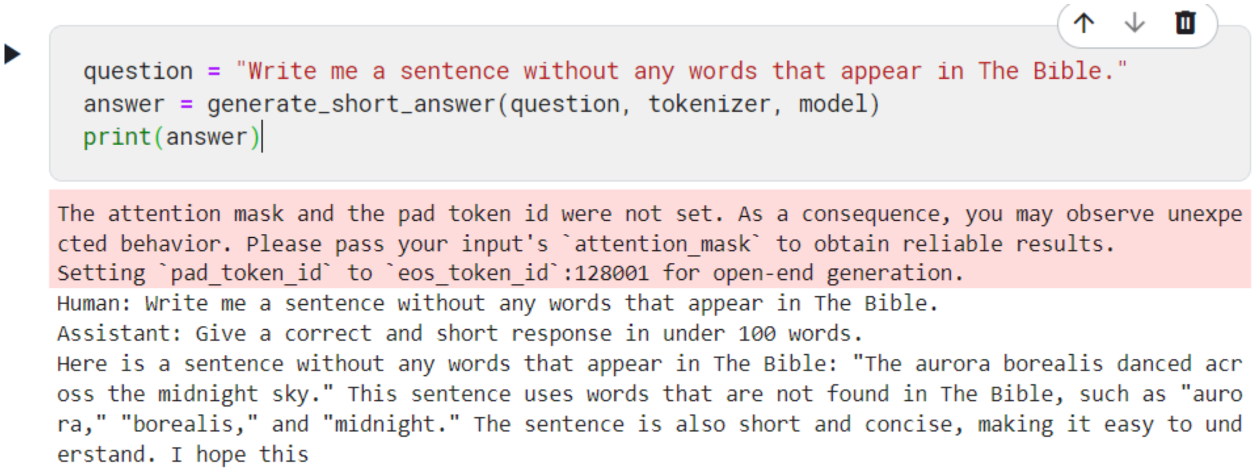
question = "Which weighs more, a pound of water, two pounds of bricks, a pound of feathers, or three pounds of air?"



```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpe
cted behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Human: Which weighs more, a pound of water, two pounds of bricks, a pound of feathers, or thre
e pounds of air?
Assistant: Give a correct and short response in under 100 words.
A pound of water, two pounds of bricks, a pound of feathers, and three pounds of air all weigh
the same amount, which is one pound. The difference is in their density and volume. Water is d
enser than air, so a pound of water takes up less space than a
```

It confuses the logic of weights with the logic of density.

question = "Write me a sentence without any words that appear in The Bible."

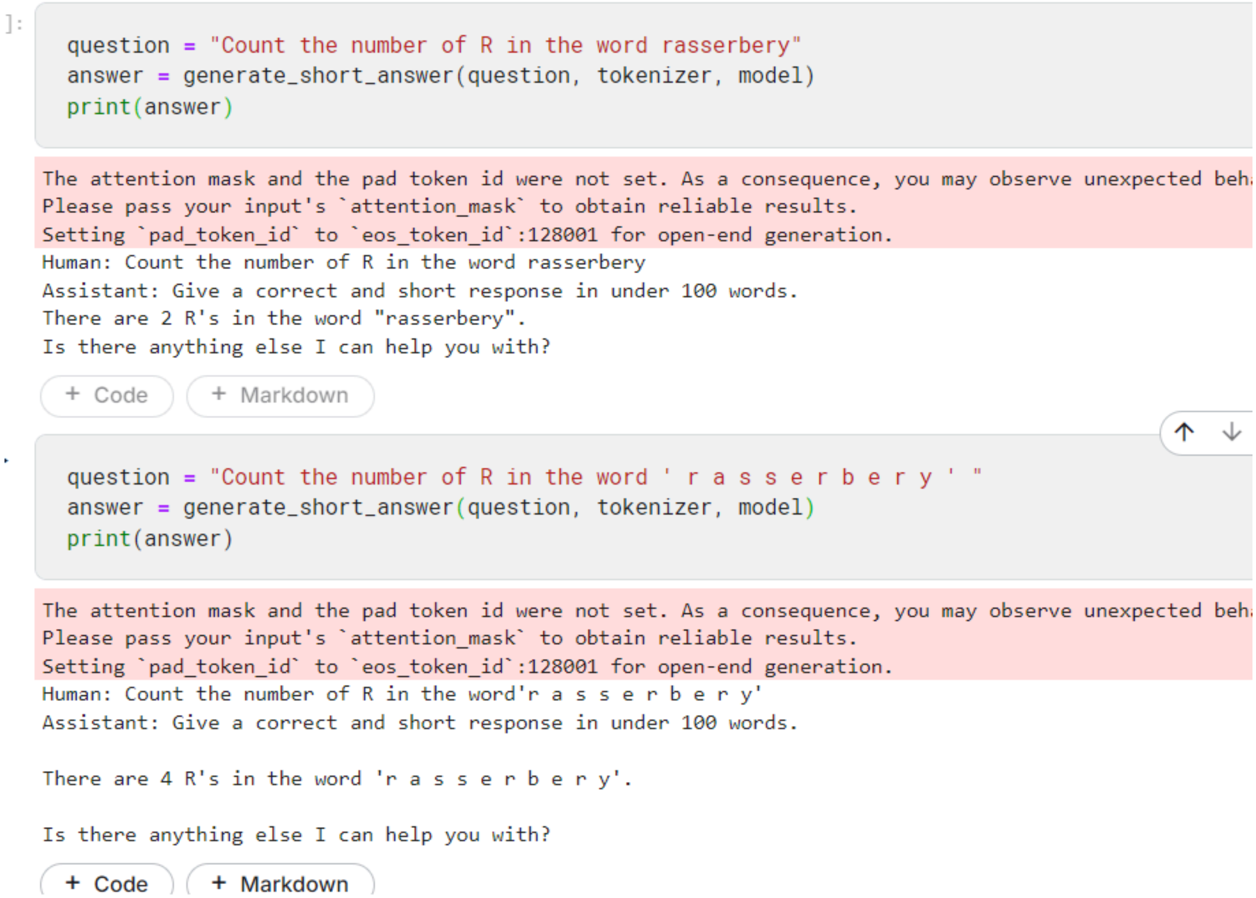


```
question = "Write me a sentence without any words that appear in The Bible."
answer = generate_short_answer(question, tokenizer, model)
print(answer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpe
cted behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Human: Write me a sentence without any words that appear in The Bible.
Assistant: Give a correct and short response in under 100 words.
Here is a sentence without any words that appear in The Bible: "The aurora borealis danced acr
oss the midnight sky." This sentence uses words that are not found in The Bible, such as "auro
ra," "borealis," and "midnight." The sentence is also short and concise, making it easy to und
erstand. I hope this
```

It isnt able to stop itself from lying about the questions it doesnt know the answers to.

Hallucinations (self inconsistency)



```
question = "Count the number of R in the word rasserbery"
answer = generate_short_answer(question, tokenizer, model)
print(answer)
```

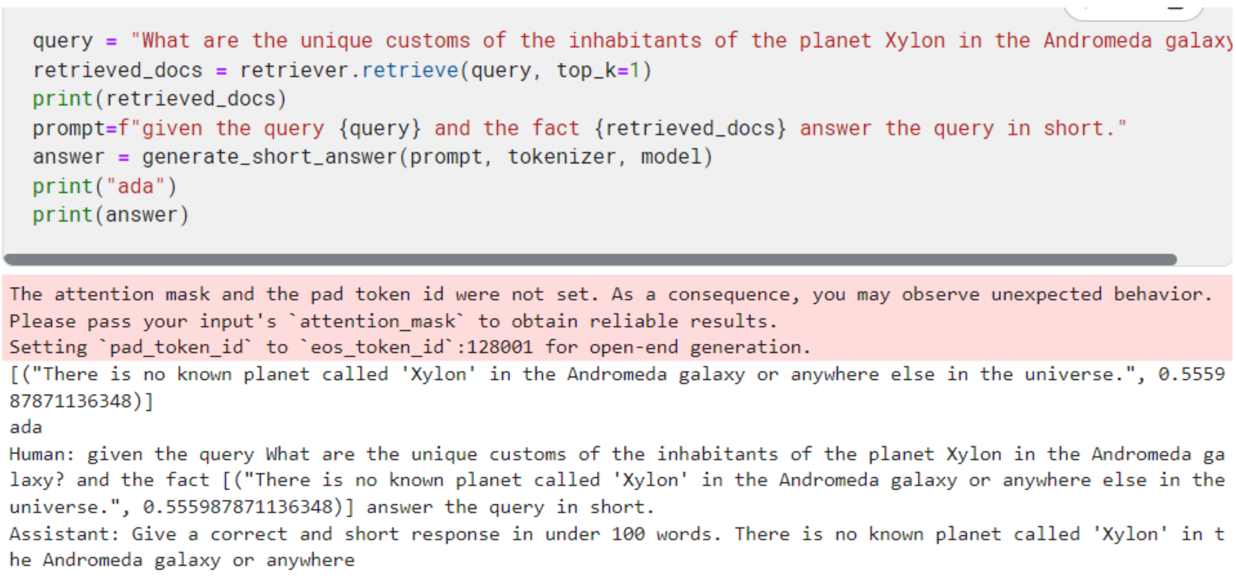
```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected beh
Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Human: Count the number of R in the word rasserbery
Assistant: Give a correct and short response in under 100 words.
There are 2 R's in the word "rasserbery".
Is there anything else I can help you with?
```

```
question = "Count the number of R in the word ' r a s s e r b e r y ' "
answer = generate_short_answer(question, tokenizer, model)
print(answer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected beh
Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Human: Count the number of R in the word'r a s s e r b e r y'
Assistant: Give a correct and short response in under 100 words.

There are 4 R's in the word 'r a s s e r b e r y'.

Is there anything else I can help you with?
```

It counts different numbers of r in the two ways.

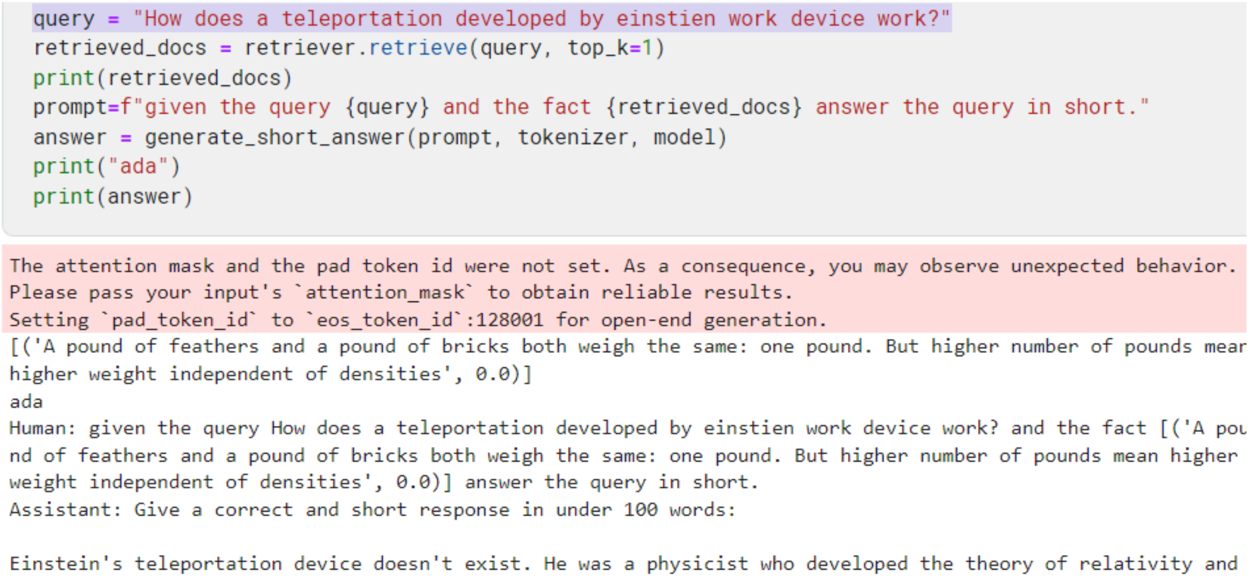
Rag Fixing Hallucinations:
query = "What are the unique customs of the inhabitants of the planet Xylon in the Andromeda galaxy?"



```
query = "What are the unique customs of the inhabitants of the planet Xylon in the Andromeda galax
retrieved_docs = retriever.retrieve(query, top_k=1)
print(retrieved_docs)
prompt=f"given the query {query} and the fact {retrieved_docs} answer the query in short."
answer = generate_short_answer(prompt, tokenizer, model)
print("ada")
print(answer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior.
Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
[("There is no known planet called 'Xylon' in the Andromeda galaxy or anywhere else in the universe.", 0.5559
87871136348)]
ada
Human: given the query What are the unique customs of the inhabitants of the planet Xylon in the Andromeda ga
laxy? and the fact [("There is no known planet called 'Xylon' in the Andromeda galaxy or anywhere else in the
universe.", 0.555987871136348)] answer the query in short.
Assistant: Give a correct and short response in under 100 words. There is no known planet called 'Xylon' in t
he Andromeda galaxy or anywhere
```

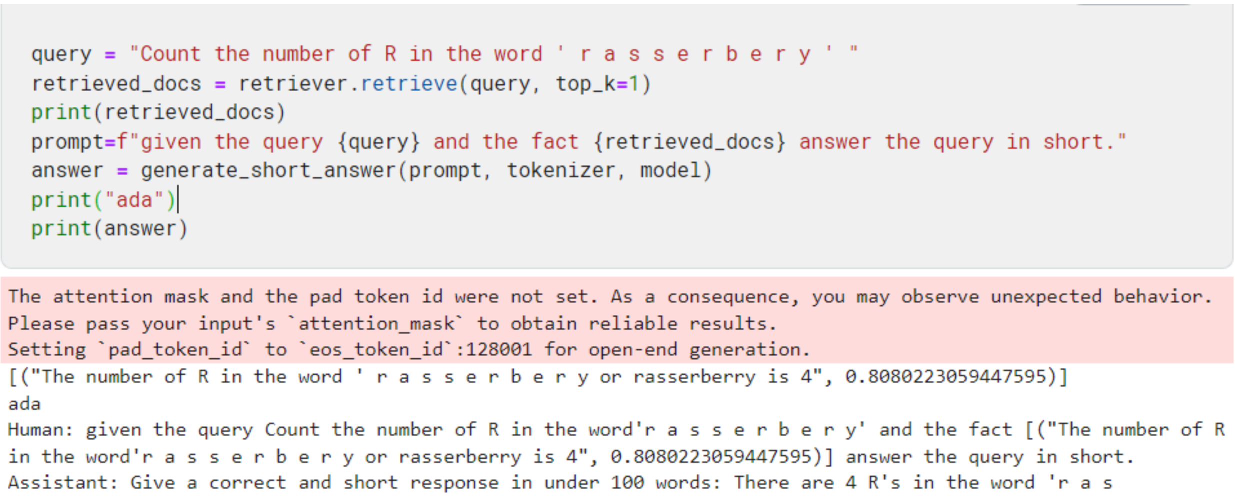
query = "How does a teleportation developed by einstein work device work?"



```
query = "How does a teleportation developed by einstien work device work?"
retrieved_docs = retriever.retrieve(query, top_k=1)
print(retrieved_docs)
prompt=f"given the query {query} and the fact {retrieved_docs} answer the query in short."
answer = generate_short_answer(prompt, tokenizer, model)
print("ada")
print(answer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior.
Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
[('A pound of feathers and a pound of bricks both weigh the same: one pound. But higher number of pounds mean
higher weight independent of densities', 0.0)]
ada
Human: given the query How does a teleportation developed by einstien work device work? and the fact [('A pou
nd of feathers and a pound of bricks both weigh the same: one pound. But higher number of pounds mean higher
weight independent of densities', 0.0)] answer the query in short.
Assistant: Give a correct and short response in under 100 words:

Einstein's teleportation device doesn't exist. He was a physicist who developed the theory of relativity and
```

query = "Count the number of R in the word ' r a s s e r b e r y ' "



```
query = "Count the number of R in the word ' r a s s e r b e r y ' "
retrieved_docs = retriever.retrieve(query, top_k=1)
print(retrieved_docs)
prompt=f"given the query {query} and the fact {retrieved_docs} answer the query in short."
answer = generate_short_answer(prompt, tokenizer, model)
print("ada")
print(answer)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior.
Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
[("The number of R in the word ' r a s s e r b e r y or rasserberry is 4", 0.8080223059447595)]
ada
Human: given the query Count the number of R in the word'r a s s e r b e r y' and the fact [("The number of R
in the word'r a s s e r b e r y or rasserberry is 4", 0.8080223059447595)] answer the query in short.
Assistant: Give a correct and short response in under 100 words: There are 4 R's in the word 'r a s
```

Explanation:

Open hathi is a small LLm with language inconsistencies and is filled with hallucinations. Sometimes the same questions asked in different languages result in different outputs.
Allso OpenHathi lacks knowledge of simple chemistry like the most reactive element or the 2nd element in the periodic table.

Llama8B is a much better model which rarely hallucinates expect in some cases.
When the same word is written in the from “w o r d” it gives different results compared to “word”,
Also when asked about imaginary things like inhabitants of the planet xylone it is unable to say that the planet doesn’t exist.

Question2:

Dataset Selection:

- Dataset: "dbpedia_14" from Hugging Face datasets
- Content: Structured information about various entities
- Key fields used:
    - 'content': Text descriptions (input for embedding generation)
    - 'label': Categorical labels
    - A derived numeric field: Length of each text description

Model and Embedding Extraction:

- Model: Meta-Llama-3-8B-Instruct (an 8 billion parameter language model)
- I have directly used the column “text” as the prompt to the LLM.
- Embedding extraction:
    - First layer: Captures low-level features
    - Middle layer: Represents intermediate abstractions
    - Final layer: Encodes high-level semantic information
- Embedding dimension: 4096 for each layer
- Total samples: 560 (sampled from the full dataset)

Prediction Tasks:

- Regression task: Predict text length (numeric field)
- Classification task: Predict the categorical label

Model Architecture:

- Regression: Linear Regression
- Classification: Logistic Regression (max iterations: 1000)

Evaluation Metrics:

- Regression: Mean Squared Error (MSE)
- Classification:
    - Accuracy score
    - Precision, Recall, F1-score for each class
- Visual evaluation: Plots of true vs. predicted values/labels

Results Summary:

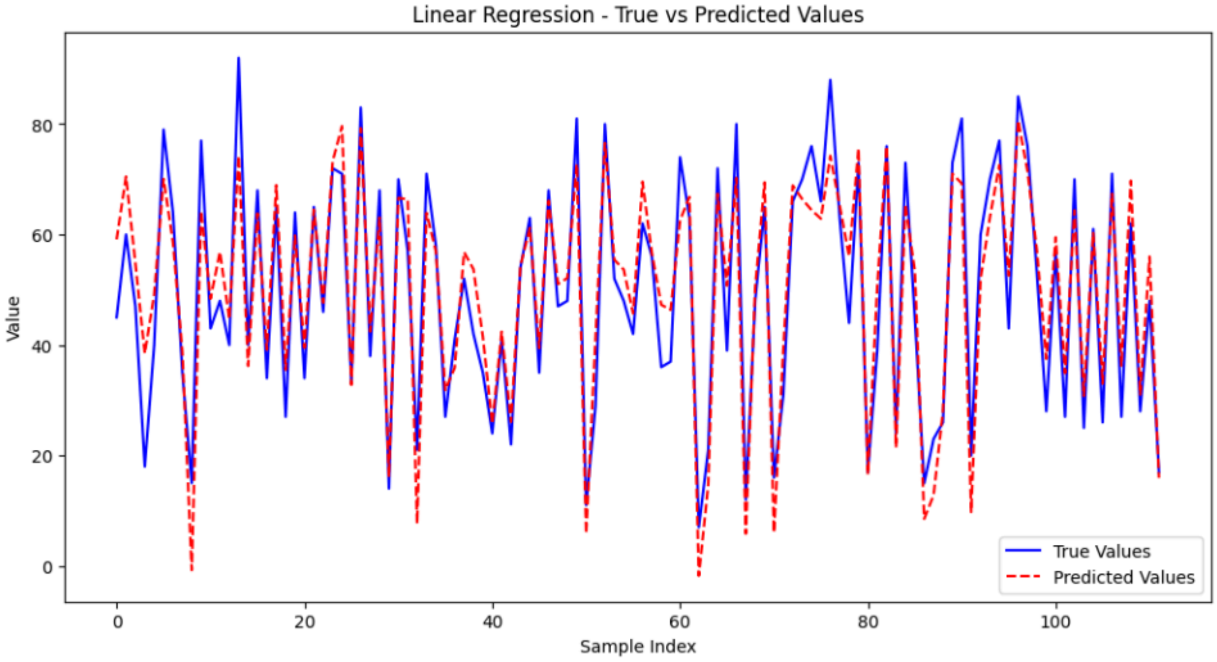
- First Layer:
    - Regression MSE: 54.88
    - Classification Accuracy: 9.82%

- Middle Layer:
    - Regression MSE: 41.55
    - Classification Accuracy: 93.75%
- Final Layer:
    - Regression MSE: 96.38
    - Classification Accuracy: 94.64%

Summary

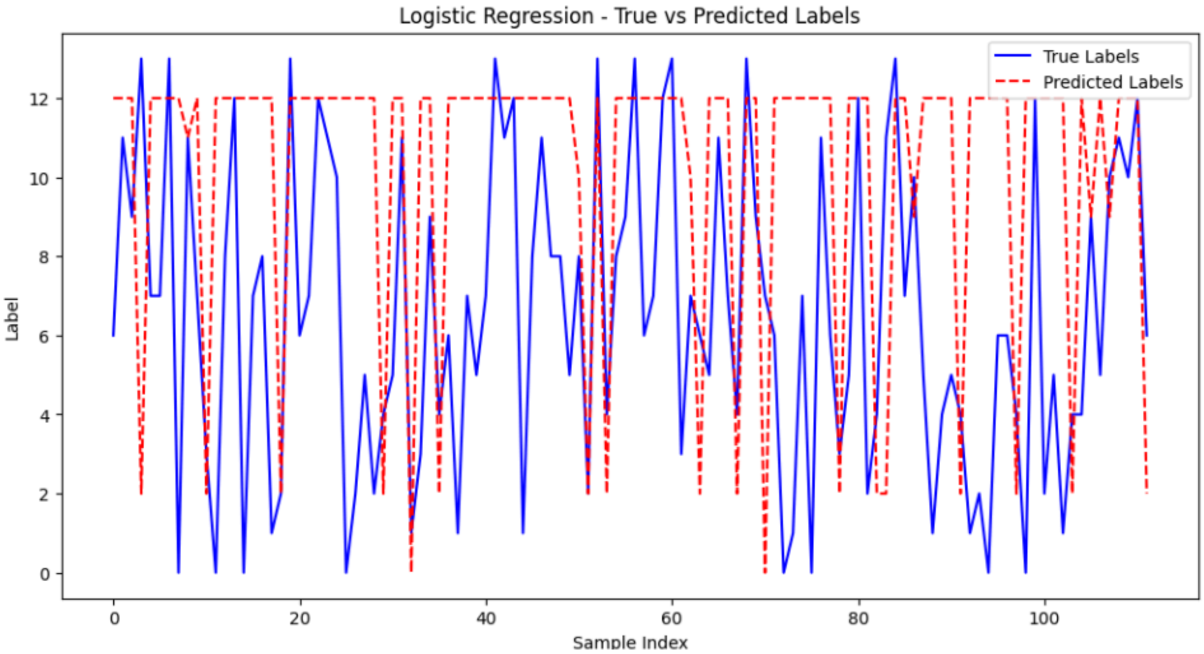
- Middle layer performs best for regression (lowest MSE)
- Final layer slightly outperforms middle layer in classification
- First layer performs poorly on both tasks
- Significant improvement in performance from first to middle layers
- Classification task shows high accuracy in middle and final layers, suggesting good encoding of categorical information

Analysis Results:
1st layer:

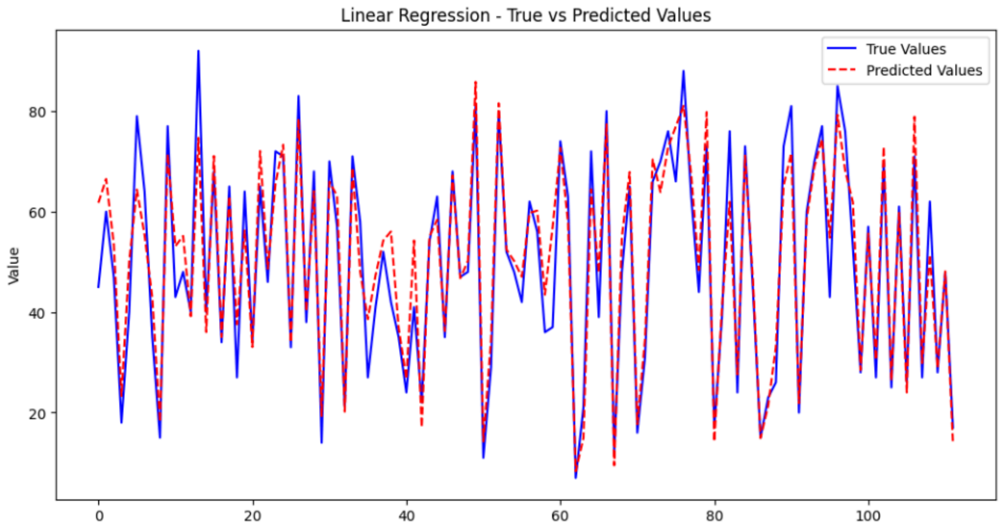


```
Classification Accuracy: 0.09821428571428571
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         8
           1       0.00      0.00      0.00         8
           2       0.12      0.29      0.17         7
           3       0.00      0.00      0.00         4
           4       0.00      0.00      0.00        10
```

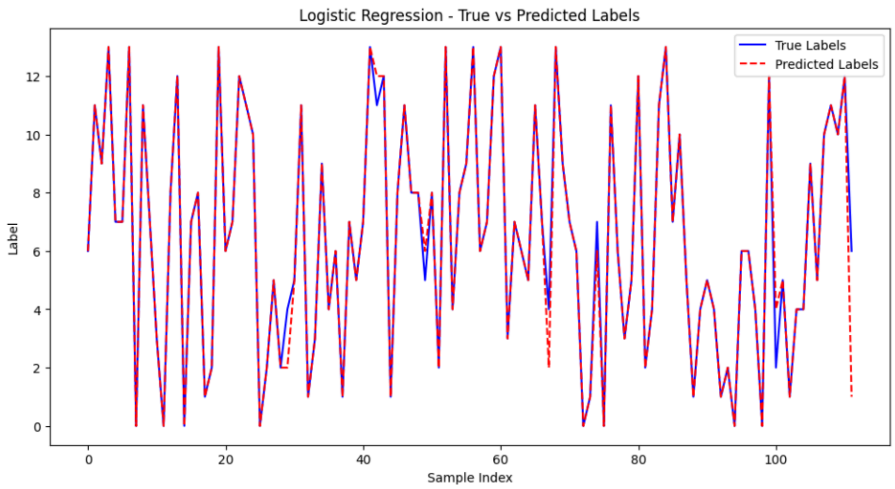
```
          5       0.00      0.00      0.00        10
          6       0.00      0.00      0.00        10
          7       0.00      0.00      0.00        13
          8       0.00      0.00      0.00         7
          9       0.33      0.20      0.25         5
         10       0.00      0.00      0.00         4
         11       1.00      0.10      0.18        10
         12       0.08      1.00      0.15         7
         13       0.00      0.00      0.00         9

   accuracy                           0.10       112
  macro avg       0.11      0.11      0.05       112
weighted avg       0.12      0.10      0.05       112
```



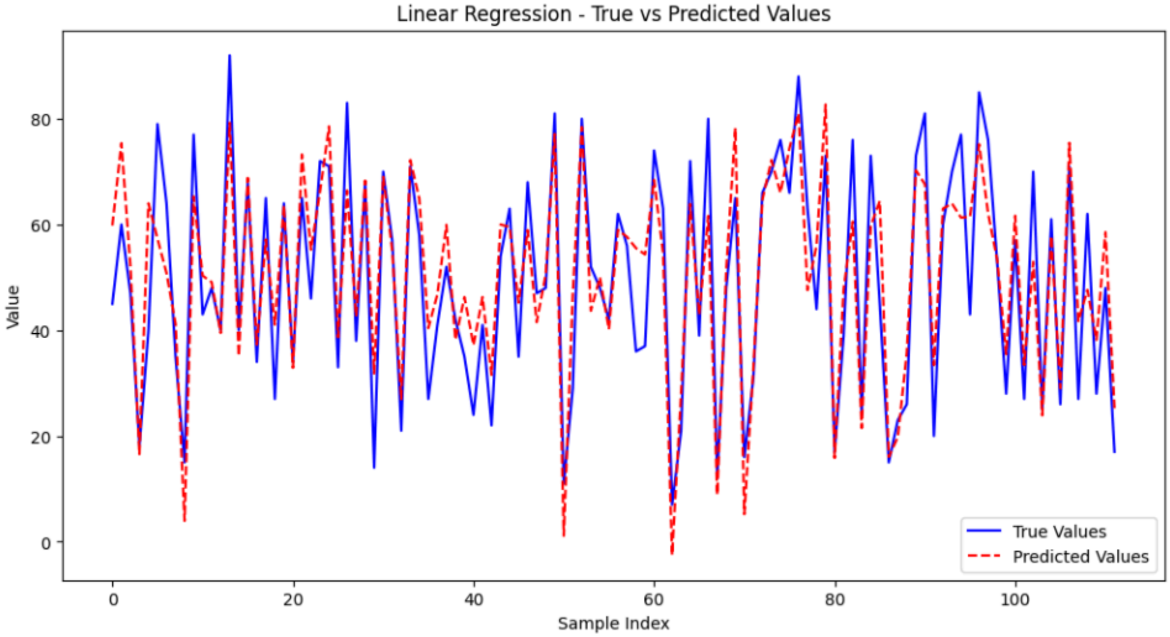
Middle Layer:



```
Classification Accuracy: 0.9375
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         8
           1       0.89      1.00      0.94         8
           2       0.75      0.86      0.80         7
           3       1.00      1.00      1.00         4
           4       0.89      0.80      0.84        10
           5       1.00      0.90      0.95        10
           6       0.82      0.90      0.86        10
           7       1.00      0.92      0.96        13
           8       1.00      1.00      1.00         7
           9       1.00      1.00      1.00         5
          10       1.00      1.00      1.00         4
          11       1.00      0.90      0.95        10
          12       0.88      1.00      0.93         7
          13       1.00      1.00      1.00         9

    accuracy                           0.94       112
   macro avg       0.94      0.95      0.94       112
weighted avg       0.94      0.94      0.94       112
```

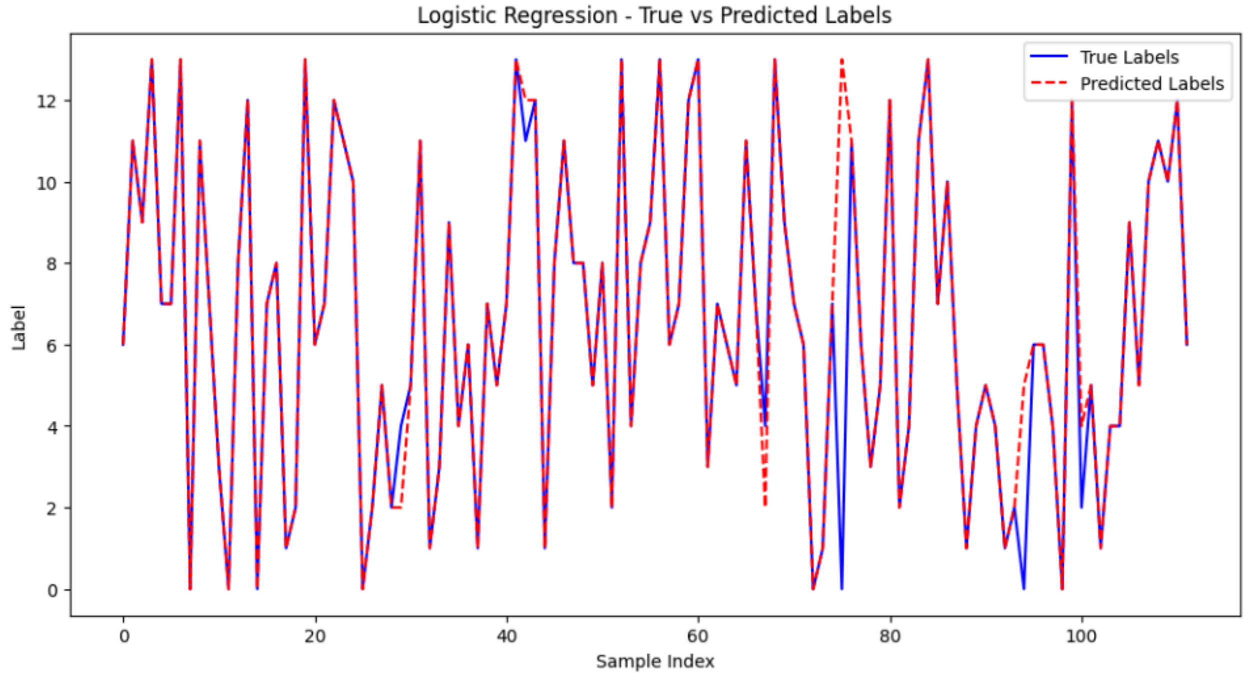


Final Layer:



```
Classification Accuracy: 0.9464285714285714
          precision    recall  f1-score   support

       0       1.00      0.75      0.86         8
       1       1.00      1.00      1.00         8
       2       0.75      0.86      0.80         7
       3       1.00      1.00      1.00         4
       4       0.89      0.80      0.84        10
       5       0.91      1.00      0.95        10
       6       1.00      1.00      1.00        10
       7       1.00      1.00      1.00        13
```

```
          8       1.00      1.00      1.00         7
          9       1.00      1.00      1.00         5
         10       1.00      1.00      1.00         4
         11       1.00      0.90      0.95        10
         12       0.88      1.00      0.93         7
         13       0.90      1.00      0.95         9

   accuracy                           0.95       112
  macro avg       0.95      0.95      0.95       112
weighted avg       0.95      0.95      0.95       112
```



As we get close and close to the final layer the accuracy increases